中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-37

论文引用格式： Chen Zhineng, Yuan Zhaoquan, Yang Xiaoshan, Cao Yixin, Li Liang, Wu Xiao, Bao Bing-Kun. Cross-modal 3D Generation: principles, methods and recent advances[J/OL]. Journal of Image and Graphics, XXXX: 1-37. DOI: 10.11834/jig.250655. (陈智能, 袁召全, 杨小汕, 曹艺馨, 李亮, 吴晓, 鲍秉坤. 跨模态3D生成：原理、方法与前沿进展[J/OL]. 中国图象图形学报, XXXX: 1-37. DOI: 10.11834/jig.250655.)
[DOI:10.11834/jig.250655]

跨模态3D生成：原理、方法与前沿进展

陈智能[1]，袁召全[2]，杨小汕[3]，曹艺馨[1]，李亮[4]，吴晓[2]，鲍秉坤[5]

1. 复旦大学可信具身智能研究院，上海 200438；2. 西南交通大学计算机与人工智能学院，四川成都 611756；3. 中国科学院自动化研究所，北京 100080；4. 中国科学院计算技术研究所，北京 100080；5. 南京邮电大学计算机学院，江苏南京 210023

摘 要： 随着虚拟现实、增强现实与数字内容创作等领域对高质量三维模型需求的快速增长，传统的人工建模与扫描方式逐渐暴露出效率低、成本高的不足，已难以满足实际应用需求。近年来，深度学习与预训练多模态大模型的发展显著推动了跨模态3D生成的性能提升与应用拓展。跨模态3D生成技术通过将文本、图像等多模态信息映射到三维表示，实现了从语义到3D内容的自动化生成等，为智能化三维内容生产提供了新的可能。该方向融合了多媒体分析、计算机视觉、自然语言处理与计算机图形学等多领域的前沿技术，但当前方法在模态间语义对齐、3D数据表示、高质量几何与纹理生成，以及生成结果的可控性与多样性等方面仍面临诸多挑战。本报告梳理了跨模态3D生成中的3D数据表示方式，涵盖显式、隐式与混合三大类别；分析了文本到3D和图像到3D的典型数据集、语义对齐机制、主流模型架构及技术路线。进一步，本文系统梳理了文本驱动三维对象生成、图像驱动三维对象生成以及三维场景生成三大方向的发展脉络与核心技术路线，总结了各方向的核心机制、代表性方法及其优势与局限。在此基础上，本文深入探讨了跨模态3D生成的未来发展趋势，指出其正加速迈向具备时空理解与交互表达能力的世界模型时代。总体而言，本综述对跨模态3D生成领域进行了系统综述，涵盖从数据表示到模型架构多个方面，旨在为后续研究提供知识框架，推动跨模态3D内容生成在世界理解与创造任务中的应用与发展。本文提及的数据集、算法已汇总至https://github.com/L-Matilda/Cross-modal-3D-Generation。

关键词： 跨模态3D生成；文本-3D生成；图像-3D生成；三维场景生成；语义对齐

Cross-modal 3D Generation: principles, methods and recent advances

Chen Zhineng[1], Yuan Zhaoquan[2], Yang Xiaoshan[3], Cao Yixin[1], Li Liang[4], Wu Xiao[2], Bao Bing-Kun[5]

1. Institute of Trustworthy Embodied Artificial Intelligence, Fudan University, Shanghai 200438, China; 2. School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu Sichuan 611756, China; 3. Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, China; 4. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China; 5. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu, 210023, China

Abstract: Cross-modal 3D generation aims to automate the synthesis of high-fidelity three-dimensional geometry and texture from 1D or 2D modalities such as text and images to bridge the semantic gap between the virtual and physical worlds. In recent years, it has become a pivotal technology in the fields of multimedia analysis, natural language processing, computer vision and computer graphics, especially in emerging industries like Virtual Reality (VR), Augmented Reality

收稿日期： 2025-12-29；**修回日期：** 2026-01-16

基金项目： 国家自然科学基金项目（项目编号：62325206；62532003；U25B2070；62322211），四川省自然科学基金项目（项目编号：2024NSFSC0508）

Supported by: National Natural Science Foundation of China (Grant Nos: 62325206; 62532003; U25B2070; 62322211), Natural Science Foundation of Sichuan Province (Grant Nos: 2024NSFSC0508)

© 中国图象图形学报版权所有

(AR), the Metaverse, digital twin manufacturing, and autonomous robotics, where the requirement of high-quality grows rapidly. Recently, the advancement of Deep Learning and Pre-trained Multimodal Large Models has significantly promoted the performance of 3D content generation and its applications. In particular, the emergence of advanced techniques, including implicit neural representations, 2D-to-3D knowledge distillation, and large-scale feed-forward reconstruction frameworks, has led to a qualitative leap in generation efficiency and visual realism. However, a comprehensive review connecting data representations, semantic alignment mechanisms, and model architectures is required to clarify the complex technical routes in this rapidly evolving field. Thus, we develop a systematic and timely review to explore the principles, methods, and recent advances of cross-modal 3D generation. First, a systematic analysis of 3D data representations is presented, categorized into explicit, implicit, and hybrid classes. Specifically, explicit representations offer compatibility with graphics engines and efficient geometry processing, while implicit representations enable continuous topology handling and photorealistic view synthesis. Second, we dissect the core mechanisms of semantic alignment and model architectures, and introduce typical datasets for these tasks. The alignment strategies are classified into contrastive and generative approaches, while the architectures cover adversarial, variational, diffusion-based, and autoregressive models. From the perspective of generation tasks, the existing methods can be divided into three major categories: Text-to-3D generation, Image-to-3D generation, and 3D Scene generation. Specifically, Text-to-3D generation has evolved from optimization-based methods to feed-forward native 3D generation, focusing on resolving multi-view inconsistency and enhancing geometry-texture decoupling. Image-to-3D generation has transitioned from single-view reconstruction using 2D priors to multi-view consistent generation, and direct regression models that infer 3D structures in seconds. 3D Scene generation extends these capabilities to complex spatial layouts, utilizing video priors and procedural generation to handle large-scale environments. In addition, we summarize mainstream datasets and analyze how the shift from large-scale synthetic repositories to real-world scanned collections impacts model generalization. Finally, this review highlights the remaining challenges in the community, including semantic ambiguity, physical inconsistency, and high computational cost, and carries out forecasting analysis. Prospects are recommended further, pointing out that cross-modal 3D generation is accelerating towards the era of "World Models," characterized by 4D spatiotemporal understanding, physical extractability, unified multimodal architectures, and 3D-native foundation models. To sum, this paper presents a systematic survey of the cross-modal 3D generation field, covering multiple aspects ranging from data representation to model architectures. It aims to provide a knowledge framework for future research and promote the application and development of cross-modal 3D content generation in tasks related to world understanding and creation. The datasets, algorithms, and evaluation metrics mentioned are linked at:https://github.com/L-Matilda/Cross-modal-3D-Generation.

Key words: cross-modal 3D generation; Text-to-3D generation; Image-to-3D generation; 3D scene generation; semantic alignment

0 引 言

随着虚拟现实(VR)、元宇宙(metaverse)以及数字孪生(digital twin)等前沿技术的迅猛发展,三维内容生成已然成为多媒体与人工智能领域的核心研究方向之一。这一趋势不仅推动了沉浸式交互体验与虚拟空间构建的需求增长,也对3D内容生成的效率、精度与语义一致性提出了更高要求。在虚拟现实领域,用户沉浸式体验所依赖的场景与交互对象几乎全部由三维生成技术驱动。无论是逼真的自然景观、复杂的建筑环境,还是栩栩如生的虚拟角色,其视觉表现力均取决于高质量的3D内容生成。在此基础上,元宇宙更进一步依赖海量且多样化的三维内容,以支撑虚拟世界的构建与持续演化。为了满足元宇宙中丰富的活动与交互需求,生成的3D内容不仅要具备真实感,还要能够实现实时交互与动态更新。相较之下,数字孪生技术则更注重对现实世界的精确映射与仿真。通过生成高精度的三维模型,数字孪生可实现对物理实体的数字化复制,从而在工业制造、城市规划、航空航天等领域中支撑复杂系统的优化设计、故障预测与远程监控。

传统的三维建模技术主要包括人工建模、三维激光扫描建模、数字近景摄影测量建模以及倾斜摄

© 中国图象图形学报版权所有

影测量建模等。然而，这些方法普遍存在建模过程耗时、对人工依赖度高、空间精度有限以及对复杂物体几何与语义信息捕获不完整等问题，难以满足对大规模、高保真三维内容的高效生成需求（Zhan等，2023）。随着人工智能生成内容（AI generated content，AIGC）技术的发展，三维内容生成逐渐发展为跨模态3D生成方法。该类方法通过在生成过程中融合语义、几何与结构等多模态特征，实现三维对象与场景的联合建模，不仅弥补了传统方法和单模态生成的不足，还显著提升了生成内容的语义一致性、结构精度与生产效率。跨模态3D生成技术为虚拟现实、元宇宙和数字孪生等应用提供了坚实的技术支撑，并为大规模、高质量及动态可交互的三维内容生成提供了新的可能性。

根据输入模态和任务目标，跨模态3D生成技术可以分为三类：文本到3D（Text-to-3D）、图像到3D（Image-to-3D）以及复杂3D场景生成。文本到3D指根据自然语言描述生成三维对象或场景。这类方法依赖文本提供的语义信息，因此主要挑战在于如何将语言语义映射为三维结构，同时保持语义与几何的一致性。图像到3D指根据单张或多张二维图像，利用其提供的视觉信息（如形状、纹理、颜色等）生成三维对象或场景。该方法的核心挑战在于如何从二维图像中准确理解物体形状并进行高质量的三维重建。复杂3D场景生成区别于前述的单一模态输入方法，其输入可为图像结合文本等多模态混合形式。该方法的核心挑战在于场景生成需要处理大尺度、复杂空间布局与多对象间的交互关系。通过整合不同模态的信息，模型不仅需要生成独立的三维对象，更要将它们合理组织成一个全局一致、结构有序的整体，以确保场景在空间上真实和在逻辑上合理。不同输入模态在此过程中各有侧重：图像提供丰富的视觉细节但缺乏高层语义，文本提供明确的语义指导但缺少具体视觉信息，而多模态融合则能兼顾两者，为构建尺度更大、对象关系更复杂的三维环境提供全面支持。图像到3D的关键问题是图像理解与三维重建，文本到3D侧重语义理解与三维生成，而复杂场景生成则需处理场景布局与多对象交互问题。

综上，尽管跨模态3D生成在不同输入模态下展现出丰富的应用潜力，但在实际生成过程中仍面临语义理解、表示学习及生成效率等方面的难题，可以归纳为以下几个核心挑战：

1）跨模态语义对齐与一致性。在跨模态3D生成中，模型需要将来自不同模态的信息（如文本描述、二维图像或其他多模态信号）准确映射到三维结构空间，这一过程存在显著挑战。首先，当输入文本包含多个抽象概念、复合属性或复杂逻辑关系时，模型往往难以准确理解其语义，并在三维生成中保持一致性。其次，不同模态在信息表达上具有天然差异：文本能够传达丰富语义但缺失几何细节，而图像等视觉输入虽包含形状、纹理等外观信息，语义抽象能力却相对有限。这种模态差异可能导致生成结果在语义或几何上偏离预期。再次，自然语言歧义及模态间信息不一致也会进一步增加对齐难度。现有评价指标在一定程度上可衡量生成结果与输入语义的对齐程度，但大多侧重于几何或像素匹配，难以全面反映跨模态语义一致性和逻辑合理性。因此，如何在跨模态映射中保证语义理解准确、信息融合有效且生成结果与多模态输入高度一致，是当前跨模态3D生成的重要挑战。

2）多模态特征与三维表示的融合。跨模态3D生成中，文本与图像等模态在信息表达上各有侧重。为实现高质量生成，关键在于将此类多源异构特征有效映射至统一的语义空间，而这一融合过程与三维表示形式的选择紧密耦合：不同的3D表示（如显式的网格、点云，或隐式的神经辐射场）因其各自在结构精度、计算效率与可编辑性上的特点，直接影响着融合后特征的表达能力与生成结果的质量。例如，网格表示利于表现细节表面但难以处理拓扑变化，神经辐射场虽能生成高质量新颖视图，其训练与优化却严重依赖有效的语义-几何对齐。当前核心问题在于，如何根据多模态输入特征，协同优化特征融合策略与三维表示方法，从而在复杂的语义约束下生成既语义符合又结构合理的三维场景。

3）计算资源与生成效率。高质量、高分辨率的3D内容生成通常需要消耗大量计算资源和内存，尤其在处理复杂三维表示（如高分辨率网格或体素）时。部分3D生成方法在训练阶段耗时较长，而基于分数蒸馏采样（score distillation sampling，SDS）（Poole等，2022）的推理方法也通常需要多轮优化，导致生成延迟增加，难以满足实时或交互式应用的需求。此外，高质量三维数据集的稀缺也限制了模型的训练效率和泛化能力，使得在保持生成精度的

© 中国图象图形学报版权所有

同时提升速度成为跨模态3D生成的关键难题之一。

本文旨在系统梳理该领域的研究进展:首先,阐述跨模态3D生成的研究背景与核心挑战;继而,剖析其基础原理,涵盖3D表示、语义对齐与生成模型;进而,分类综述基于文本、图像及复杂场景生成的关键技术路线;最后,总结当前挑战并展望构建3D生成基础模型、加强模态对齐等未来方向,为推动高效、泛化的3D内容生成提供参考。

1 跨模态3D生成原理

1.1 3D表示方式

在跨模态3D生成任务中(如从图像或文本到三维形状的生成),选择合适的3D表示至关重要。不同的3D表示会直接影响模型的表达能力、计算效率以及生成精度。如图1所示,本文将3D表示划分为三种类型:显式表示、隐式表示和混合表示。显式表示通过离散数据结构描述三维物体的几何形状,例如点云、网格和体素等,能够直观表达物体几何,但在处理复杂细节或连续表面时可能受限。隐式表示通过数学函数或神经网络表示三维形状,包括符号距离函数(signed distance function, SDF)、神经辐射场(neural radiance fields, NeRF)以及3D高斯泼溅(3D gaussian splatting, 3DGS)等,能够表达连续形状并支持梯度优化,但在直接渲染或操作时计算开销较大。混合表示结合显式与隐式的优势,在结

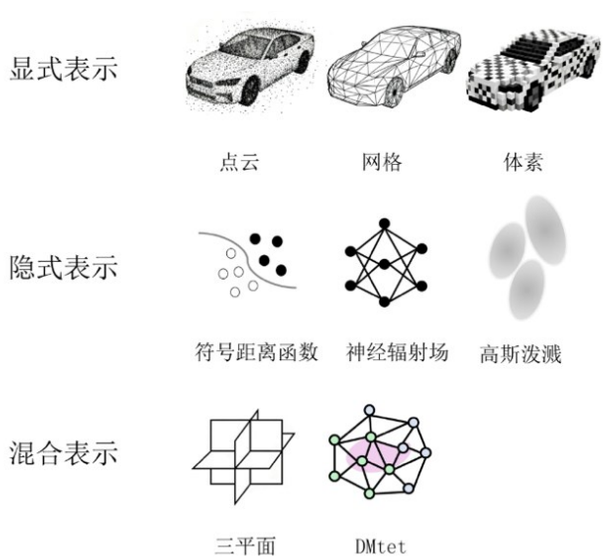

图1 3D表示方式

Fig. 1 3D Representation

构化几何基元(如体素网格或网格顶点)上引入可学习的隐式特征场,从而兼顾几何表达和可优化性。例如,三平面(tri-plane)和DMTet(deep marching tetrahedra)方法利用混合表示,在保持几何结构的同时支持高效梯度优化。

1.1.1 显示表示

1)点云是三维数据的一种基本表示形式,由欧几里得空间中无序的离散采样点构成,每个点通常以笛卡尔坐标(x,y,z)参数化,并可扩展附加属性,如RGB颜色值、表面法线或反射率等。其作为深度传感器的直接输出形式,在三维重建与生成任务中被广泛应用。

然而,点云的非结构化与无序特性使传统卷积神经网络难以直接处理,可能导致生成结果在几何细节上出现模糊性或不连续性。为此,研究者提出了多种点云特征提取与表示学习方法。PointNet(Qi等,2017)利用多层感知机(multilayer perceptron, MLP)提取全局特征向量,并通过最大池化实现点序不变性,但忽略了局部几何关系,对复杂结构建模能力有限。PointNet++(Qi等,2017)则通过分层特征学习,在局部邻域内逐步聚合特征,有效捕获了点云的局部上下文信息。随着注意力机制的发展,基于Transformer(Vaswani等,2017)的点云网络(如点云Transformer(zhao等,2021))通过自注意力机制建模点之间的长程依赖,从而显著增强了特征表达能力。为减少对标注数据的依赖,自监督学习也被引入点云建模,例如Point-BERT(Yu等,2022)和Point-MAE(Pang等,2023)通过掩蔽点建模任务学习潜在表示,实现无标注点云的高效预训练。此外,状态空间模型(state space model, SSM)被用于提升建模效率与可扩展性,PointMamba(Liang等,2024)等架构在保持全局建模能力的同时显著降低了计算复杂度。

总体而言,在跨模态3D生成任务中,点云作为显式几何表示具有天然优势。一方面,与其他3D表示相比其结构简单,点云与扩散模型和其他生成框架兼容性高,适合用于从文本或图像生成三维形状的显式表示;另一方面,高质量的点云表征可以作为多模态信息融合的桥梁,将视觉与语义特征映射到三维空间,实现语义与几何的一致性。因此,点云已成为跨模态3D生成中最具代表性的显式几何表示之一。

© 中国图象图形学报版权所有

2)网格是三维几何建模与计算机图形学中经典且应用广泛的表面表示形式,由顶点、边和面三种基本拓扑元素组成,以离散化方式精确描述物体表面几何。相比点云,网格显式编码了邻接关系,能够准确表达曲面拓扑结构和局部几何细节,使得物体表面形态可被高保真重建与渲染。其特点包括表面连续性强、拓扑信息丰富和几何表达精细。在跨模态3D生成任务中,网格能够承载从文本、图像等多模态输入映射而来的几何与语义信息,是实现高精度、结构一致三维对象生成的重要表示形式。然而,网格通常定义在非欧流形上,且拓扑高度不规则,使深度模型在处理时必须同时考虑顶点几何位置与复杂的拓扑依赖,增加了特征聚合和卷积算子设计的难度。

为应对网格表示在非欧流形上处理复杂拓扑结构的挑战,研究者主要探索了两类技术路线。第一类方法将网格建模为图结构,将顶点视为节点、边视为连接,并利用图神经网络(graph neural networks, GNNs)的消息传递机制进行特征聚合与更新,以捕获局部几何上下文。典型工作如MeshCNN(Hanocka等,2019)以“边”为基本处理单元,在网格上定义边卷积与可学习池化操作,其池化阶段通过可学习的边重要性度量选择性执行边塌陷(edge collapse),实现结构简化和层次化特征提取。MeshGraphNet(Song等,2020)则展示了图网络在物理模拟场景中的潜力,可在基于网格的流体或结构力学系统中学习节点间的时空动力学关系。第二类方法通过将不规则网格参数化到规则域,使其能够充分利用为图像等规则数据设计的成熟神经网络架构。例如,SubdivNet(Hu等,2022)框架通过重新采样,使输入网格兼容环状细分序列,形成由粗到细的网格金字塔结构,从而在网格面上直接定义类似图像卷积的操作,并支持池化与上采样机制,使经典二维卷积网络(如VGG、ResNet)可迁移至三维网格学习任务中。此外,基于流形参数化的Geodesic CNN(Masci等,2015)、Surface Network(Ran等,2022)和Spherical CNNs(Cohen等,2018)等方法也在流形域上定义卷积算子,实现拓扑一致且可扩展的特征提取。这些方法为跨模态3D生成中的网格表示处理提供了理论和实践基础,使模型能够更有效地融合来自文本、图像等多模态的语义与几何信息,实现高保真三维对象与场景生成。

总体而言,在跨模态3D生成任务中,网格作为显式几何表示的优势在于其结构化拓扑信息能够帮助生成模型更好地保持形状完整性与表面连贯性。在从图像或文本生成三维对象的过程中,网格不仅提供精确的表面几何描述,还可作为多模态特征融合的桥梁,将语义信息与视觉特征映射到三维结构中。这种特性有助于提升生成内容的几何精度与语义一致性,使跨模态3D生成在复杂场景与多对象布局中实现更高的可控性和达到更好的真实感。

3)体素是三维空间中的基本体积单元,其概念源于二维图像中的像素,通过在欧几里得三维空间中构建规则立方网格实现几何形体的离散化表示。相比网格和点云,体素显式地提供了规则化的空间结构,每个体素可存储丰富属性,如几何占用状态、颜色、材质特征或符号距离函数值,从而有效表征物体的空间结构与物理属性。其特点包括规则化结构、易于卷积操作以及对空间连续性和体积信息的天然编码能力。然而,体素的空间分辨率与存储开销呈立方增长,高精度或大规模场景下的计算成本显著增加,限制了其直接应用。为缓解该问题,研究者提出多种层次化与稀疏化改进策略,例如基于八叉树(Octree)的体素表示,仅在几何表面附近细化建模,以在保证精度的同时降低计算负担。

在跨模态3D生成任务中,三维对象的表示形式直接影响生成精度、计算效率和多模态信息融合能力。体素规则化结构天然适配3D卷积网络,适合特征学习和生成,早期如3D ShapeNets(Wu等,2015)、3D-R2N2(Choy等,2016)利用体素实现三维物体识别与多视图重建,但其存储与计算开销随分辨率立方级增长,研究者提出八叉树等层次化与稀疏化策略提升可扩展性。在渲染与生成中,VoxNet(Maturana等,2015)、Neural Volumes(Lombardi等,2019)和MPI(Zhou等,2018)等方法利用体素进行视觉合成和点云特征提取。

总体而言,体素作为规则化的三维表示,通过提供结构化的空间嵌入,有助于在跨模态3D生成任务中实现语义对齐与多模态融合。其规则网格结构不仅适合深度卷积网络进行高效特征学习,也便于将来自文本、图像等不同模态的信息映射到三维空间,从而提升生成内容的结构一致性和语义准确性。结合层次化与稀疏化策略,体素能够在保证高精度建模的同时降低计算与存储开销,使其成为连接显式

© 中国图象图形学报版权所有

几何表示与连续场生成的重要桥梁。

1.1.2 隐式表示

1)符号距离函数(SDF)是一种隐式三维几何表示方法,通过连续函数将空间中任意点映射到与物体表面最近的距离,并以符号区分点的位置关系:负值表示点在物体内部,正值表示在外部,零水平集对应物体表面。SDF的隐式特性使其天然支持平滑几何建模与布尔运算,便于生成连续且高质量的表面形状。同时,结合行进立方体(marching cubes)或行进四面体(marching tetrahedra)等等值面提取算法,可以高效将隐式表示转换为显式网格,便于渲染和后续处理。在跨模态3D生成任务中,SDF能够将来自文本或图像的语义信息映射为连续三维几何结构,实现精细的形状生成和高保真重建,尤其适用于需要平滑曲面和复杂拓扑的场景建模。具体地,DeepSDF(Park等,2019)首次将自编码器框架引入符号距离函数学习,通过网络隐式建模形状的有符号距离场,将复杂几何形状映射至低维潜空间,实现紧凑的连续形状编码与生成,奠定了基于神经隐式函数的三维形状建模基础。随后,AutoSDF(Mittal等,2022)在DeepSDF基础上引入VQ-VAE将连续的SDF表示离散化,以增强模型的稳健性与生成多样性;Diffusion-SDF(Chou等,2023)将扩散模型(diffusion model)引入SDF生成框架,在隐式距离场空间中实现高保真、多样化的三维形状生成,显著提升了生成式模型的几何细节表达能力。尽管SDF在封闭表面(closed surfaces)表示方面表现优异,但对非封闭结构(open surfaces)及复杂拓扑存在局限。为此,研究者提出无符号距离函数(unsigned distance function, UDF)及其神经扩展形式。NeuralUDF(Long等,2023)将无符号距离场与体渲染框架结合,通过学习可微密度函数实现任意拓扑结构的表面重建,从而突破了传统SDF对闭合表面的依赖;NeUDF(Liu等,2024)通过改进体渲染权重函数与采样策略,使模型仅凭多视角图像监督即可重建复杂非封闭结构的高质量表面。这些工作展示了神经隐式表示在跨模态3D生成任务中对于连续、复杂形状的高精度建模能力。

总体而言,距离函数类表示(如SDF及其变体)以其连续性、可微性和拓扑灵活性,推动了三维几何建模从显式表面向隐式场的范式转变。通过将几何形状嵌入神经网络潜空间,这类表示不仅显著提升了三维形状的重建与生成能力,还为跨模态3D生成提供了统一的几何先验,使模型能够更好地融合图像、文本等多模态信息,实现高保真、连续且拓扑灵活的三维对象生成。

2)神经辐射场(NeRF)是一种隐式神经表示方法,用于连续场景的三维建模。其核心思想是利用多层感知机学习一个函数,将三维空间坐标及观察方向映射为体积密度与辐射颜色。通过对该隐式函数进行体积渲染积分,可以沿光线累积采样点的颜色和密度,实现新视角下的高保真图像生成。相比于显式几何表示(如点云或网格),NeRF无需明确的拓扑结构即可建模复杂场景,特别适合在稀疏多视图监督下生成逼真图像。然而,原始NeRF存在计算成本高、渲染速度慢,以及对动态或大规模场景处理能力有限等挑战。

针对原始NeRF的局限,研究者在模型结构、渲染加速及应用拓展等方面提出了多种改进。模型结构方面,Mip-NeRF(Barron等,2021)通过引入圆锥台采样(cone tracing)替代点采样,缓解了多尺度渲染的混叠问题,提升了场景细节表现力;Ref-NeRF(Verbin等,2022)结合反射辐射亮度建模,增强了高光与镜面反射场景下的渲染质量。效率优化方面,Instant-NGP(Müller等,2022)采用多分辨率哈希编码,大幅降低了特征查询与网络计算开销,实现实时训练与渲染;Kilo-NeRF(Reiser等,2021)通过分治策略将场景划分为数千个小规模MLP并行推理,实现了百万级光线实时渲染。在动态与大规模场景建模方面,D-NeRF(Pumarola等,2021)引入时间编码捕捉物体随时间变化的形变,实现了动态场景的时序一致性;NeRF-Flow(Du等,2021)和TiNeuVox(Fang等,2022)等方法利用自注意力或分层时间特征处理复杂运动与遮挡;Block-NeRF(Tancik等,2022)将城市级环境划分为独立训练区块,并结合外观嵌入与相机姿态优化,实现大规模场景可扩展建模与实时渲染,为自动驾驶与虚拟街景提供基础。生成与编辑方面,GRAF(Schwarz等,2020)首次将生成对抗网络(generative adversarial networks, GAN)与NeRF结合,实现从无姿态图像学习类别级三维形状与外观分布;EditNeRF(Yuan等,2022)通过模块化网络实现形状与外观可分解建模,使用户能够通过二维草图、局部遮罩或语义指令对三维场景进行交互式编辑与生成。

© 中国图象图形学报版权所有

总体而言，神经辐射场通过将三维几何与光照信息隐式编码为连续可微场，并结合体积渲染生成多视角图像，实现了从稀疏视图到高保真场景重建的能力。其可扩展的架构、丰富的渲染细节捕捉能力及对动态与大规模场景的适应性，使NeRF不仅在传统三维重建中表现突出，也为跨模态3D生成提供了统一的几何与光照先验，成为实现语义驱动、高精度、多模态三维内容生成的重要基础。

3）3D高斯泼溅（3DGS）是一种高效的隐式三维场景表示与实时渲染技术，通过高斯分布间接建模场景几何和外观。其核心思想是将场景离散为若干带有位置、协方差、不透明度和颜色信息的高斯椭球体，在渲染时通过可微光栅化投影至二维图像并进行深度顺序的Alpha混合，实现高保真的新视角生成。相比NeRF，3DGS显著提升了渲染效率，并通过协方差自适应优化精确拟合局部几何，同时结合球谐函数建模视点相关反射，在渲染速度与质量之间取得平衡。

然而，3DGS的性能高度依赖于初始点云质量，在输入稀疏或噪声较大的情况下，重建结果容易出现伪影或几何不完整，同时其显存开销随高斯数量线性增长，复杂或动态场景的扩展仍存在挑战。为提升效率与适用性，研究者提出多种改进方法：在内存效率与几何质量优化方面，Scaffold-GS（Lu等，2024）通过引入锚点网格结构（anchor grid）对高斯参数施加结构化约束，在保证几何精度的同时显著降低内存占用；基于向量量化的参数压缩方法通过离散化高斯属性表征，不仅减少了模型存储规模，还在提升了渲染细节表现力。针对稀疏视角输入的重建问题，FSGS（Zhu等，2024）通过深度正则化与几何平滑约束，从极度稀疏的SfM（structure from motion）点云中逐步扩展并优化高斯分布，实现高完整度场景重建；Sparse2DGS（Wu等，2024）结合多视图立体（multi-view stereo，MVS）估计的密集几何先验，通过几何优先的高斯更新机制，在稀疏输入条件下获得更精确且一致的重建结果。

总体而言，3D高斯泼溅以其显式几何表示、可微优化能力和实时渲染特性，为高效三维场景建模与新视角生成提供了强大支持。在跨模态3D生成任务中，3DGS不仅能够在保持几何细节与光照真实性的同时显著提升渲染效率，还可通过各类改进方法应对稀疏输入、噪声点云及内存开销大的问题，成为连接三维重建与生成式建模的重要桥梁，为未来交互式、实时的3D生成应用奠定了技术基础。

1.1.3　混合表示

1）三平面（Triplane）是一种高效且结构化的三维场景表示方法，其核心在于将三维隐式场信息分解到三个轴对齐的正交二维平面（XY、XZ、YZ）上。每个平面存储可学习的二维特征图，对任意三维点通过双线性插值查询其平面特征，并将三个平面特征聚合后输入轻量级多层感知机解码器，以预测点的颜色、密度等属性。该方法最早由EG3D（Chan等，2022）引入三维生成领域，实现了三维隐式场的二维分解建模。其显著优势在于将原本三维体素级别的表示压缩为三个二维平面，使内存复杂度从$O(N^3)$降至$O(N^2)$，在保持高保真渲染质量的同时显著提升了训练与推理效率，从而在显式结构化特征查询与隐式连续函数建模之间实现良好平衡。

针对基础三平面框架在不同任务场景下的局限性，研究者提出了多种改进与扩展方向。例如，SemCity（Wang等，2023）将三平面表示与扩散模型相结合，面向真实户外环境的三维语义场景生成任务。该方法利用三平面特征在稀疏输入下的高效空间编码能力，有效缓解了户外场景数据稀缺与遮挡严重的问题，实现了对复杂城市环境的语义修补、扩展与结构补全。InstantMesh（Zhou等，2024）在三平面框架中集成可微分等值面提取模块，使模型能够直接将三平面特征解码为显式网格，从而取代原有体积渲染流程。这种优化设计不仅允许在网格表面直接施加深度与法线等几何约束，避免了体素渲染的高内存消耗，同时显著提升了重建精度与几何一致性。

此外，近期研究还探索了三平面与3DGS的融合（Zou等，2024；Ju等，2025）。通过在三平面上引入可学习的位置编码模块以增强高斯分布的参数化能力，该类方法进一步拓展了三平面在实时渲染与动态场景建模中的应用潜力，展示了其作为统一中间表征在3D生成任务间迁移的可行性。

三平面在跨模态3D生成中提供了结构化与连续性兼具的表达框架，已经逐渐成为连接隐式场建模与可控三维生成的重要中间表征，为多模态条件下的高保真3D内容生成奠定了基础。

2）DMTet是一种融合显式与隐式特征的混合三维几何建模框架，其核心思想是将三维空间离散化为可变形的四面体网格结构，并在每个顶点上关联

© 中国图象图形学报版权所有

距离函数值及其梯度信息。通过神经网络联合优化SDF场与顶点位移，DMTet能够在训练过程中动态调整网格形状，并借助可微分的行进四面体算法持续提取显式表面，实现了几何的端到端优化。该方法兼具显式与隐式表示的优点：一方面，其隐式函数建模能力允许网络在连续空间中表达复杂的拓扑变化；另一方面，其显式网格提取机制支持直接施加基于表面的几何与感知损失，从而有效提升重建结果的平滑度与结构精度。

在此基础上，研究者进一步提出了多种基于DMTet的改进方法以提升生成质量与适用性。例如，Magic3D（Lin等，2023）采用“粗到细”的两阶段优化流程，将隐式NeRF场转换为DMTet网格，并结合扩散模型与分数蒸馏采样实现从文本生成高质量三维内容。FlexiCubes（Liao等，2023）在等值面参数化中引入可学习的局部自由度，使网格几何在训练中自适应调整，从而兼顾几何保真度与可控性。

DMTet以其可微分的网格生成机制与隐式场的连续表达能力，在跨模态3D生成中实现了几何结构与语义条件的高效耦合。它通过显隐式融合的建模方式，不仅提升了三维重建与生成的几何保真度，也为文本到三维、图像到三维等多模态生成任务提供了稳定且可控的几何支撑。

1.2 多模态语义对齐机制

在跨模态3D生成任务中，文本、图像与三维几何数据在数据结构、语义粒度与表征空间上存在显著差异。为了实现从语言与视觉信息到三维空间的统一理解与生成，核心在于构建一个共享的语义潜空间，使不同模态的特征能够在其中通过相似度或距离度量反映语义一致性，如图2所示。

1.2.1 预训练多模态大模型对齐

多模态预训练模型（如CLIP（Radford等，2021）、BLIP-2（Li等，2023）、EVA-CLIP（Fangl等，2023）、ALIGN（Jia等，2021）、Florence-2（Yuan等，2021））通过大规模图文数据学习到统一的跨模态语义空间，为三维生成提供了稳定且可泛化的语义表征。尽管这些模型本身并不直接处理三维数据，但在Text-to-3D或Image-to-3D任务中，3D生成方法会将渲染出的多视角图像输入这些模型，与文本或图像特征进行对比，从而形成一个外部的优化循环，用语义反馈引导三维形状与外观的改进。基于这种“利用多模态模型构造的语义监督回路”，三维生成模型能够借助纯2D/文本知识实现可控、语义一致且结构稳定的三维重建或生成，大幅减少对大规模三维数据的依赖。

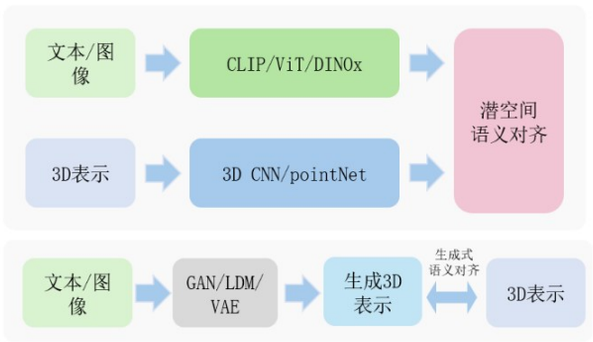

图2 多模态语义对齐机制概述

Fig. 2 Overview of the Multimodal Semantic Alignment Mechanism

该范式的优势在于其强大的零样本泛化能力与数据效率。它绕过了对稀缺且昂贵的配对3D标注数据的依赖，直接利用从互联网规模数据中学习到的丰富语义先验。例如，CLIP模型对开放词汇的理解能力，使得生成模型能够响应广泛的、甚至未见过的文本描述。然而，其核心局限源于其二维本质。预训练模型学习的语义空间建立在二维图像投影之上，缺乏对三维空间连续性和物理结构的内在理解。这导致优化过程极易陷入局部最优，产生诸如“Janus效应”、几何模糊或空间结构不合理等问题，因为优化目标仅在于匹配二维渲染图的语义，而非构建一个自洽的三维实体。

1.2.2 图像-3D语义对齐

图像-3D特征映射旨在建立二维视觉与三维几何的语义对应，实现跨模态特征对齐。其核心是在共享潜空间中统一表示图像的外观、纹理、光照与深度信息及三维对象的几何与拓扑，使视觉相似的图像与结构相似的三维形态在语义空间中一致。近年来，研究者在利用深度特征与图结构进行跨模态对齐方面取得了进展（Liu等，2024）。主流方法借助视觉-语言预训练模型（如CLIP、DINOv2（Caron等，2021）、BLIP-2、EVA-CLIP）提取视觉语义特征，并与三维编码器（PointNet、VoxelNet、3D CNN、Point Transformer（Zhao等，2021）、SparseConvNet（Graham等，2018））提取的几何特征对齐，通过跨模态映射或共享投影层嵌入统一潜空间。研究主要分两类：对比式表征对齐和生成式隐空间对齐方法。

© 中国图象图形学报版权所有

对比式表征对齐利用跨模态对比损失（如InfoNCE或NT-Xent Loss）在配对样本上优化语义距离，以最大化匹配样本的相似度、最小化非匹配样本的相关性，从而构建全局一致的视觉-几何语义分布。其关键技术在于如何设计有效的正负样本对以及特征投影网络，以克服二维图像与三维形状在数据结构和信息密度上的固有差异。CLIP-Forge提出利用CLIP提取文本/图像嵌入，并通过可逆流模型映射到3D潜空间，以最大化正样本相似度、最小化负样本相关性实现语义对齐（Sanghi等，2022），其后续CLIP-Sculptor（Sanghi等，2023）在多分辨率生成中保持这一语义一致性。ISS（Liu等，2022）方法提出通过文本-图像对齐优化嵌入，再以图像条件生成3D，支持零样本文-3D生成；CISP/IC3D提出先用Image-Shape对比式预训练构建联合嵌入，再训练图像条件的3D扩散模型，实现单视图重建与遮挡补全（Sbrolli等，2022）。PointE提出采用两阶段流水线，利用渲染图像嵌入与文本对齐，再通过图像条件扩散生成点云（Nichol等，2022）；OpenShape（Liu等，2023）提出通过大规模三模态对比学习，将3D嵌入对齐至CLIP图像-文本空间，实现文本或图像直接条件化3D表征。类似地，针对点云数据，国内研究者提出了基于层次化对比学习的跨模态预训练框架，有效提升了点云与文本的对齐精度（Wang等，2022）。CLIP2Scene（Chen等，2023）提出将CLIP图像特征映射至3D场景表示，用于语义分割与跨模态迁移，CLIP-FO3D提出类似方法（Zhang等，2023）；大重建模型（large reconstruction model）（Hong等，2024）提出通过图像-3D对比式预训练优化单视图-3D Transformer，使隐式3D表示高效预测。

对比式对齐方法的优势在于其框架简洁、训练相对稳定，且能构建一个度量意义明确的跨模态共享空间。这使得其非常适合于跨模态检索、零样本分类等任务。例如，OpenShape展示的强大零样本3D分类能力。然而，其局限性也十分明显。首先，性能高度依赖大规模、高质量的图像-3D配对数据，数据集的偏差和噪声会直接影响对齐质量。其次，全局对比损失倾向于学习样本级别的语义相似性，难以建立像素级或局部特征级的精细对应关系，导致在需要高精度几何重建的任务中细节保留不足。最后，该方法通常是一种判别式框架，在生成多样化和高保真3D形状方面能力有限。

生成式隐空间对齐在此基础上引入生成式建模框架，通过扩散模型、GAN或变分自编码器（variational autoencoders，VAE）等机制，将图像嵌入作为条件输入，引导三维形状的生成与重建过程，从而在生成阶段实现隐式语义对齐。这类方法的核心优势在于能够学习从图像到三维结构的连续、概率性映射，通过生成过程中的多步去噪或对抗训练，隐式地编码了复杂的跨模态对应关系。ImageDream（Wang等，2023）提出单张图像-3D映射通过多视角扩散网络生成一致的多视图渲染，再利用SDS优化隐式3D表示（如NeRF）；Zero-1-to-3提出基于扩散模型的条件多视图生成，通过多视角图像构建潜空间并生成隐式三维模型（Liu等，2023）；Zero123++也提出类似方法（Shi等，2023）。Human-SGD提出针对单人像生成，通过估计粗略几何与背面信息，再利用扩散模型逐步生成多视图图像，并通过逆向渲染获得高分辨率3D网格（AlBahar等，2023）；CAT3D（Gao等，2024）提出模拟拍摄流程，从少量图像生成一致的新视图，再经NeRF重建3D场景；Acc3D（Liu等，2025）提出在扩散生成与SDS优化中引入边缘一致性约束与对抗学习，实现低采样步长下的高质量3D重建；Lyra（Bahmani等，2025）提出利用视频扩散潜空间的3D信息自蒸馏为显式3D高斯场景，实现大规模、多样化且几何一致的生成；GET3D（Gao等，2022）提出通过GAN从2D图像集合直接生成高保真纹理化3D网格。在纹理生成与几何生成的解耦研究方面，国内工作通过引入分离的几何与外观潜码，提升了生成模型的可控性与编辑能力（Zhang等，2023）。

生成式对齐方法的优势在于其强大的数据分布建模能力和高保真输出潜力。通过扩散模型或GAN，该方法能够学习从图像到复杂3D表示的连续、非线性映射，在生成过程中隐式地融合了三维几何先验，因此能产生细节丰富、视觉逼真的结果。例如，Zero-1-to-3展示了从单图生成几何一致新视图的强大能力。其局限性主要体现在两方面：一是训练复杂度高，通常需要海量数据且训练不稳定（尤其是GAN）；二是“黑箱”特性明显，生成过程的可解释性和可控性较差，用户难以精确干预生成结果的特定属性。此外，基于SDS的方法还存在优化速度慢、易出现过平滑等固有问题。

1.2.3 文本-3D语义对齐

文本-3D语义对齐的核心目标在于通过建立语

© 中国图象图形学报版权所有

言描述与三维结构之间的语义映射，实现跨模态的语义一致性与特征互通。这一过程旨在使自然语言中的高层语义概念与三维几何结构的形态、拓扑及空间关系在共享语义空间中得到统一表征。基于此，现有研究在训练策略上主要分为两类：一类是对比式语义对齐，通过InfoNCE或类似的跨模态损失在配对数据上最大化匹配文本-三维样本的相似度，同时最小化不匹配样本的相关性，从而构建全局一致的语义分布；另一类是生成式语义对齐，利用扩散模型或变分自编码器，将文本嵌入作为条件输入引导三维形状生成，使模型在生成阶段隐式完成语义对齐。

在多模态语义对齐研究中，对比式学习机制成为主流范式，其核心思想是在共享语义空间中通过相似度最大化与差异度最小化实现跨模态特征一致性。其关键进展体现在如何构建更有效的三模态（文本-图像-3D）对比学习框架，以及如何处理自然语言中的组合语义。CLIP^2（Zeng等，2023）模型提出通过构建文本-图像-点云三元代理空间，在语义级与实例级施加双重对比约束，有效缓解了跨模态语义对齐中的数据稀缺问题。ULIP模型提出借助CLIP的图文先验，通过少量图文-点三元组实现点云与图像-文本空间的一致性（Xue等，2022）；其后续ULIP-2摒弃人工标注文本，引入视觉语言模型生成全局描述，从而实现端到端三模态预训练（Xue等，2023）。JM3D（Wang等，2023）框架提出通过结构化多模态组织器与联合对齐模块，以层级化方式在点云、图像与文本之间构建统一语义空间，显著增强了三模态融合能力。近期国内研究也探索了基于知识图谱增强的文本-3D对齐方法，通过引入外部常识来理解复杂文本描述中的隐含空间关系（Li等，2024）。Text4Point（Huang等，2023）提出以RGB-D图像为中介建立点图对应关系，利用CLIP提取跨模态特征并通过文本查询模块引导点云表示学习，实现隐式点文映射。SceneForge（Sbrolli等，2025）提出采用几何组合增强策略，在合成多物体场景与文本描述的同时维持几何排列一致性，从而扩充数据样本并提升生成的语义一致性。OpenShape（Liu等，2023）提出整合多个3D数据集与文本描述，构建三模态对比学习框架，使三维特征直接嵌入至CLIP的语义空间，支持零样本3D分类与生成任务。GPT4Point提出采用两阶段语义对齐与生成框架：第一阶段通过点-文对比与Caption生成实现语义嵌入映射，第二阶段利用已对齐特征输入大型语言模型与PointE生成器，实现文本条件下的三维生成推理（Qi等，2024）。CLIP2Point提出针对深度图域的图像-深度联合预训练方案，通过将深度图特征对齐至CLIP视觉嵌入，增强点云语义理解与结构可分辨性（Huang等，2023）。

文本-3D对比式对齐的优势在于能够有效地将离散的语言符号与连续的几何空间进行关联，为3D模型赋予丰富的语义信息。这使得模型能够执行以文本为查询的3D检索、零样本分类等任务，极大地扩展了3D模型的可用性。然而，其局限性在于对齐粒度较为粗糙。自然语言描述常包含组合语义（如“一个带扶手的天蓝色布艺沙发”）和空间关系（如“在桌子下面”），全局对比损失难以精确捕捉和解析这些复杂、细粒度的语义组合，导致生成的3D形状可能在颜色、部件或布局上与文本描述存在偏差。

相较于对比式方法，生成式语义对齐强调在跨模态联合空间中直接学习从文本到三维表示的生成映射。其研究重点在于设计能够充分理解文本语义并转换为三维几何参数的生成架构，如基于Transformer的扩散模型或自回归模型。Text2Shape（Chen等，2018）提出首次尝试将自然语言描述直接映射到三维形状，通过CNN-GRU提取文本语义并结合3D-CNN获取体素特征，设计了往返一致性损失（TST/STS）以增强语义对齐；AutoSDF提出引入P-VQ-VAE将T-SDF离散化为块层次的隐式词元（token），并通过Transformer与乘积分布融合实现文本-形状潜空间对齐，在局部几何与语义约束方面显著优于早期体素方法（Mittal等，2022）；ShapeCrafter（Fu等，2022）提出递归文本条件生成策略，通过短语级描述分解与多对多映射实现可控形状生成。在潜空间扩散建模方向，SDFusion（Cheng等，2023）提出利用BERT与3D VQ-VAE结合交叉注意力实现语义融合；Diffusion-SDF（Li等，2023）提出将TSDF拆分为块并通过UinU-Net扩散生成，显著提升几何细节；3DQD（Li等，2023）提出采用块级VQ-VAE建立统一离散码本并通过交叉注意力注入文本语义，实现轻量高效的三维生成；A3D（Ignatyev等，2024）提出在共享潜在空间中实现多物体连续过渡的语义一致生成；HyperSDFusion（Leng等，2024）提出利用超球面

© 中国图象图形学报版权所有

空间进行层级语义建模以强化语义-几何对齐；Zero3D（Han等，2023）提出通过CLIP预训练模型实现文本、图像与三维形状的语义统一，支持多类别生成；Sketch-and-Text Diffusion（Wu等，2023）提出通过联合草图与文本的扩散过程生成彩色点云，实现形状与外观的协同语义约束；Geometry Image Diffusion（Elizarov等，2024）提出通过几何图像有效表征三维形状，实现高效的文本到三维生成。在国内，研究者针对中文语境和特定物体类别（如中式家具）进行了生成模型的优化与探索（Chen等，2022；Zhou等，2024）。

生成式语义对齐方法的优势在于能够端到端地学习从抽象语言到具体几何的生成过程，从而实现对复杂语义更精细的控制和更高保真度的输出。扩散模型和自回归模型能够建模多模态的条件分布，生成多样化且符合语言描述的形状。其挑战主要在于数据与计算方面。训练此类模型需要大规模高质量的文本-3D配对数据，而此类数据依然稀缺。同时，模型架构复杂，参数量大，训练和推理成本高昂。此外，如何确保生成结果严格遵循文本中的每一个约束（尤其是空间关系），仍是未完全解决的难题。

总体来看，多模态预训练模型推动了三维生成从显式特征对齐向隐式潜空间建模的演进，实现了跨模态统一表征、语义一致生成与泛化理解的闭环。这类模型不仅提升了生成质量、语义连贯性和多视角稳定性，也为开放世界三维理解、零样本生成以及多任务三维推理提供了坚实的语义基础。未来挑战在于实现更高精度和更细粒度的对齐、处理动态与交互语义，以及构建更具可解释性和可控性的对齐机制。

1.3 跨模态3D生成式模型架构

近年来，生成式人工智能的快速发展推动了三维生成领域的显著突破。如图3所示，3D生成所使用的生成式模型大致可分为四类：生成对抗网络（GAN）、变分自编码器（VAE）、扩散模型（diffusion models）与自回归模型（autoregressive models）。它们分别从对抗博弈、概率推断、噪声建模与序列生成等不同范式出发，为三维内容生成提供了多样化的理论基础与实现路径。

1.3.1 生成对抗网络（GAN）

GAN（Goodfellow等，2014）自问世以来，凭借其在图像合成等领域的卓越表现引起了广泛关注。GAN的核心思想源于一种对抗性博弈过程。其框架包含一个生成器和一个判别器：生成器将随机噪声作为输入，旨在生成与真实数据分布难以区分的合成数据；判别器则致力于准确判别输入样本源自真实数据分布还是生成器。在训练过程中，二者通过最小最大化博弈进行联合优化：生成器力求提升生成数据的真实性以迷惑判别器，而判别器则力求增强其鉴别能力。此内在的竞争机制最终驱使生成器产生高度逼真的数据。

随着GAN在二维图像生成领域取得显著成果，研究者们自然将其强大的生成能力拓展至三维空间，主要形成了两条技术路径。第一条路径是直接生成三维表示，如图3中a1所示，该方法紧密遵循经典GAN范式，其关键在于将生成器与多样化的三维数据结构相结合。在这一思路下，生成器直接输出如点云、体素网格、网格模型或符号距离场等显式或隐式的三维表示，并由判别器以真实三维数据为监督进行判断。例如，l-GAN（Achlioptas等，2018）与3D-GAN（Wu等，2016）便是分别生成点云与体素的先驱性工作；而随后的SurfGen（Luo等，2021）等工作则演进为先生成高质量的隐式SDF表示，再转换为显式几何，从而获得了更优的表面质量。然而，此路径的进一步发展受限于高质量三维标注数据的稀缺与生成分辨率的计算瓶颈。

为突破三维数据依赖的限制，如图3中a2所示，第二条路径利用二维监督的3D-Aware GAN应运而生，并迅速成为研究热点。该路径巧妙地利用海量二维图像作为监督信号，通过一个可微分渲染器搭建起三维生成与二维判别之间的桥梁。其基本流程是，生成器首先产生一个三维表示（可为显式或隐式），随后该表示被从特定视角渲染成二维图像，最终判别器通过比较渲染图像与真实图像的真伪来为生成器提供训练梯度。这一范式使得模型能够从无尽的二维图像中逆向学习三维结构知识。早期的HoloGAN（Nguyen等，2019）通过将学得的三维特征投影至二维进行渲染，奠定了基础；而神经辐射场等隐式表示的兴起，则为此路径带来了革命性进展。GRAF（Schwarz等，2020）首次提出了生成式辐射场，pi-GAN（Chan等，2021）通过引入SIREN网络进一步提升了生成质量与视角一致性，GIRAFFE（Niemeyer等，2021）实现了组合式场景建模，直至EG3D

© 中国图象图形学报版权所有

(Chen 等,2022)提出高效且强大的三平面混合表示,将此方向推向了新的高度。

生成式对抗网络为跨模态3D生成提供了一个极具潜力的框架,其优势在于能够灵活适配多种三维表示,并能从易于获取的二维图像中学习,从而生成细节丰富、视觉逼真的三维内容。然而,其应用也面临着固有的挑战,包括训练过程的不稳定性与模式崩塌风险、高昂的计算成本,以及在仅依赖二维监督时可能出现的三维几何合理性等问题。未来的研究有望在提升训练稳定性、探索更高效紧凑的三维表示、以及确保几何精确性等方面继续深化,推动生成式三维人工智能不断向前发展。

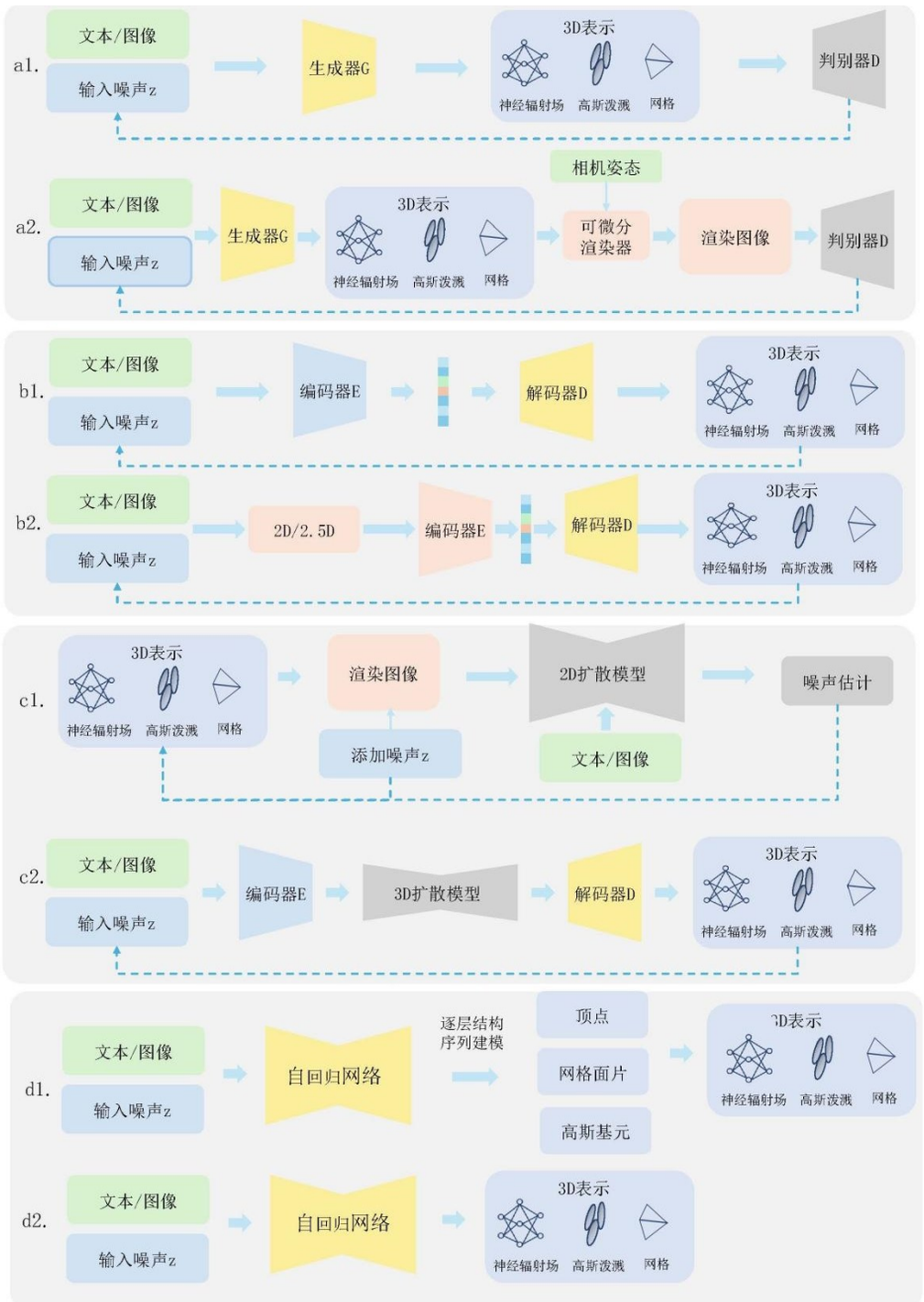

图3 生成式模型架构

Fig. 3 Generative model architecture

© 中国图象图形学报版权所有

1.3.2　变分自编码器(VAE)

在三维生成任务中，VAE将三维对象视为体素、点云或网格集合，并假设潜在向量服从标准正态分布。编码器将输入对象映射到潜在空间以近似其潜在分布，解码器根据潜在向量生成对应的三维结构。训练过程中，通过最大化边际似然的证据下界进行优化：一方面通过重建误差保证生成对象与原始形状一致，另一方面通过KL散度约束潜在分布接近先验，同时通常采用重参数化技巧以保证梯度可传递。三维对象可拆分为多个元素(体素、点或网格部件)，VAE通过潜在向量捕捉全局结构信息，并将对象生成建模为各元素的条件分布，从而实现潜空间到三维对象的映射。结合部件或层次化建模时，可为每个元素引入局部潜在向量，以刻画局部结构，使生成兼顾全局拓扑一致性与局部细节表达，从而提升生成精度与丰富性。在深度学习驱动的三维生成任务中，针对“从二维输入生成三维形状”的问题，现有研究主要形成了两类具有代表性的建模范式。

如图3中b1所示，第一类为直接的二维到三维生成范式，其核心思想是在不依赖显式中间表示的情况下，直接从输入的二维图像中学习到可映射至三维空间的潜在特征表示。此类方法通常以二维卷积神经网络为特征提取器，通过特征嵌入或三维转置卷积结构将二维特征映射到三维形状空间。代表性工作包括将单/多视角图像经二维卷积与三维反卷积生成体素网格的Pix2Vox(Xie等，2019)，通过图卷积网络在模板网格上实现渐进式形变生成三维网格的Pixel2Mesh(Wang等，2018)，以及利用多层感知机预测连续隐式距离场实现形状重建DISN(Xv等，2019)。该范式具备端到端可训练、结构紧凑、生成效率高等优点，但由于直接从二维图像中推断三维几何结构，仍受到视角遮挡、深度歧义及缺乏空间约束等问题的制约，容易导致生成形状的几何失真与拓扑不一致。

如图3中b2所示，第二类为经由中间二维表征的间接生成范式。该范式在二维输入与三维输出之间引入显式的几何中间表征，以缓解深度推断的不确定性并提升几何一致性。具体而言，模型首先从输入图像预测包含三维结构线索的中间二维几何信息(如深度图、法向图、2.5D草图、多视角掩码等)，再将其编码为潜在三维表示用于重建或生成完整的三维形状。代表性工作包括 通过预测2.5D草图并以自编码器生成体素表示的MarrNet(Wu等，2017)，基于图像与渲染图学习二维嵌入特征并生成体素结构的TL-Net(Li等，2020)，以及融合多视角图像特征，通过三维卷积LSTM生成体素形状的3D-R2N2(Choy等，2016)。相比直接生成范式，该类方法的优势在于中间表征提供了显式的几何约束，可有效降低三维生成歧义、提升在少样本与遮挡场景下的鲁棒性；但其局限在于多阶段训练流程较复杂，且中间表征的近似性可能导致细节丢失，从而影响最终的形状精度。

总体而言，这两类范式在生成效率、几何鲁棒性与细节精度之间形成了互补。直接生成范式以端到端结构实现了从图像到三维形状的高效映射，具备较好的计算效率与实时生成能力，适合应用于大规模监督场景及工业级三维重建任务。然而，其依赖二维特征直接推断三维几何，使得在输入视角受限、形状遮挡或复杂拓扑场景中，生成结果往往存在深度歧义与局部几何扭曲等问题。相对应，间接生成范式通过引入中间几何表征，在生成过程中显式注入空间一致性约束，有效提升了模型在少样本、遮挡恢复与结构重建任务中的鲁棒性和可解释性，但也因多阶段映射导致训练复杂度上升与中间信息损失，从而限制了高精度形状建模与局部细节还原的能力。

1.3.3　扩散模型

扩散模型是一类基于概率生成理论的深度生成模型，其核心机制由前向扩散与反向去噪两个阶段构成。在前向过程中，原始数据通过多步高斯噪声扰动逐渐退化为近似各向同性的标准高斯分布。该过程固定且无需学习，旨在将复杂的真实数据分布映射至易于建模的潜在空间。反向过程则通过学习到的神经网络逐步去除噪声、恢复原始数据分布，从而实现由随机噪声生成结构化样本的能力。凭借其在高维复杂分布建模中的稳定性与可控性，扩散模型近年来已成为跨模态3D生成的关键生成范式。

自去噪扩散概率模型(denoising diffusion probabilistic models，DDPM)(Ho等，2020)与改进的DDPM(Nichol等，2021)奠定理论基础以来，扩散模型被迅速引入三维生成领域，并形成了两条主要技术路径：一是基于二维扩散模型的先验迁移，二是基于三维空间的原生扩散建模。

如图3中c1所示，前者主要利用预训练的二维

© 中国图象图形学报版权所有

扩散模型来弥补三维数据稀缺与语义标注不足的问题，成为跨模态3D生成研究的主流方向。典型代表为 DreamFusion（Poole 等，2022），其提出了 SDS 机制，通过以文本条件的二维扩散模型作为外部优化指导器，在可微渲染循环中优化三维表示参数，从而实现从文本直接生成三维形状。此后，大量研究在该框架上进行了改进与扩展。例如，Magic3D（Lin 等，2023）采用粗到细的双阶段生成策略以提升几何与纹理的分辨率；Fantasia3D（Chen 等，2023）与 ProlificDreamer（Wang 等，2023）分别在几何---外观解耦与优化稳定性方面取得突破。后续工作如 Text-Mesh（Hong 等，2023）、Make-It-3D（Tang 等，2023）以及 GaussianDreamer（Liu 等，2024）等进一步拓展了 SDS 的适用范围，将其推广至多种三维表示（如网格、高斯泼溅与隐式场）中，实现了跨表征的一致性生成。这一系列研究表明，复用二维扩散模型的语义先验已成为跨模态3D生成中最具实践价值的范式，可在缺乏大规模三维语义标注的条件下生成高保真且语义一致的三维内容。

如图3中c2所示，另一条研究路径是在三维域中直接建立原生扩散过程，即在点云、体素、三平面或隐式场空间内进行端到端的三维去噪生成。PointE（Nichol 等，2022）与 ShapE（Jun，2023）是代表性工作，分别在点云与隐式函数空间中实现了条件扩散生成，为三维空间的直接采样提供了新的思路。进一步的研究将扩散过程与三维可感知表征相结合，如 EG3D（Chan 等，2022）中的三平面结构被扩展为条件扩散模型，以同时兼顾几何一致性与跨模态可控性。近期研究亦尝试将扩散机制与高效显式表征融合，如基于高斯泼溅的扩散方法（Wu 等，2024）在稀疏场景下实现了高效的连续辐射场生成。这些工作标志着三维原生扩散逐渐从“依赖二维先验”迈向“自主建模三维空间分布”的阶段，为跨模态3D生成提供了更为统一且物理一致的建模框架。

总体来看，扩散模型推动了跨模态3D生成从基于优化的单实例生成向可训练、可控的概率生成框架转变。基于二维先验的方法在语义一致性与纹理细节方面具有显著优势，而原生三维扩散在几何精度与采样效率上更为突出，能够适应多样化场景与复杂空间结构的生成需求。

1.3.4　自回归模型

自回归模型在三维生成中是一类基于条件概率的逐步生成方法，其核心思想是将三维对象表示为按序排列的元素或潜变量序列，并通过学习每个元素在已生成元素条件下的分布来建模整体三维结构。训练阶段通常采用教师强制机制，以真实历史元素指导模型学习条件概率；推理阶段则通过逐步采样生成完整对象，实现对三维结构的可控递归建模。在具体研究中，三维自回归方法可分为两类技术范式。一类是基于结构序列建模的显式几何生成范式，另一类是基于潜空间条件化建模的自回归范式。

如图3中d1所示，在三维生成中，基于结构序列建模的显式几何生成范式将三维对象直接分解为离散元素序列，如点、体素、网格面片或高斯基元，并通过条件概率的链式分解方式逐步生成整个三维结构。该方法能够显式捕捉局部几何依赖关系和拓扑结构的连续性，因此在生成精细结构和保持全局一致性方面表现优异。典型方法包括 PolyGen（Nash 等，2020），其利用 Transformer 对网格顶点和面片序列进行建模，实现可控网格生成；Octree Transformer（Ibing 等，2023）和 OctGPT（Wei 等，2025）则通过层级体素结构进行递归生成，提高生成效率与空间一致性；G3pt（Zhang 等，2024）在点云生成中采用顺序采样机制，有效捕捉局部-全局几何依赖。

与之不同，如图3中d2所示，基于潜空间条件化建模的自回归范式先将三维结构映射至潜空间表示，再在潜向量序列上进行条件概率建模，以捕捉更高层次的语义和全局结构依赖。这类方法能够在生成过程中保持语义一致性，并支持跨模态条件生成，如文本、图像或动作序列驱动的三维建模。代表性工作包括 T2M-GPT（Zhang 等，2023），通过文本嵌入生成连续人体动作帧；HiT-DVAE（Bie 等，2022）和 HuMoR（Rempe 等，2021）通过潜变量与多尺度解码建模动作序列的时间依赖；Uni-3Dar（Lu 等，2025）和 Tar3D（Zhang 等，2025）采用统一的自回归策略提高生成效率与多模态泛化能力；在医学三维生成中，BrainSynth（Tudosiu 等，2024）利用潜空间自回归实现高分辨率解剖结构生成和特定病灶条件建模。

总体来看，自回归建模在三维生成中的特殊性体现在对空间结构和语义条件的精细捕捉能力，其逐步生成策略能够兼顾局部细节与全局一致性，并通过潜空间方法实现语义与形态的紧密耦合。这类方法适用于动作序列、点云、复杂网格、场景布局及

© 中国图象图形学报版权所有

医学影像等多种三维生成任务，支持多模态条件输入与多尺度解码。

1.4 主流3D数据集

高质量三维数据集是跨模态3D生成发展的核心基础设施。受限于采集成本高、标注复杂、格式异构等瓶颈，早期三维数据长期面临规模小、偏差大、模态单一的困境，严重制约生成模型的训练与泛化。近年来，随着众包建模、逆向工程与合成渲染技术的进步，三维数据集逐步从小规模人工标注（如ShapeNet（Chang等，2015））迈向大规模社区共建（如Objaverse（Paszke等，2022））、真实感多模态配对（如MVImgNet（Yu等，2023）、谷歌扫描物体（Google scanned objects，GSO（Downs等，2022））乃至开放世界物理感知（如OmniObject3D（Wu等，2023）），显著推动了生成范式从“优化驱动”向“原生生成”的跃迁。针对生成对象的空间尺度差异，主流数据集可划分为对象级与场景级两类，如表1所示。

1.4.1 对象级数据集

对象级数据主要聚焦于单个物体的几何结构与外观纹理，是物体生成模型的基础。ShapeNet（Chang等，2015）作为结构化语义3D数据的奠基者，包含约5.1万CAD模型，覆盖55个类别，提供部件级分割、对称性与功能标签等细粒度语义。其高精度与拓扑规范性使其成为MeshGPT（Siddiqui等，2024）、TRELLIS（Xiang等，2024）等可控生成方法的验证基准，但其规模有限、工业设计偏态明显，难以支撑大模型训练。为突破规模瓶颈，Objaverse（Paszke等，2022）通过聚合Sketchfab等平台的CC-BY许可模型，构建了超1000万个对象的开放资产库，首次实现亿级三维语义样本覆盖。尽管其噪声较大（含非流形几何与低质拓扑），需依赖自动清洗管线，但其弱文本标注（标题/标签）天然适配文本驱动生成，已成为LRM（Hong等，2023）、DMV3D（Xu等，2023）等端到端模型的预训练基石。随后推出的Objaverse-XL（Deitke等，2023）进一步将规模扩展至1000万以上，并引入更丰富的数据对齐策略，标志着3D生成正式迈入大模型时代。然而，仅靠模型规模无法解决域偏移问题——合成模型与真实世界存在显著外观与几何鸿沟。为此，MVImgNet（Yu等，2023）与GSO（Downs等，2022）应运而生。MVImgNet基于ImageNet类别体系，为120万3D模型渲染12视角真实感图像（含PBR材质与复杂光照），提供严格对齐的“图像-NeRF-相机位姿”三元组，使模型得以从真实世界光照与材质先验中学习；GSO则通过激光扫描获取241个日常物体的毫米级精度几何与HDR纹理，成为评估生成保真度的“黄金真值”，其扫描-真实特性暴露了合成数据训练模型在细节还原上的系统性不足。二者共同将监督信号从“模型相似性”升级为“视觉一致性”，直接推动了Image-to-3D性能的质变。面向更复杂的场景与世界建模需求，OmniObject3D（Wu等，2023）进一步融合激光扫描、无人机摄影与UE5仿真，构建6，000+高质量对象与200+完整场景，提供物理属性（质量/摩擦）、动力学约束（铰链/刚体分组）及多语言描述，首次实现几何—语义—物理的全模态标注。此外，针对语言描述的精细化，Cap3D（Luo等，2024）为Objaverse生成了近百万条高度描述性的3D字幕，有效缓解了文本-3D对齐中的“语义模糊”问题。

1.4.2 场景级数据集

场景级数据集不仅包含多个物体，还涉及复杂的空间布局、光照遮挡及物体间关系，是室内外场景生成与具身智能的关键。早期研究多依赖SUNCG（Song等，2017）和3D-FRONT（Fu等，2021）等合成数据集。其中3D-FRONT包含约6，800个专业设计的室内场景布局，纹理丰富且布局合理，广泛应用于ATISS（Paschalidou等，2021）等场景布局生成任务，但合成数据的非真实感纹理限制了其在照片级渲染生成中的应用。

为捕捉真实世界的复杂性，ScanNet（Dai等，2017）采集了1，513个RGB-D扫描的室内场景，提供密集的语义分割与相机轨迹，成为场景重建与理解的事实标准。针对更高保真度的需求，ScanNet++（Yeshwanth等，2024）进一步利用激光雷达与高分辨率DSLR相机，捕捉了亚毫米级几何细节与反射材质，直接推动了高保真神经渲染与逆向渲染生成的发展。此外，Matterport3D（Chang等，2017）提供了更大规模（90个建筑级场景）的全景深度数据，支持了从房间级到建筑级的生成探索。

在室外与大尺度场景方面，自动驾驶数据集占据主导。Waymo Open Dataset（Sun等，2020）和nuScenes（Caesar等，2020）提供了海量城市街景的LiDAR点云与同步图像，成为StreetGaussians（Yan等，2024）等城市级场景生成模型的核心数据源。近期，DL3DV-10K（Ling等，2024）发布了包含1万个场

© 中国图象图形学报版权所有

景的视频级众包数据，不仅覆盖室内外多样化环境，还针对视图合成进行了深度优化，为通用场景生成模型（large scene models，LSM）的训练提供了前所未有的数据规模支持。

总体而言，3D数据集正经历从单体对象向复杂场景，从几何形状向物理外观，从单一模态向语言-视觉-3D统一的演进，这一进程为跨模态3D生成模型提供了从可学习到可泛化的基础条件。随着数据规模、质量与模态不断提升，未来的跨模态3D生成研究将更多依赖多源数据的联合建模与自监督学习，为构建统一的世界级三维生成模型奠定基础。

表1　主流3D数据集汇总

Table 1　Overview of Mainstream 3D Datasets

类型	数据集名称	发布年份	发布地址	数据规模	数据来源	主体表示
Object	ShapeNet	2015	https://shapenet.org	~5.1 万模型	合成	网格
	Objaverse (-XL)	2022/23	https://objaverse.allenai.org	~1000 万+模型	众包	网格
	MVImgNet	2023	https://gaplab.cuhk.edu.cn/projects/MVImgNet	120 万模型/650 万图	合成渲染	多视图图像
	GSO	2022	https://goo.gle/scanned-objects	1,030+模型	真实扫描	网格
	OmniObject3D	2023	https://omniobject3d.github.io	6,000+模型	扫描+仿真	网格/点云
Scene	3D-FRONT	2021	https://huggingface.co/datasets/huanngzh/3D-Front	~6,800 场景	合成（设计）	布局/网格
	ScanNet (++)	2017/24	http://www.scan-net.org	1,500+ / 460 场景	真实扫描	图像 /重建网格
	Matterport3D	2017	https://niessner.github.io/Matterport	90 建筑/10,800 全景	真实扫描	多视图图像/ 网格/ 全景图
	Waymo Open	2020	https://waymo.com/open	1,150+路段	真实采集	激光雷达点云
	DL3DV-10K	2024	https://dl3dv10k.github.io/DL3DV-10K	10,000+视频	众包视频	多视图图像

2　文本驱动的三维对象生成

三维模型的生成长期以来依赖专业的几何建模与渲染流程，而文本到3D生成的兴起，使自然语言成为三维内容建模的重要接口。该方向的研究目标是：通过输入文本描述，自动生成具有结构合理、语义一致和视觉真实感的三维对象。其发展历程可大致分为三个阶段：基于CLIP的优化方法、基于扩散模型的SDS框架，以及基于3D原生数据的直接生成方法。

2.1　基于CLIP的优化方法

早期研究借鉴了CLIP在图像-文本对齐上的成功经验，通过优化三维表示（如体素、网格、隐式神经辐射场NeRF），使从不同视角渲染的图像与输入文本在CLIP嵌入空间中相似。代表性方法包括CLIP-Forge、CLIP-Sculptor、CLIP-NeRF、CLIP-Mesh 、Text2Mesh、DreamFields和TANGO等（Li等，2023）。这些方法不依赖文本-3D配对数据，只需文本提示即可驱动三维形状优化，从而在零样本条件下实现语义一致的3D建模。

其中，DreamFields（Jain等，2022）是极具代表性的开创性工作。如图4所示，DreamFields首次将NeRF与CLIP相结合，实现了从纯文本提示中零样

© 中国图象图形学报版权所有

本生成3D对象。它通过优化一个NeRF，使得从多个随机视角渲染出的2D图像在CLIP嵌入空间中与目标文本高度相似。为了解决单纯CLIP监督下NeRF容易产生“漂浮物”、近场伪影和几何不连贯等问题，DreamFields引入了多项关键的几何先验：包括鼓励场景稀疏性的透射率正则化、用于约束物体位置的场景边界以及改进的MLP网络架构。这些设计共同确保了生成的3D对象具有合理的几何结构和多视角一致性。由于CLIP对空间结构与视角一致性的理解有限，这类方法常出现几何不稳定（如“Janus效应”）或纹理模糊问题。同时，整个优化过程依赖于耗时的梯度下降迭代，计算代价高、生成速度慢。

尽管如此，这一阶段的研究奠定了跨模态语义对齐的基础，成功验证了“语言控制三维形状”的核心理念，也为后续基于扩散模型的高效3D生成研究铺平了道路。

2.2 基于扩散模型的优化方法

随着文本到图像扩散模型（如Imagen和Stable Diffusion）在2D生成任务中展现出极强的细节刻画能力与语义一致性，研究者开始尝试将这种“2D生成先验”迁移到3D领域。DreamFusion（Poole等，2023）在这一方向上提出了具有里程碑意义的分数蒸馏采样框架。如图5所示，SDS的核心理念是：它并不直接训练一个3D生成模型，而是借用一个已经训练好的、固定的2D文本到图像扩散模型作为老师。在生成过程中，系统会不断从当前3D模型的

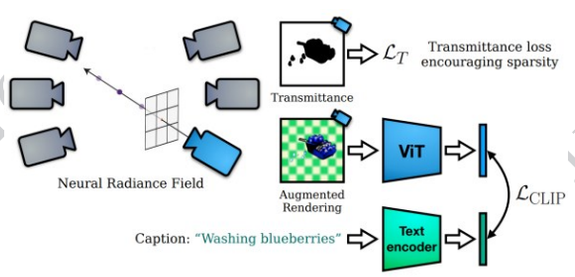

图4 DreamFields方法流程图（Jain等，2022）

Fig. 4 The pipeline of DreamFields (Jain et al, 2022)

多个随机视角渲染出2D图像，然后将这些图像输入到2D扩散模型中，询问“这张图在多大程度上符合给定的文本描述”。2D模型会返回一个优化信号，告诉3D模型应该如何调整其几何和外观，才能让渲染出的图像更贴近文本语义。通过反复执行这一“渲染---评估---更新”的循环，3D模型逐渐被蒸馏出与文本高度一致的形状和纹理。

相较于早期的CLIP优化方法，SDS有两个显著优势：一是生成先验更强。扩散模型直接掌握了真实图像分布的统计规律，而不是仅依赖语义相似度；二是优化信号更稳定。通过蒸馏扩散模型的“去噪方向”，3D表示可以获得细粒度的形状与纹理指导，从而生成具有高保真度和多视角一致性的三维结果。因此，DreamFusion的提出标志着文本到三维生成从判别式相似度约束转向生成式知识蒸馏的新阶段。然而，SDS在实际应用中仍存在若干问题。由于扩散模型本身是二维的，它在不同视角下的指导信号并不完全一致，容易导致生成几何出现不稳定或不连续的现象，例如“Janus效应”——物体的每个角度都像是正面。此外，SDS还容易出现纹理过饱和、表面过度光滑或几何细节丢失的问题，且由于优化过程依然是逐步渲染与更新，生成效率较低、可控性不足。为解决这些问题，后续研究主要沿着三个关键路径展开：引入分阶段的精炼策略、改进蒸馏损失函数本身，以及开发多视角一致的2D先验。

1）分阶段与粗到精的生成策略。在提升生成效率和分辨率方面，“分阶段与粗到精”的生成策略被证明是行之有效的（Liu等，2024）。代表性工作Magic3D（Lin等，2023）针对DreamFusion优化慢和分辨率低的问题，提出了一个两阶段框架：第一阶段使用稀疏NeRF与低分辨率扩散先验快速生成粗糙几何；第二阶段则将该表示转换为高分辨率的纹理网格，并利用高分辨率潜空间扩散模型进行精细优化。DreamMesh（Yang等，2024）同样采用了粗到精的流程，但其从始至终专注于显式的三角网格，其粗糙阶段使用文本引导的雅可比矩阵来变形网格，精细阶段则联合优化网格几何与纹理贴图。Fantasia3D（Chen等，2023）引入了更深刻的解耦思想。它将几何与外观分开建模，通过显式表面表示（如DMTet）并利用提取的表面法线来指导几何生成；同时，它将基于物理渲染的材质模型（如BRDF）引入Text-to-3D任务中以学习外观，这使得生成结果具有高真实感，并原生支持重新打光。DreamCraft3D（Sun等，2023）则将此过程进一步模块化为几何塑形与纹理增强两个环节。在粗糙3D模型的渲染图上训练个性化扩散模型DreamBooth（Ruiz等，2023），使其具备场景的3D感知能力，再反过来用这个3D感知的先验指导

© 中国图象图形学报版权所有

精细纹理优化，实现了高效且一致的细节增强。

2）改进的蒸馏损失与优化目标。研究者发现原始SDS梯度存在过强的收敛偏置，易导致过饱和、过平滑或多样性不足的问题。分类分数蒸馏（Yu等，2024）重新评估了无分类器指导的作用，并发现仅使用指导部分（即条件分数与无条件分数的差值）就足以进行有效的3D生成，这种更简洁的梯度被证明比完整SDS更稳定。为了解决SDS中因伪基准真相（pseudo-GTs）不一致而导致的“平均效应”，区间分数匹配（Interval Score Matching）（Liang等，2024）建议采用确定性的DDIM扩散轨迹，并在轨迹的区间而非单步之间进行匹配，以减少累积偏差。为了从根本上解决多样性问题，Variational Score Distillation（Wang等，2024）不再将3D模型参数视为一个固定值，而是将其建模为一个分布，并使用基于粒子的变分推断框架进行优化，显著提升了生成的多样性与质量。此外，Asynchronous Score Distillation（Ma等，2024）观察到扩散模型在早期时间步（噪声较多时）的预测误差更低，因此提出异步地调整蒸馏所处的时间步。这种方法无需修改预训练的2D模型，有效减少了噪声积累，显著提高了训练稳定性。

3）多视角一致的扩散先验。原始SDS依赖的单视角2D先验缺乏3D空间知识，导致了严重的“Janus效应”。多视角一致的扩散先验是解决此问题的关键。MVDream（Shi等，2024）是这一方向的开创性工作。该模型通过在3D资产渲染出的多视图数据集上进行训练，使其能够根据输入的相机姿态生成一组几何和纹理上保持一致的多视角图像。当MVDream被用作3D优化的先验时，它提供了强大的空间约束，极大地缓解了“Janus效应”，提升

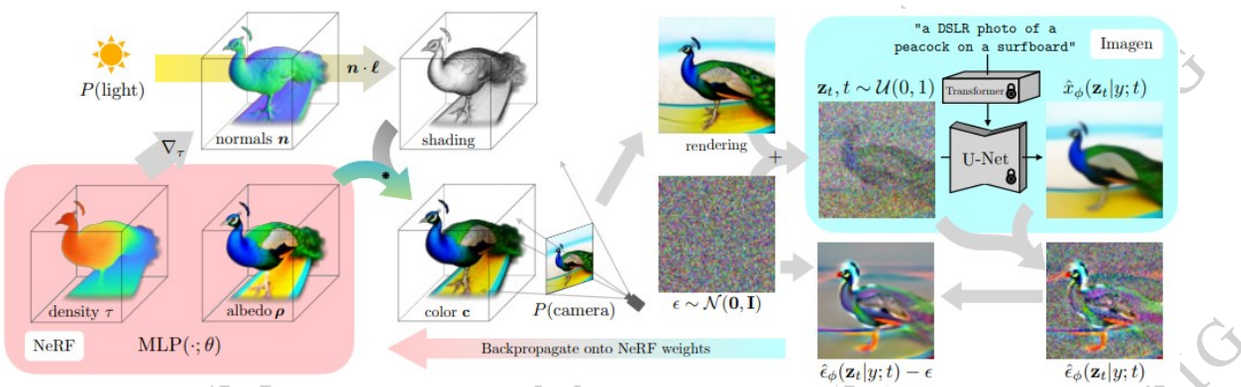

图5　DreamFusion方法流程图（Poole等，2022）

Fig. 5　The pipeline of DreamFusion (Poole et al., 2022)

了生成结果的整体三维一致性。

总体而言，SDS方法的提出使跨模态3D生成进入了一个“借助强大2D模型进行知识蒸馏”的新时代。它不仅显著提升了文本驱动三维生成的视觉质量和语义一致性，也奠定了后续研究（如Magic3D、Fantasia3D、MVDream等）的理论与技术基础。SDS及其改进变体的不断演化，使得基于扩散模型的三维内容生成逐渐走向高分辨率、高真实感与高一致性的方向。

2.3　基于3D原生数据的生成模型

随着ShapeNet、Objaverse等大规模三维数据集的建立，研究者开始探索直接在3D原生空间中训练生成模型，从而摆脱依赖2D扩散模型蒸馏的限制。这类方法通常利用已有的3D标注数据（如点云、网格、体素或神经辐射场等）进行端到端训练，能够实现快速的前馈式生成，在推理阶段直接输出三维形状或场景表示，而无需再通过渲染优化循环。根据采用的生成范式不同，主要可分为基于GAN、扩散模型与自回归模型的三类。

GAN最早在2D图像生成中展现出卓越的建模能力，随后被引入3D领域，用以直接学习三维形状分布（Hu等，2025）。早期研究如3D-GAN（Wu等，2016）与PointGAN（Achlioptas等，2018）分别在体素与点云空间中实现了从潜向量到3D形状的映射，验证了GAN在3D空间建模的可行性。后续的Mesh-GAN（Cheng等，2019）与Tree-GAN（Shu等，2019）进一步拓展到显式网格与层次化点云结构，使网络能够生成更精细的几何细节。与此同时，研究者提出了3D-Aware GANs框架，利用可微渲染器将隐式3D表示投影到2D图像空间，再通过2D判别器进行对

© 中国图象图形学报版权所有

抗学习。代表性工作包括 HoloGAN（Nguyen-Phuoc 等，2019）、BlockGAN（Nguyen-Phuoc 等，2020）与 EG3D（Chan 等，2022）。其中，EG3D 提出了显式-隐式结合的三平面表示，使生成器在保持高效的同时能表达复杂的几何与纹理关系，成为随后众多3D生成工作的核心结构。该系列方法的优势在于高效与可控，可通过潜空间插值与相机姿态控制实现多视角一致的图像与几何生成。

扩散模型在2D图像生成领域的成功也启发了研究者在3D空间中建立端到端的去噪生成过程。这类方法直接在3D表示（如点云、网格、神经距离场或神经辐射场）上执行正向加噪与反向去噪，从而学习三维形状的生成分布。典型工作包括 PVD（Zhou 等，2021）、Diffusion-SDF（Cho 等，2023）、MeshDiffusion（Liu 等，2023）、Point-E（Nichol 等，2022）以及 Shap-E（Jun 等，2023）等。其中，Point-E 是 OpenAI 提出的大规模文本到点云的扩散系统。它采用两阶段流程：首先利用文本到图像扩散模型 GLIDE（Nichol 等，2021）生成一个合成视图；随后，以此图像为条件，使用第二个扩散模型生成彩色点云。相比传统优化型方法，Point-E 能在单张 GPU 上用1到2分钟生成一个三维样本，速度提升了一个数量级。虽然在精细度上略低于 DreamFusion 等蒸馏方法，但在效率与多样性上具有显著优势，为快速3D原生生成奠定了基础。随后，Shap-E（Jun 等，2023）进一步扩展了生成目标，不再局限于点云，而是直接生成可渲染为“带纹理网格”与“神经辐射场”的隐式函数参数。它首先训练一个编码器，将3D资产映射为隐式函数参数；再基于这些参数训练条件扩散模型，实现多表示统一建模。Shap-E 在效率与表现力上均优于 Point-E，能够以更少的采样步数生成复杂的三维几何与纹理。与前者相比，Shap-E 不仅在隐式空间内实现了更高维的连续建模，还展示了单一模型统一输出 NeRF 与网格两种形式的

潜力。

受大型语言模型成功的启发，一些工作将3D资产离散化为一系列词元，将3D生成问题转化为序列建模任务。ShapeGPT（Yin 等，2025）首先通过 3D VQ-VAE 将形状编码为离散词元序列，然后利用类似 GPT 的 Transformer 模型，根据文本提示自回归地生成这些词元，从而实现文本驱动的三维形状生成。这种方式不仅能生成几何形状，还能借助大语言模

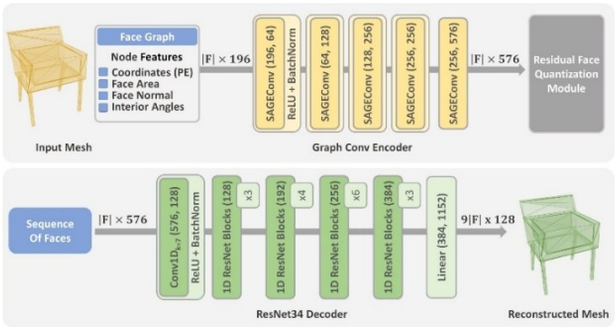

图6　MeshGPT方法流程图（Siddiqui 等，2024）

Fig. 6　The pipeline of MeshGPT (Siddiqui et al, 2024)

型的推理能力完成3D内容编辑与语义理解任务。最新的 MeshGPT（Siddiqui 等，2024）则进一步推动了该方向。它提出一种基于 Decoder-Transformer 的三角网格生成框架：首先使用图卷积网络学习局部几何与拓扑的量化嵌入词表，然后将这些嵌入序列输入解码器进行自回归预测，逐步生成完整的三角面片。该方法展示了自回归 Transformer 在3D内容生成中的强大潜力，为三维生成迈向语言化建模与结构化生成提供了新的思路。

总体来看，基于3D原生数据的生成模型代表了跨模态3D生成从“2D知识蒸馏”到“3D直接建模”的关键转变。它们摆脱了2D教师模型的限制，能够直接在3D分布上学习结构、几何与外观，从而在生成速度、结构一致性与物理可解释性上取得显著提升。未来的发展方向包括：构建更大规模、更均衡的3D数据集；探索统一的多模态-多表示生成框架；以及将几何先验、物理约束和语义理解进一步融入3D生成过程，为通用三维基础模型的构建奠定基础。

3　图像驱动的三维对象生成

与文本到3D（Text-to-3D）从抽象语义生成全新对象不同，图像到3D（Image-to-3D）旨在从一张或少量给定的参考图像出发，重建或生成其对应的三维模型。单视图3D生成尤其具有挑战性，因为它需要模型具备强大的“脑补”能力，从有限的2D信息中推断出完整的3D结构和外观，特别是那些在输入视角中被遮挡或不可见的部分。与 Text-to-3D 的发展路径相似，Image-to-3D 的研究也经历了从依赖2D先验优化到3D原生直接生成的演进。

© 中国图象图形学报版权所有

3.1 基于2D扩散先验的优化方法

与Text-to-3D方法类似，早期及主流的Image-to-3D方法大量借鉴了预训练2D扩散模型的强大生成先验，通过分数蒸馏采样及其变体来优化三维表示（如神经辐射场NeRF或3D高斯泼溅）。在这一框架下，输入的参考图像被用于约束生成内容的核心外观、身份和结构。这一路径的早期探索包括NeuralLift-360（Xu等，2023）、RealFusion（Melas-Kyriazi等，2023）和NeRDi（Deng等，2023）等工作。它们通常首先对输入图像进行文本反演，获得一个代表该图像核心语义的特殊文本嵌入，然后将这个嵌入和输入图像作为SDS优化的强条件，辅以单目深度估计、法线图等几何先验，共同指导一个神经辐射场NeRF的优化过程，从而将单张"野外"图像提升为一个完整的360度3D对象。

然而，通用的2D扩散模型（如Stable Diffusion）缺乏对3D空间一致性的内在理解，直接将其用于SDS优化容易导致严重的“Janus效应”和几何失真。为了解决这一根本性问题，研究界开发了专门用于新视角合成的扩散模型，其中Zero-1-to-3（Liu等，2023）是具有里程碑意义的工作，其核心创新在于构建了一个视角条件化的扩散模型。该模型在一个大规模的合成3D物体数据集Objaverse上进行训练，学习如何根据输入的单张RGB图像和指定的相对相机姿态，生成该物体在新视角下的图像。其关键洞察是：尽管训练数据是合成的，但通过在互联网规模的图像数据上进行预训练，扩散模型已经内化了关于物体几何和外观的强大先验知识。因此，Zero-1-to-3展现出了卓越的零样本泛化能力，能够成功处理分布外的数据集，甚至是风格迥异的"野外"图像，如印象派绘画。这一能力使其成为Image-to-3D任务中一个极其强大且可靠的3D感知先验。

Zero-1-to-3的出现极大地推动了Image-to-3D的发展，它很快被用作一种强大的“3D感知先验”集成到优化流程中，并催生了两种主流的改进范式：

1）基于“伪图像”的监督：以One-2-3-45（Liu等，2023）为代表的方法，利用Zero-1-to-3首先从输入图像生成一组几何一致的多视角“伪图像”。然后，它们不再使用不稳定的SDS损失，而是转而使用更传统的重建损失（如L2损失或LPIPS损失），迫使3D表示（如NeRF）在渲染到相应视角时，能与这些“伪图像”相匹配。

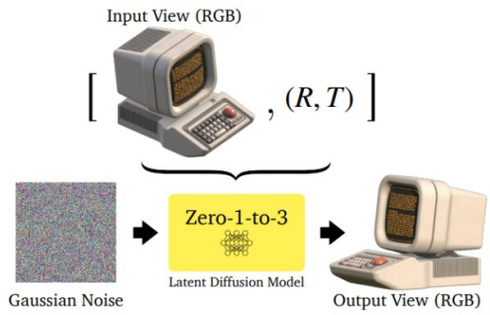

图7 Zero-1-to-3方法流程图（Liu等，2023）

Fig. 7 The pipeline of Zero-1-to-3 (Liu et al, 2023)

2）基于“混合先验”的蒸馏：以Magic123（Qian等，2023）为代表的方法，提出了一种更精细的双重先验优化策略。Magic123团队发现，单独使用Zero-1-to-3作为3D先验，虽然能保证良好的一致性，但由于其在合成数据上训练，生成的纹理细节和真实感有时会弱于通用的2D先验。因此，Magic123同时使用了两个“教师”：3D先验（Zero-1-to-3）：提供视角一致的几何指导，作为主要的形状约束；2D先验（Stable Diffusion）：通过文本反演，提供高频纹理细节和真实感。该方法引入了一个权衡参数，在优化过程中巧妙地融合了两种先验的SDS梯度，实现了3D一致性与纹理真实感的兼得。

在优化效率方面，上述依赖神经辐射场NeRF表示的方法通常需要较长的优化时间。为了解决NeRF优化速度慢的问题，DreamGaussian（Tang等，2024）提出将3D高斯泼溅（3DGS）作为新的三维表示，结合SDS优化框架实现快速高效的3D生成。3DGS将场景或物体表示为一组可学习的三维高斯核，每个高斯具有位置、尺度、颜色和透明度等参数，能够通过高效的可微分渲染器直接生成视图图像。与传统NeRF相比，3DGS的显著优势在于其渲染复杂度低、收敛速度快且内存占用小。DreamGaussian进一步引入密度渐增策略和两阶段优化流程：第一阶段通过稀疏高斯分布快速拟合整体几何结构；第二阶段将结果转换为高分辨率纹理网格并在UV空间进行精细纹理优化。该方法在单张图像输入条件下即可在数分钟内生成高质量的三维模型，效率较基于神经辐射场NeRF的优化方法提升一个数量级。

3.2 基于多视图一致性增强的方法

基于优化的Image-to-3D方法虽能生成细节丰富的3D资产，但其迭代过程通常耗时较长，并且严

© 中国图象图形学报版权所有

重依赖于预训练2D扩散模型的稳定性与泛化能力。

为克服这些瓶颈，研究者提出了基于多视图一

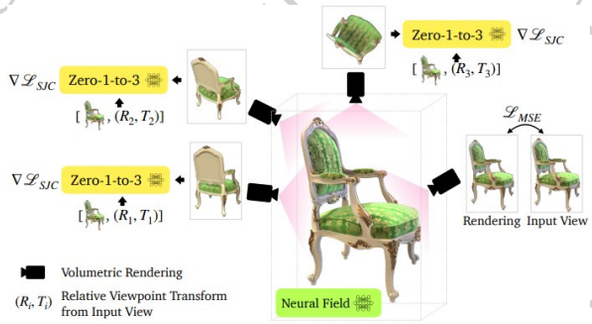

图8　利用Zero-1-to-3重建3D（Liu等，2023）

Fig. 8　Using Zero-1-to-3 to reconstruct 3D (Liu et al, 2023)

致性增强的生成范式。该类方法通常采用两阶段流程：首先利用专门设计的生成模型，从单张输入图像推理出一组几何与纹理一致的多视角图像；随后，再借助传统的结构重建方法（如结构光恢复或多视图立体）将这些多视图图像转换为高质量的三维网格。此类方法的核心在于第一阶段生成的多视角图像必须具有强烈的三维一致性，以确保后续重建过程能够得到稳定的几何与纹理映射。根据核心技术路线的不同，这一方向的研究主要沿着两条路径演化：（1）基于多视图扩散模型的方法，以及（2）基于视频扩散模型的方法。

1）基于多视图扩散模型的方法。这类方法通过改进扩散模型的结构设计与训练机制，使其在生成阶段即可显式地建模不同视角间的几何对应与外观一致性。其核心思想是利用多视图特征共享与相机位姿条件化机制，使模型在一次扩散过程中同时预测多个视角下的图像，实现跨视角的同步一致生成（Wu等，2025）。Zero123++（Shi等，2023）是继Zero-1-to-3之后的重要里程碑。该模型以Stable Diffusion为基础，通过联合生成六个固定相机姿态下的图像，实现多视角分布的联合建模，而非独立预测各个视角。为了消除物体朝向模糊问题，Zero123++使用固定绝对俯仰角与相对方位角组合的相机布局，从而确保所有视图在统一姿态下生成。此外，它改进了噪声调度策略与局部条件机制，有效强化了视图间的全局一致性与细节对齐。SyncDreamer（Liu等，2023）进一步从生成机制层面解决视角解耦问题，提出了同步多视图扩散框架。该方法在反向扩散的每一步同时生成所有目标视图，并通过3D感知的特征注意力模块在不同视角之间进行状态同步，从而显式地建模多视图间的联合分布。SyncDreamer的关键在于3D-aware feature attention：在去噪过程中，模型在三维体素空间内对齐各视角的特征，再通过深度方向的注意力机制维

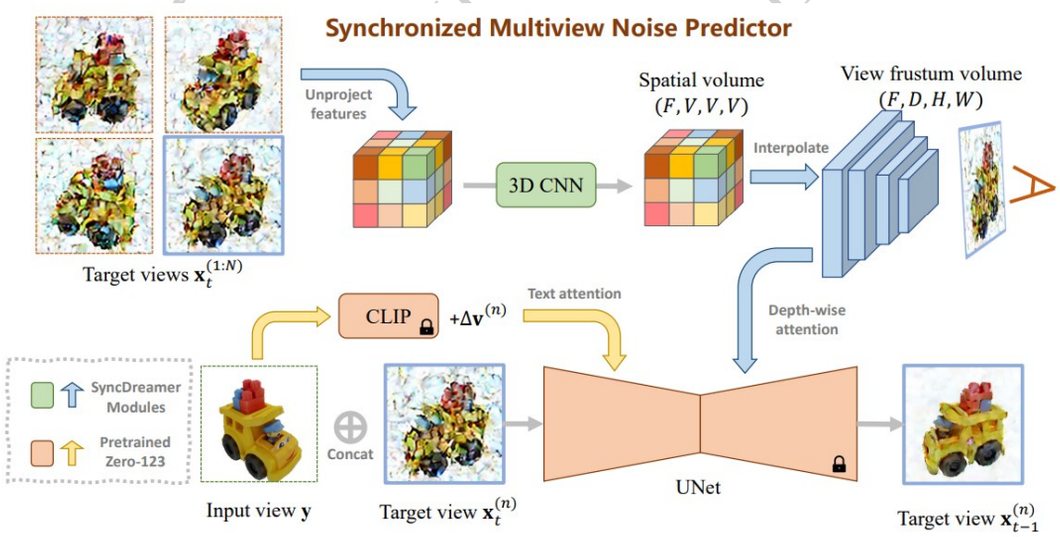

图9　SyncDreamer方法流程图（Liu等，2023）

Fig. 9　The pipeline of SyncDreamer (Liu et al, 2023)

持跨视角的几何对应关系。这使得模型能够直接生成高度一致的多视图图像，进而无需SDS优化即可使用NeRF或NeuS（Wang等，2021）等方法直接重建高质量网格。Wonder3D（Long等，2024）则提

© 中国图象图形学报版权所有

出了跨域扩散模型，同时生成多视图的彩色图像与法线图。其核心创新是引入跨域注意力与域切换器机制，使模型在"颜色域"和"法线域"之间共享特征并互相约束，从而保证几何与纹理的一致性。Wonder3D还设计了几何感知的法线融合算法，能够在稀疏多视图条件下恢复平滑且高保真的3D网格表面。与SyncDreamer相比，Wonder3D显著提升了几何细节恢复能力与重建效率，仅需数分钟即可生成高质量纹理模型。

2）基于视频扩散模型的方法。另一条极具前景的路径是重塑预训练的视频扩散模型（video diffusion models，VDM）。研究者发现，VDM（如stable video diffusion，SVD）在训练中为确保视频帧的时间一致性所学习到的强大先验，可以被巧妙地"转译"为3D空间一致性。其核心思想是：将环绕物体的相机轨道序列视为一个"视频"，从而将VDM的时间先验迁移到3D视角生成任务中。Stable Video 3D（SV3D）（Voleti等，2024）是这一方向的标志性工作。它并非简单生成固定轨道，而是通过微调SVD，使其能够接受显式的相机姿态轨迹作为条件，生成高分辨率的轨道视频。其核心机制是将相机轨迹（高程和方位角）的正弦嵌入与噪声时间步的嵌入相加，然后注入到UNet的残差块中，从而实现对视角的精确控制。SV3D的应用具有双重性：一方面，它可以作为优化方法的直接升级，充当一个强大的3D感知视频先验，通过"掩码得分蒸馏采样（Masked SDS）"损失来指导NeRF和DMTet网格的由粗到精优化。另一方面，它也可以作为"生成-再重建"模型，直接利用其生成的NVS结果进行3D重建。此外，为解决轨道视频首尾（即0°和360°）的衔接问题，SV3D还提出了"三角波分类器无关指导"策略，取代了SVD原本的线性缩放，显著提升了环绕一致性。Hi3D（Yang等，2024）则专注于解决超高分辨率的纹理生成问题。该工作认为，由于训练稳定性的限制，现有方法难以生成高分辨率的图像。为此，Hi3D提出了一个两阶段的级联范式：基础多视图生成：首先，利用一个微调的VDM（注入相机高程作为条件），从单张图像生成一组中等分辨率（如512x512）的3D感知序列（轨道视频）。3D感知多视图精炼：随后，设计一个3D感知的视频到视频精炼器。该精炼器（同样基于VDM）以低分辨率视频、原始输入图像、相机高程以及估计的深度图序列作为多重条件，将视频上采样至1024x1024的分辨率，并显著增强3D细节和一致性。Hi3D是一种纯粹的"生成-再重建"方法，其最终重建流程还结合了3DGS（用于视图增强）和SDF（用于网格提取），以实现高保真网格。

V3D（Chen等，2024）同样将多视图生成视为视频生成，通过在360度轨道视频上微调SVD，从单张正面视图中生成多帧（如18帧）轨道视频。该工作敏锐地指出，VDM的输出帧之间仍存在轻微的不一致，如果直接使用像素级损失进行重建，会导致模糊或伪影。为此，V3D的核心贡献在于其定制的

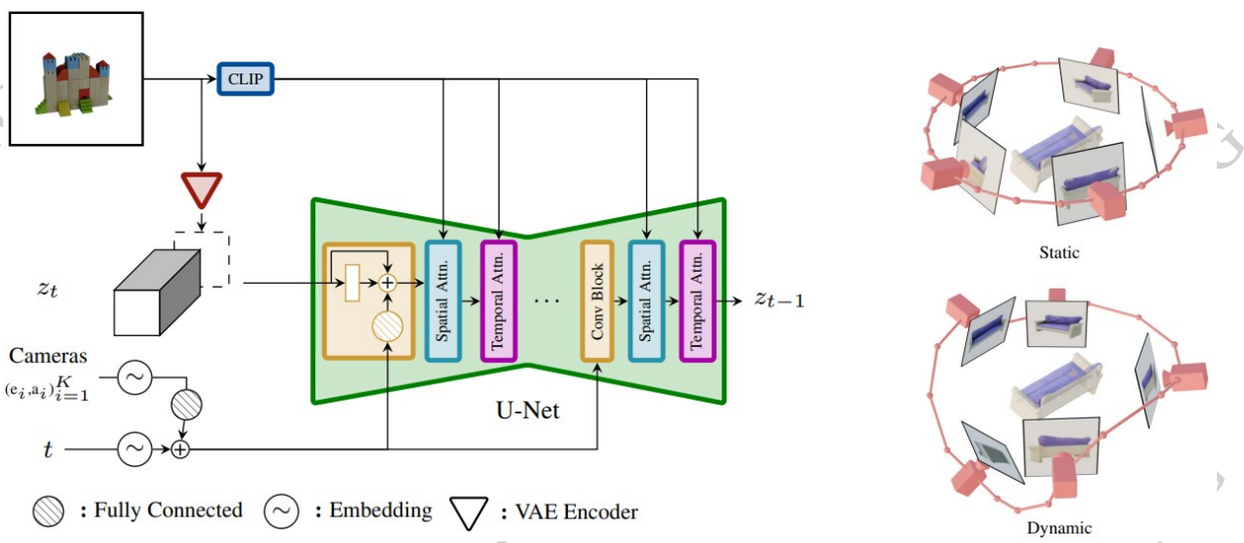

图10　SV3D方法流程图（Voleti等，2024）

Fig. 10　The pipeline of SV3D (Voleti et al, 2024)

© 中国图象图形学报版权所有

重建流程，以增强对不一致性的鲁棒性：它采用3DGS作为表示，但在优化时使用感知损失来替代传统的像素损失。同时，为加速3DGS收敛，它提出了一种空间雕刻的初始化方法，利用多视图的2D掩码来大致定位高斯核，避免了在空白区域的无效优化。V3D还通过集成PixelNeRF（Yu等，2021）编码器，将此框架扩展到能处理稀疏视图输入的场景级新视角合成。

3.3 基于3D原生数据的直接生成方法

尽管基于2D扩散模型的优化方法在质量上不断提升，基于多视图一致性增强的方法中的"生成-再重建"方法在效率上取得了突破，但这两种路径都存在各自的局限性。前者（如SDS优化）速度缓慢且逐例优化过程不稳定；后者（如Wonder3D）虽然实现了前馈式生成，但两阶段流程中，重建步骤的质量完全依赖于第一阶段生成的多视图图像的一致性，任何视图间的微小不一致都可能导致重建伪影和误差累积。为彻底解决这些问题，研究界探索了第三条路径：构建端到端的直接生成模型。这类方法的目标是训练一个单一的、前馈式的网络，使其能够直接从输入的单张图像映射到一个显式或隐式的三维表示，从而在一次推理中输出最终的3D模型，彻底摒弃耗时的迭代优化或两阶段重建。早期的通用3D生成模型（He等，2022），如OpenAI的Point-E和Shap-E，已初步验证了这一路线的可行性。它们不仅支持文本输入，也能接受图像作为条件，直接生成点云或隐式函数。然而，受限于当时的数据规模和模型架构，其生成结果在几何细节和纹理真实感方面仍有较大提升空间。近期，随着大规模3D数据集（如Objaverse）的涌现和Transformer等强大架构的引入，直接生成方法迎来了爆发式发展，主要形成了两大技术范式。

1）大型重建模型范式。该范式的核心思想是借鉴"大数据+大模型"的成功经验，通过在百万级3D数据集上训练高容量模型，学习强大的通用3D先验知识。LRM（large reconstruction model）（Hong等，2023）是这一范式的奠基性工作。它首次提出利用大规模Transformer编码器-解码器结构（约5亿参数）从单张图像直接回归三平面隐式场。LRM采用预训练的视觉Transformer（DINO）提取2D图像特征，并通过一个图像到三平面解码器利用交叉注意力机制将图像特征"投影"到三维潜空间，再由共享MLP完成体渲染。模型在约百万规模的Objaverse与MVImgNet（Yu等，2023）数据上进行端到端训练，仅依赖简单的图像重建损失即可实现高质量三维重建。在单张A100上，LRM可在约5秒内从单图生成可交互的三维网格，具有出色的速度与泛化能力。相比依赖SDS的优化方法，LRM不需要预训练2D扩散模型的指导信号，而是通过大规模数据和强表达能力直接学习通用的3D先验。Instant3D（Li等，2023）继承并扩展了LRM的思想，提出了一个两阶段的前馈式Text-to-3D框架。其第一阶段在微调的SDXL（Podell等，2023）上进行结构化稀疏视图生成，一次性产生四个几何一致的视图（2×2网格）。第二阶段则采用基于Transformer的稀视图重建器，将这些多视图图像直接回归为三平面NeRF。Instant3D的重建器架构与LRM类似，但在编码端加入相机姿态调制，使每个图像词元具备视角感知能

力。该方法在不依赖SDS或多阶段优化的情况下，实现了仅需20秒的高质量3D生成，在保持视觉保真度的同时，推理速度比传统优化方法快约200倍。DMV3D（Xu等，2023）则进一步将LRM的重建能力嵌入到扩散模型的去噪器中，提出了基于重建的多视图去噪机制。在扩散过程的每一步中，模型利用一个LRM变体从带噪声的多视图图像重建出干净的三平面NeRF，再通过体渲染生成去噪后的图像。这样的设计使DMV3D成为首个单阶段的3D扩散模型，能够直接生成NeRF表示，无需依赖SDS优化或显式两阶段重建。得益于扩散框架的概率性建模，DMV3D不仅支持单图和文本条件生成，还能在30秒内生成具有多样性和高保真的3D资产。

2）基于VAE和潜空间扩散/流模型的方法。另一条重要的技术路线是在压缩的3D潜空间中进行生成，类似于隐式扩散模型在图像领域的成功。这类方法通常包含两个核心组件：一个高效的3D变分自编码器用于将3D形状压缩/解压缩为潜码，以及一个在潜空间中操作的生成模型（通常是扩散Transformer、DiT或流匹配模型）。

CLAY（Zhang等，2024）是这一方向的大规模探索。它采用基于3DShape2VecSet（Zhang等，2023）的多分辨率变分自编码器架构，以点云为输入、占用场为输出，实现连续表面的高效编码与解码；同时在潜空间中引入极简的扩散Transformer（DiT）作为生

© 中国图象图形学报版权所有

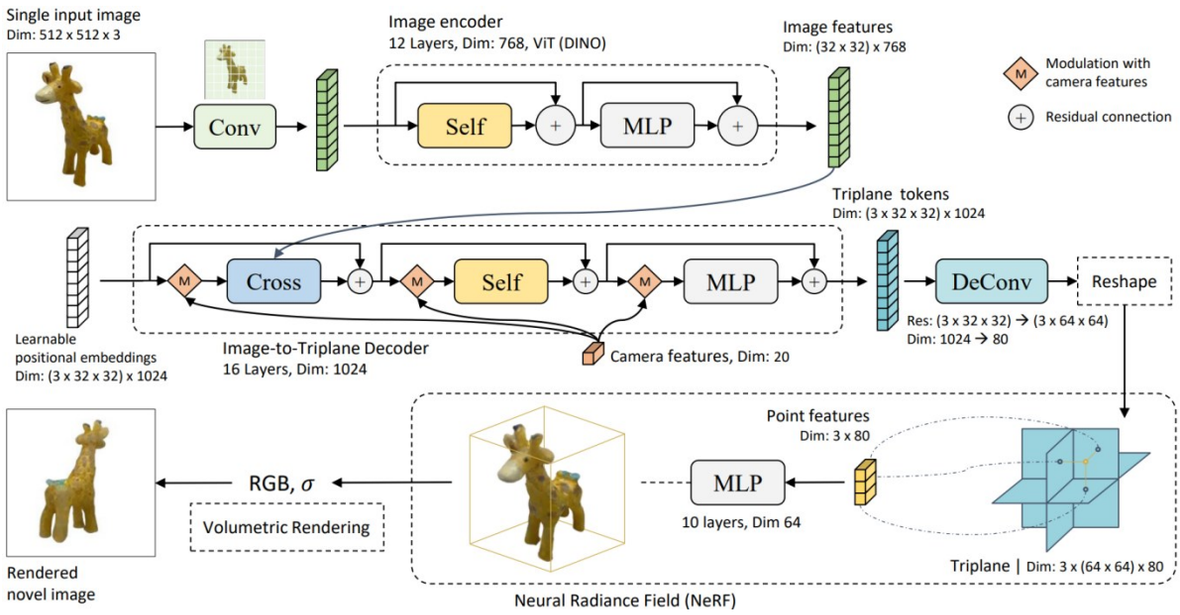

图11　LRM方法流程图(Hong等,2023)

Fig. 11　The pipeline of LRM (Hong et al, 2023)

成器,通过渐进式训练策略(逐步扩展潜码长度与模型规模)将模型参数扩展至15亿,实现了高质量三维几何的直接生成。为支撑大模型训练,CLAY构建了高质量3D数据处理管线,包括重网格化、数据标准化及基于GPT-4V的语义标注,从而显著提升了数据多样性与结构一致性。值得注意的是,CLAY将几何生成与PBR材质合成解耦,前者生成可编辑网格,后者通过独立的多视图材质扩散模型生成2K分辨率的漫反射、粗糙度与金属度贴图,实现几何与外观的一体化生成。CLAY还支持文本、图像、体素、点云等多模态条件输入,展现出在概念建模与生产级资产生成间的良好平衡,标志着3D原生大模型在可控生成与高保真几何建模方面的重要进展。

TRELLIS(Xiang等,2024)进一步提出了结构化潜变量表示(structured latent, SLAT),在稀疏三维网格上为每个活跃体素分配局部潜向量,从而在潜空间中同时编码几何与纹理信息。通过融合来自多视图图像的DINOv2视觉特征,SLAT以统一方式表达不同3D表示形式(如NeRF、3DGS与网格),构建了跨表示的共享潜空间。TRELLIS采用两阶段生成流程:首先生成稀疏结构以确定全局几何布局,其次生成局部潜变量以刻画细节与外观。整个模型基于矫正流Transformer实现潜空间建模,并在50万规模的高质量3D资产上训练,参数量高达20亿。得益于结构化潜空间的高效性,TRELLIS仅需约10秒即可从文本或图像提示生成多格式3D资产。此外,SLAT的局部性特征使模型天然支持区域化与细节级编辑,无需重新训练即可通过文本或图像提示实现局部修改。该工作以统一潜空间打破了不同三维表示间的

壁垒,兼具可扩展性、可控性与高效率,为未来"表示无关"的三维基础模型奠定了重要基础。TripoSG(Li等,2025)将该范式推向新的规模与质量高度。其核心提出基于矫正流Transformer的高保真形状生成框架,首次在3D领域实现与2D生成模型相当的训练与推理能力。TripoSG采用VAE联合Flow的两阶段结构:首先利用改进的3D VAE将形状编码为多尺度潜向量集,结合神经距离场表示、法线引导与Eikonal约束以提升几何细节与表面平滑性;随后在潜空间中使用矫正流Transformer进行生成建模,以线性流动轨迹替代传统扩散过程,从而显著提升训练稳定性与采样效率。模型通过混合专家(Mixture-of-Experts, MoE)扩展至40亿参数规模,并结合CLIP与DINOv2提取的全局与局部特征进行跨模态条件控制,实现从单张图像生成高保真的三维网格。为支撑大模型训练,TripoSG构建了包含200万高质量样本的图像--神经距离场对数据集,并通过自动评分、筛选与修复机制确保数据一致性。实验表明,TripoSG在几何精度、细节保真度及图像一致性方面均显著优于现有方法,仅需约10秒即可

© 中国图象图形学报版权所有

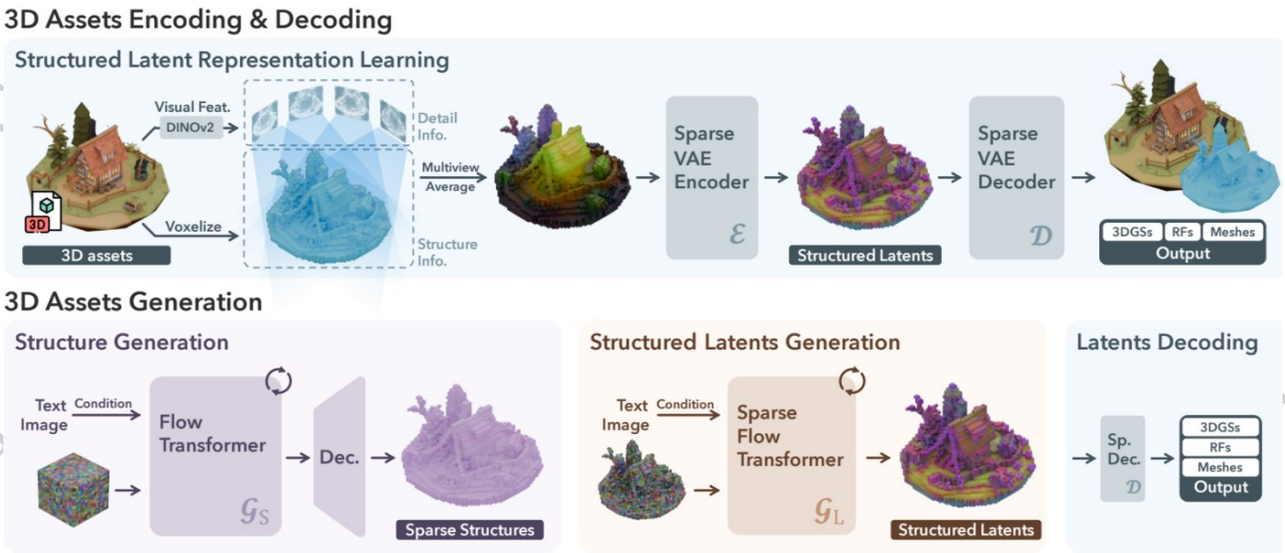

图12　TRELLIS方法流程图(Xiang等,2024)

Fig. 12　The pipeline of TRELLIS (Xiang et al, 2024)

生成复杂多物体三维场景,成为当前最具代表性的大规模流式3D生成模型。

Hunyuan3D 2.1(Hunyuan3D等,2025)进一步将该技术路线扩展至生产级三维资产生成。系统采用"形状--纹理解耦"的两阶段生成框架:形状生成部分(Hunyuan3D-DiT)基于流匹配与3D VAE的混合架构,通过Hunyuan3D-ShapeVAE实现高质量潜空间编码,并在潜空间中利用扩散Transformer进行流式去噪生成,从单图像输入生成细节丰富、几何一致的高保真网格;纹理生成部分采用多视图扩散网络,同时生成Albedo、Metallic与Roughness贴图,结合3D-Aware RoPE(Su等,2024)与空间对齐多重注意力模块,确保视图一致性与光照无关性。整个系统具有高度模块化设计,可独立输出无纹理网格或为外部资产上色。

这些基于VAE和潜空间生成的方法,通过不断改进VAE的表达能力、增大潜空间生成模型的规模、以及优化训练数据和策略,正快速提升直接3D生成的效率和质量,成为构建通用、可控、高保真3D基础模型的重要途径。

4　3D场景生成的进展

在单体对象的跨模态生成取得进展后,研究界自然地将目光投向了更宏大、更复杂的场景级生成。跨模态3D场景生成旨在根据文本描述、图像或视频等输入,合成结构化、语义丰富且视觉逼真的三维环境。与单体对象建模不同,场景级生成需要处理更大尺度、更复杂的空间布局与多对象交互。这要求模型不仅要理解单一模态的语义,还必须将其解析为全局一致的3D空间结构。近年来,随着大语言模型、2D/3D扩散模型以及NeRF和3DGS等新型三维表示的出现,跨模态场景生成呈现出多种技术路径,成为三维内容生成与世界建模(world modeling)的关键支撑方向。

4.1　基于文本驱动的程序化生成

场景生成的一个重要分支是程序化技术。传统的程序化方法,CityEngine(Parish等,2001)主要通过显式规则(如L-System)来构建环境,而基于优化的方法(如ProcTHOR(Deitke等,2022))则通过约束求解来实现合理布局。虽然这类方法一致性高,但严重依赖人工规则设计,并非跨模态方法。随着大语言模型的兴起,程序化生成迎来了语义驱动的新阶段,使其成为一种跨模态技术。大型语言模型能够理解高层语义规划,甚至直接编写控制渲染引擎的代码,从而实现自然语言到三维世界的自动映射。例如:LayoutGPT(Feng等, 2023)能根据文本提示生成场景布局参数。3D-GPT(Sun等, 2025)则通过大语言模型控制Blender或Infinigen,实现从自然语言到可视化场景的端到端生成。这一方向显著提升了生成的智能性与灵活性,标志着程序化场景生成正从规则驱动向"文本-到-代码-到-场景"的跨模态语

© 中国图象图形学报版权所有

义驱动转变。

然而,该类方法仍面临着显著的“语义-空间”鸿沟挑战。首先,LLM虽然擅长逻辑推理,但在处理精确的三维空间坐标和几何约束(如避免物体穿模、维持比例协调)时往往表现不佳,常导致布局不合理。其次,生成的视觉质量受限于底层资产库或程序化基元,难以达到生成式模型那样丰富多变的纹理细节和有机形态。最后,自然语言描述的模糊性与程序化代码的精确性之间存在天然矛盾,如何将抽象的风格描述转化为具体的渲染参数仍是一个未解决的难题。

4.2 基于2D图像先验的场景生成

受限于高质量三维数据的稀缺,近年的研究逐渐转向利用强大的二维图像生成模型,通过图像或多视图序列合成三维场景。这类方法通常能生成具有照片级真实感和丰富细节的图像,但深度与视图一致性较弱。整体生成方法倾向于一次性生成覆盖整个场景的图像,常以全景(Panorama)形式呈现。早期工作多基于GAN,而MVDiffusion(Tang等,2023)与PanoDiff(Wang等,2023)等近期方法则借助扩散模型生成360°全景图,并通过多视图一致性约束确保空间连贯性。部分研究(如LayerPano3D(Yang等,2025))进一步将全景扩散与3DGS结合,实现了高效的新视角渲染与交互式探索。另一类迭代生成方法沿相机轨迹逐步外绘新视图,实现场景的渐进扩展与延展性重建。代表性工作Infinite Nature(Liu等,2021)首次提出“渲染---精炼---重复”的迭代范式,GFVS(Rombach等,2021)利用Transformer建模长期一致性,Pose-guided Diffusion(Tseng等,2023)通过姿态调控实现精确视角控制,而Text2Room(Höllein等,2023)与SceneScape(Fridman等,2023)则将2D图像生成与Mesh/点云重建相结合,WonderJourney(Yu等,2024)借助多模态语言模型实现语义一致的场景延展,LucidDreamer(Chung等,2023)进一步通过优化NeRF或3DGS提升几何一致性。

尽管基于2D先验的方法在视觉效果上表现出色,但其在几何准确性和全局一致性上存在固有局限。主要挑战包括:1)几何模糊性与纹理投影问题,2D模型难以推断复杂的遮挡关系,导致生成的场景往往是“空心”的表面,一旦用户偏离既定轨迹,便会出现拉伸、伪影或空洞;2)多视角不一致,即物体在不同视角下呈现出相互矛盾的外观;3)迭代生成的累积误差,在长距离漫游生成中,场景的尺度、风格和几何结构容易随时间推移发生漂移,难以维

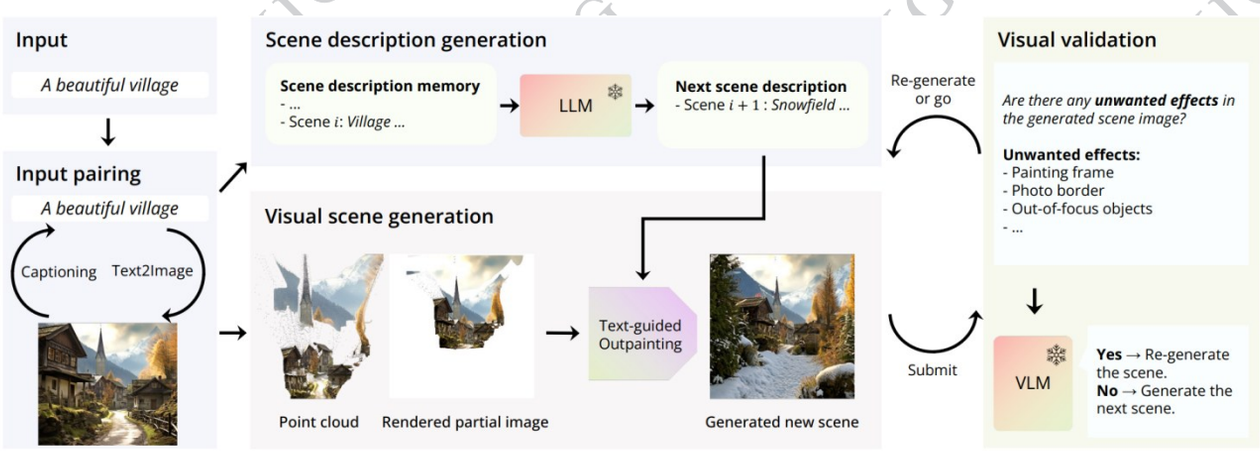

图13 WonderJourney方法流程图(Yu等,2024)

Fig. 13 The pipeline of WonderJourney (Yu et al, 2024)

持全局闭环的一致性。

4.3 基于视频先验的“世界建模”

随着视频扩散模型的快速发展,3D场景生成逐渐被统一视作一种视频生成问题,即通过时间建模来隐式实现空间一致性。这种方法将跨模态生成从静态图像提升到了动态时空。两阶段生成方法通常先生成参考视频,再将其扩展或优化为动态三维表示(如4D Gaussians Splatting)。例如,VividDream(Lee等,2025)首先生成静态场景视频并进一步渲染动画,4Real(Yu等,2024)与DimensionX(Sun等,

© 中国图象图形学报版权所有

2024)以单视频为条件合成多视角动态场景,实现空间连续与时间一致。GenXD(Zhao等,2025)与CAT4D(Wu等,2025)则采用多视点--时间联合建模策略,实现端到端4D场景合成,而GameGen-X(Che等,2025)与MagicDrive(Gao等,2024)则将用户动作、语义指令与BEV表示结合,实现可交互的动态驾驶与游戏场景生成。近期的4K4DGen(Li等,2025)与360DVD(Wang等,2024)进一步在全景视频条件下生成球面一致的动态世界,展现出连续的沉浸式世界建模能力。

虽然从图像迈向视频显著增强了动态真实性,但该类"世界建模"方法仍面临严峻挑战:首先是高昂的计算成本与显存需求,处理高分辨率、长时间序列的4D数据极大地限制了模型的普及与实时性;其次是物理一致性的缺失,当前的视频生成模型主要基于像素级的统计规律,而非底层的物理引擎,因此常出现违反物理常识的现象(如物体凭空消失、非刚体运动扭曲等);最后,长程交互的可控性依然较弱,在生成长序列时,如何精准响应用户的细粒度交互指令并保持环境的永久性仍是通向真正的世界模型必须跨越的障碍。

总体来看,从图像生成向视频生成的延伸不仅显著增强了生成结果的空间一致性与动态真实性,也为从"视觉生成"迈向"世界建模"奠定了技术基础,标志着跨模态3D生成正向具备时空理解与交互表达能力的世界模型时代迈进。

5 挑战及发展趋势展望

尽管跨模态3D生成技术在文本到3D(Text-to-3D)、图像到3D(Image-to-3D)以及复杂场景建模等方向上取得了突破性进展,但在统一表征、语义对齐、几何一致性、物理可解释性及高效落地等方面仍面临诸多挑战,严重制约了其向通用三维世界建模与智能生成的进一步发展。综合分析当前研究现状,未来跨模态3D生成的发展需重点突破以下瓶颈:

1)统一表征体系的构建仍是实现通用3D生成模型的关键难题。当前NeRF、3DGS、SDF与网格等表示各具优劣,隐式方法可微但训练成本高,显式方法高效却难以表达拓扑变化。开发能兼容多种表征并在潜空间中统一编码几何、纹理和语义信息的结构化表示,将成为未来趋势。

2)跨模态语义对齐与空间推理能力亟待提升。现有模型主要依赖CLIP或Stable Diffusion等2D语义先验,缺乏对"语言--视觉--几何"三模态的对齐机制,难以准确推理物体在三维空间中的位置、尺度及遮挡关系。如何在大规模训练数据中融入空间逻辑与物理常识,使模型具备真实三维世界的语义理解与推理能力,是提升生成质量与可控性的关键。

3)数据规模与分布偏差问题依然突出。高质量3D数据获取成本高、分布单一,导致模型泛化性受限;仿真数据与真实场景之间的域差也使生成结果在真实应用中表现退化。构建多模态、多场景、多尺度的开放3D资产库,并探索自监督或合成-真实域迁移学习,是未来必经之路。

4)生成物理一致性与可交互性不足。当前方法多聚焦视觉逼真,而缺乏对光照、材质、动力学等物理属性的约束,生成结果虽看起来真实,却难以在物理引擎或具身智能中直接使用。融合PBR渲染模型、物理仿真约束及可微分渲染机制,实现几何、纹理与物理特性协同生成,将成为迈向可交互世界模型的重要方向。

5)高效计算与模型落地机制尚不完善。SDS优化及大规模Transformer模型推理成本高、时延长,限制了在生产级和实时场景中的应用。未来需在模型量化、蒸馏与分布式推理框架上开展系统研究,结合云-边-端协同和流式生成机制,实现高效、低延迟的3D内容生成。

6)安全性、可解释性与人机共创机制仍需建立。跨模态3D生成在数据安全、偏差纠正及伦理约束方面尚无统一标准,生成结果存在不可预测性。未来应构建完善的模型纠错与偏差检测机制,引入人类反馈回路与可解释生成策略,使系统具备自我评估与自我优化能力。

本文系统综述了跨模态3D生成技术的研究进展,从基础原理、核心架构到具体应用场景进行了全面剖析。随着人工智能从二维内容生成向三维空间感知的跨越,跨模态3D生成已成为连接虚拟世界与物理现实的关键纽带,为元宇宙构建、数字孪生以及具身智能等前沿领域提供了核心技术支撑。

通过梳理技术演进脉络,本文总结出该领域呈现出的三个发展趋势:首先,在生成范式上,经历了从"基于2D先验的优化迭代"向"基于3D原生数据

© 中国图象图形学报版权所有

的直接生成”的根本性转变。早期的SDS蒸馏方法虽然解决了数据匮乏问题,但受限于效率与几何伪影;而随着Objaverse等大规模数据集的出现,基于Transformer和流匹配的端到端大模型(如LRM,TRELLIS)正在成为主流,显著提升了生成的推理速度与几何质量。其次,在三维表示上呈现出从“单一显式/隐式表示”向“高效混合表示”融合的趋势。传统的网格、点云与新兴NeRF、3DGS正通过三平面、SDF-Mesh等混合结构实现优势互补,在保证渲染高保真度的同时,兼顾了生成的结构化与可编辑性。最后,在任务维度上,正从“单一对象生成”迈向“复杂动态场景与世界模型”的构建。研究焦点已不再局限于静态物体几何重建,而是扩展至包含物理属性、时间动态及多对象交互的复杂场景生成,试图赋予模型对物理世界的时空理解与模拟能力。

尽管当前技术在视觉真实感与生成效率上取得了突破性进展,但在实现真正通用、物理一致且可交互的“3D世界模型”方面仍面临挑战。统一的多模态表征体系、对物理常识的语义理解、以及高效可控的生成机制,将是未来学术界与工业界共同攻克的重点。跨模态3D生成技术的持续演进,终将打破虚拟与现实的边界,推动人工智能从感知智能向生成智能与交互智能的全面升级。

致谢:本文由中国图象图形学学会多媒体技术专业委员会组织撰写

参考文献(References)

Achlioptas P, Diamanti O, Mitliagkas I and Guibas L. 2018. Learning representations and generative models for 3d point clouds//International conference on machine learning. Stockholm, Sweden: PMLR: 40-49

AlBahar B, Saito S, Tseng H Y, Lu J, Kim J, Kopf J and Huang J B. 2023. Single-image 3d human digitization with shape-guided diffusion//SIGGRAPH Asia 2023 Conference Papers. Sydney, Australia: 1-11 [DOI: 10.1145/3610548.3618163]

Bahmani S, Shen T, Ren J, Li Z, Li D, Park J J, Wetzstein G, Guibas L and Tagliasacchi A. 2025. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. arXiv preprint arXiv: 2509.19296

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 the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 5855-5864 [DOI: 10.1109/ICCV48922.2021.00581]

Bie X, Guo W, Leglaive S, Badeig F and Guedj B. 2022. HiT-DVAE: Human motion generation via hierarchical transformer dynamical VAE. arXiv preprint arXiv:2204.01565

Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P and Joulin A. 2021. Emerging properties in self-supervised vision transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9650-9660 [DOI: 10.1109/ICCV48922.2021.00951]

Chan E R, Lin C Z, Chan M A, Nagano K, Pan B, De Mello S, Gallo O, Guibas L, Tremblay J, Khamis S, et al. 2022. Efficient geometry-aware 3d generative adversarial networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16123-16133 [DOI: 10.1109/CVPR52688.2022.01568]

Chan E R, Monteiro M, Kellnhofer P, Wu J and Wetzstein G. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5799-5809 [DOI: 10.1109/CVPR46437.2021.00574]

Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, et al. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. Proceedings of the International Conference on 3D Vision (3DV), 2017: 667 - 676. [DOI:10.1109/3DV.2017.00081]

Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H, et al. 2015. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv: 1512.03012

Che H, He X, Liu Q, Wang Y, Zhang Y and Jin X. 2024. GameGen-X: Interactive open-world game video generation. arXiv preprint arXiv:2411.00769

Chen K, Choy C B, Savva M, Chang A X, Groueix T, Guibas L and Savarese S. 2018. Text2shape: Generating shapes from natural language by learning joint embeddings//Asian Conference on Computer Vision. Perth, Australia: Springer: 100-116 [DOI: 10.1007/978-3-030-20893-6_7]

Chen R, Chen Y, Jiao N and Jia K. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22246-22256 [DOI: 10.1109/ICCV51070.2023.02043]

Chen R, Liu Y, Kong L, Zheng X, Liu Y and Zheng Q. 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7020-7030 [DOI: 10.1109/CVPR52733.2023.00678]

Chen Y and Hu W. 2022. Stylized diffusion model based 3D generation method for Chinese-style furniture. Journal of Image and Graphics, 27(9): 2789-2801 (陈宇,胡伟. 2022. 基于风格化扩散模型的

© 中国图象图形学报版权所有

中式家具三维生成方法. 中国图象图形学报,27(9):2789-2801).[DOI:10.11834/jig.20220901]

Chen Y, Chen R, Lei J, Zhang Y and Jia K. 2022. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. Advances in Neural Information Processing Systems, 35: 30923-30936

Chen Z, Wang Y, Wang F, Yang Z and Liu Z. 2024. V3d: Video diffusion models are effective 3d generators. arXiv preprint arXiv:2403.06738

Cheng S, Bronstein M, Zhou Y, Kotsia I, Pantic M and Zafeiriou S. 2019. Meshgan: Non-linear 3d morphable models of faces. arXiv preprint arXiv:1903.10384

Cheng Y C, Lee H Y, Tulsiani S, Guibas L and Tagliasacchi A. 2023. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4456-4465[DOI: 10.1109/CVPR52733.2023.00431]

Chou G, Bahat Y and Heide F. 2023. Diffusion-sdf: Conditional generative modeling of signed distance functions//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2262-2272[DOI: 10.1109/ICCV51070.2023.00212]

Choy C B, Xu D, Gwak J Y, Chen K and Savarese S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction//European Conference on Computer Vision. Amsterdam, Netherlands: Springer: 628-644[DOI: 10.1007/978-3-319-46484-8_38]

Chung J, Lee S, Nam H, Lee J and Kim C. 2023. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384

Cohen T S, Geiger M, Köhler J and Welling M. 2018. Spherical cnns. arXiv preprint arXiv:1801.10130

Dai A, Niessner M, Zollhöfer M, Izadi S, and Theobalt C. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 5828 - 5839.[DOI:10.1109/CVPR.2017.261]

Deitke M, Schwenk D, Salvador J, Weihs L, Michel O, VanderBilt E, Schmidt L, Ehsani K, Kembhavi A and Farhadi A. 2023. Objaverse: A universe of annotated 3d objects//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 13142-13153 [DOI: 10.1109/CVPR52733.2023.01262]

Deng C, Jiang C, Qi C R, Yan X, Zhou Y, Guibas L, Anguelov D andet al. 2023. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20637-20647 [DOI: 10.1109/CVPR52733.2023.01986]

Downs L., Francis A., Koenig N., Kinman B., Hickman R., Reymann K., McHugh T.B. and Vanhoucke V., 2022, May. Google scanned objects: A high-quality dataset of 3d scanned household items. In 2022 International Conference on Robotics and Automation (ICRA) (pp. 2553-2560). IEEE.

Du Y, Zhang Y, Yu H X, Tenenbaum J B and Gan C. 2021. Neural radiance flow for 4d view synthesis and video processing//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14304-14314 [DOI: 10.1109/ICCV48922.2021.01405]

Elizarov S, Rowles C and Donné S. 2024. Geometry image diffusion: Fast and data-efficient text-to-3d with image-based surface representation. arXiv preprint arXiv:2409.03718

Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, Levi Y, Sauer F, Boesel F, Podell D, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis//International Conference on Machine Learning. Vienna, Austria: PMLR

Fang J, Yi T, Wang X, Xie L, Zhang X, Liu W, Nießner M and Tzionas D. 2022. Fast dynamic radiance fields with time-aware neural voxels//SIGGRAPH Asia 2022 Conference Papers. Daegu, Korea: 1-9 [DOI: 10.1145/3550469.3555402]

Fang Y, Xie X, Li Y, Zhu J, Chen L, Zhang X, Liu Y, Lu Y, Wang Z and Luo P. 2023. EVA-CLIP: Enhanced visual alignment CLIP. arXiv preprint arXiv:2303.11331

Feng W, Zhu W, Fu T J, Jampani V, Akula A, He X, Basu S, Wang X E and Wang W Y. 2023. Layoutgpt: Compositional visual planning and generation with large language models. Advances in Neural Information Processing Systems, 36: 18225-18250

Fridman R, Abecasis A, Kasten Y, Dekel T and Cohen-Or D. 2023. Scenescape: Text-driven consistent scene generation. Advances in Neural Information Processing Systems, 36: 39897-39914

Fu H, Cai B, Gao L, Zhang L, Wang C, Li J, et al. 2021. 3D-FRONT: 3D Furnished Rooms with Layouts and Semantics [DB/OL]. arXiv: 2011.09127 [cs. CV]. [2024-07-01]. https://arxiv. org/abs/2011.09127

Fu R, Zhan X, Chen Y, Ritchie D and Tu S. 2022. Shapecrafter: A recursive text-conditioned 3d shape generation model. Advances in Neural Information Processing Systems, 35: 8882-8895

Gao J, Shen T, Wang Z, Yin K, Li D, Litany O, Gojcic Z and Fidler S. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. Advances in Neural Information Processing Systems, 35: 31841-31854

Gao R, Chen K, Xie E, Hong L, Li Z, Yeung D Y, Xu Q and Luo P. 2023. Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601

Gao R, Holynski A, Henzler P, Sankar S and Kanazawa A. 2024. Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems, 27:

© 中国图象图形学报版权所有

2672-2680

Graham B, Engelcke M and Maaten L. 2018. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9224-9232 [DOI: 10.1109/CVPR.2018.00961]

Han B, Fu Y and Shen Y. 2023. Zero3D: Semantic-Driven Multi-Category 3D Shape Generation. arXiv preprint arXiv:2301.13591

Hanocka R, Hertz A, Fish N, Giryes R, Fleishman S and Cohen-Or D. 2019. Meshcnn: a network with an edge. ACM Transactions on Graphics, 38(4): 1-12 [DOI: 10.1145/3306346.3322959]

He Xinrui, Li Xiumei, Sun Junmei, Li Meiling, Yuan Long. Improved Pix2Vox Based 3D Reconstruction Network from Single Image[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(3): 364-372. (DOI: 10.3724/SP.J.1089.2022.18926

何鑫睿, 李秀梅, 孙军梅, 李美玲, 袁珑. 基于改进Pix2Vox的单图像三维重建网络[J]. 计算机辅助设计与图形学学报, 2022, 34(3): 364-372. DOI: 10.3724/SP.J.1089.2022.18926

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33: 6840-6851

Höllein L, Cao A, Owens A, Johnson J and Nießner M. 2023. Text2room: Extracting textured 3d meshes from 2d text-to-image models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 7909-7920 [DOI: 10.1109/ICCV51070.2023.00730]

Hong Y, Zhang K, Gu J, Bi S, Zhou Y, Liu Z, Liu F, Sunkavalli K, Bui T and Tan H. 2023. Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400

Hu Jiaping, Zhou Yang. 2025. 3D stylized portrait synthesis and structured face modeling. Journal of Image and Graphics, 30(04):1155-1169 DOI: 10.11834/jig.240380. (胡佳平, 周漾. 2025. 三维风格化人脸生成与结构化建模. 中国图象图形学报, 30(04):1155-1169) DOI: 10.11834/jig.240380.

Hu S M, Liu Z N, Guo M H, Wang J, Liu Y J and Wang R. 2022. Subdivision-based mesh convolution networks. ACM Transactions on Graphics, 41(3): 1-16 [DOI: 10.1145/3507907]

Huang R, Pan X, Zheng H, Liu Y, Wang Z and Lu H. 2024. Text4Point: Joint representation learning for text and 3d point cloud. Pattern Recognition, 147: 110086 [DOI: 10.1016/j.patcog.2023.110086]

Huang T, Dong B, Yang Y, Wu Y, Xu Y, Zhang S and Bao H. 2023. Clip2point: Transfer clip to point cloud classification with image-depth pre-training//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22157-22167 [DOI: 10.1109/ICCV51070.2023.02031]

Hunyuan3D T, Yang S, Yang M, Zhang Y, Wang P, Zhang Q, Liu Z, Lin D, Qiao Y and Guo B. 2025. Hunyuan3D 2.1: From Images to High-Fidelity 3D Assets with Production-Ready PBR Material. arXiv preprint arXiv:2506.15442

Ibing M, Kobsik G and Kobbelt L. 2023. Octree transformer: Autoregressive 3d shape generation on hierarchially structured sequences//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2698-2707 [DOI: 10.1109/CVPR52733.2023.00263]

Ignatyev S, Konovalova N, Selikhanovych D, Voynov O, Artemov A, Wang Q, Wang Y, Niessner M and Burnaev E. 2024. A3D: Does Diffusion Dream about 3D Alignment? . arXiv preprint arXiv:2406.15020

Jia C, Yang Y, Xia Y, Chen Y T, Parekh Z, Pham H, Le Q, Sung Y H, Li Z and Duerig T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918

Jun H and Nichol A. 2023. Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463

Lee J, Lee S, Jo C, Kim J and Choo J. 2024. SemCity: Semantic Scene Generation with Triplane Diffusion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE

Lee Y C, Chen Y T, Wang A, Tseng H Y, Chiu W C and Sun M. 2024. Vividdream: Generating 3d scene with ambient dynamics. arXiv preprint arXiv:2405.20334

Leng Z, Birdal T, Liang X, Guibas L and Tagliasacchi A. 2024. Hypersdfusion: Bridging hierarchical structures in language and geometry for enhanced 3d text2shape generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 19691-19700 [DOI: 10.1109/CVPR52733.2024.01868]

Li H R, Sun X and Wu F. 2024. Knowledge graph integrated text-driven 3D scene layout generation. Acta Automatica Sinica, 50(3): 567-578 (李浩然,孙晓,吴飞. 2024. 融合知识图谱的文本驱动三维场景布局生成. 自动化学报,50(3):567-578). [DOI:10.16383/j.aas.c240123]

Li H, Yang Z, Han J, Wang Y and Liu Y. 2020. TL-Net: A Novel Network for Transmission Line Scenes Classification. Energies, 13(15): 3910 [DOI: 10.3390/en13153910]

Li J, Lu J, Li H, Zhang C and Mei K. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597

Li J, Tan H, Zhang K, Xu Y, Bi S, Zhou Y, Sunkavalli K, Bui T, Tan H andet al. 2023. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. arXiv preprint arXiv:2311.06214

Li Jian, Yang Jun, Wang Liyan, Wang Yonggui. 2024. Incorporating variational auto-encoder networks for text-driven generation of 3D motion human body. Journal of Image and Graphics, 29(05):1434-1446 DOI: 10.11834/jig.230291. (李健, 杨钧, 王丽燕, 王永归. 2024. 融入变分自编码网络的文本生成三维运动人体. 中国图

© 中国图象图形学报版权所有

象图形学报，29(05)：1434-1446) DOI：10.11834/jig.230291.

Li M，Duan Y，Zhou J，Lu J andet al. 2023. Diffusion-sdf：Text-to-shape via voxelized diffusion//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver，Canada：IEEE：12642-12651［DOI：10.1109/CVPR52733.2023.01217］

Li R，Pan P，Yang B，Zhang Y，Liu Y，Lin D and Dai B. 2024. 4k4dgen：Panoramic 4d generation at 4k resolution. arXiv preprint arXiv：2406.13527

Li Y，Dou Y，Chen X，Qi L，Chen X，He Y，Guibas L andet al. 2023. 3dqd：Generalized deep 3d shape prior via part-discretized diffusion process. arXiv preprint arXiv：2303.10406

Li Y，Zou Z X，Liu Z，Zhang Y，Liu Y，Wang P，Lin D and Qiao Y. 2025. Triposg：High-fidelity 3d shape synthesis using large-scale rectified flow models. arXiv preprint arXiv：2502.06608

Liang D，Zhou X，Xu W，Zhu D，Li Z and Yu J. 2024. Pointmamba：A simple state space model for point cloud analysis. Advances in Neural Information Processing Systems，37：32653-32677

Liang Y，Yang X，Lin J，Qi X，Wang Y，Wang Y，Zhang H andet al. 2024. Luciddreamer：Towards high-fidelity text-to-3d generation via interval score matching//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle，USA：IEEE：6517-6526［DOI：10.1109/CVPR52733.2024.00629］

Lin C，Gao J，Tang L，Shao X，Wang F，Zou Y，Liu X and Qiao Y. 2023. Magic3D：High-Resolution Text-to-3D Content Creation//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：300-309［DOI：10.1109/CVPR52733.2023.00035］

Liu A，Tucker R，Jampani V，Makadia A，Snavely N and Kanazawa A. 2021. Infinite nature：Perpetual view generation of natural scenes from a single image//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：14458-14467［DOI：10.1109/ICCV48922.2021.01420］

Liu Jia，Pan Zihan，Chen Dapeng，Zhang Jiahui，Lu Guorui，Zhang Yangrui. Text-to-3D Object Generation with Grid Transformation and Material Illumination Separation［J］. Journal of Computer-Aided Design & Computer Graphics. DOI：10.3724/SP. J. 1089.2024-00320（刘佳，潘子涵，陈大鹏，张家辉，卢国瑞，张洋瑞. 具有网格转化和材质光照分离的文本到三维物体生成［J］. 计算机辅助设计与图形学学报. DOI：10.3724/SP.J.1089.2024-00320

Liu K，Zhu Z，Liu H，Wang Y，You Y and Liu Y. 2025. Acc3D：Accelerating Single Image to 3D Diffusion Models via Edge Consistency Guided Score Distillation//Proceedings of the Computer Vision and Pattern Recognition Conference

Liu M，Shi R，Kuang K，Xu Y and Zheng Q. 2023. One-2-3-45：Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems，36：22226-22246

Liu M，Shi R，Kuang K，Xu Y and Zheng Q. 2023. Openshape：Scaling up 3d shape representation towards open-world understanding. Advances in Neural Information Processing Systems，36：44860-44879

Liu R，Wu R，Van Hoorick B，Tokmakov P，Zakharov S and Vondrick C. 2023. Zero-1-to-3：Zero-shot one image to 3d object//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris，France：IEEE：9298-9309［DOI：10.1109/ICCV51070.2023.00856］

Liu S，Zhang X，Zhang Z，Zhang R，Zhu J Y and Russell B. 2021. Editing conditional radiance fields//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：5773-5783［DOI：10.1109/ICCV48922.2021.00573］

Liu Y T，Wang L，Yang J，Guo Y C and Liu Y J. 2023. Neudf：Leaning neural unsigned distance fields with volume rendering//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：237-247［DOI：10.1109/CVPR52733.2023.00031］

Liu Y，Lin C，Zeng Z，Long X，Liu Z，Komura T and Wang W. 2023. Syncdreamer：Generating multiview-consistent images from a single-view image. arXiv preprint arXiv：2309.03453

Liu Y，Wang X and Li J F. 2024. Multimodal 3D shape retrieval based on deep feature and graph structure alignment. Chinese Journal of Computers，47(2)：345-359（刘洋，王鑫，李俊峰. 2024. 基于深度特征与图结构对齐的多模态三维形状检索. 计算机学报，47(2)：345-359).［DOI：10.11897/SP.J.1016.2024.00345］

Liu Z，Dai P，Li R，Wang B and Zhang H. 2022. ISS：Image as stepping stone for text-guided 3D shape generation. arXiv preprint arXiv：2209.04145

Liu Z，Feng Y，Black M J，Tang J，Liu Y and Zhou K. 2023. Meshdiffusion：Score-based generative 3d mesh modeling. arXiv preprint arXiv：2303.08133

Lombardi S，Simon T，Saragih J，Schwartz G，Lehrmann A and Pfister H. 2019. Neural volumes：Learning dynamic renderable volumes from images. arXiv preprint arXiv：1906.07751

Long X，Guo Y C，Lin C，Liu Y，Liu Z，Liu Y J，Komura T，Zhang Y and Wang W. 2024. Wonder3d：Single image to 3d using cross-domain diffusion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：9970-9980［DOI：10.1109/CVPR52733.2024.00943］

Long X，Lin C，Liu L，Liu Y，Wang P，Theobalt C，Komura T and Wang W. 2023. Neuraludf：Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：20834-20843［DOI：10.1109/CVPR52733.2023.01995］

Lu S，Lin H，Yao L，Wang Y，Zhang Y，Liu Y and Jin X. 2025. Uni-3Dar：Unified 3d generation and understanding via autoregression on compressed spatial tokens. arXiv preprint arXiv：2503.16278

Lu T，Yu M，Xu L，Xiangli Y，Wang L，Lin D and Dai B. 2024.

© 中国图象图形学报版权所有

Scaffold-GS: Structured 3D Gaussians for View-Adaptive Rendering//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE

Luo A, Li T, Zhang W H, Liu Y J and Wang R. 2021. Surfgen: Adversarial 3d shape synthesis with explicit surface discriminators//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 16238-16248 [DOI: 10.1109/ICCV48922.2021.01594]

Ma Z, Wei Y, Zhang Y, Yang J, Lin G, Lin D and Qiao Y. 2024. Scaledreamer: Scalable text-to-3d synthesis with asynchronous score distillation//European Conference on Computer Vision. Milan, Italy: Springer: 1-19

Masci J, Boscaini D, Bronstein M and Vandergheynst P. 2015. Geodesic convolutional neural networks on riemannian manifolds//Proceedings of the IEEE International Conference on Computer Vision Workshops. Santiago, Chile: IEEE: 37-45 [DOI: 10.1109/ICCVW.2015.16]

Maturana D and Scherer S. 2015. Voxnet: A 3d convolutional neural network for real-time object recognition//2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg, Germany: IEEE: 922-928 [DOI: 10.1109/IROS.2015.7353481]

Melas-Kyriazi L, Laina I, Rupprecht C and Vedaldi A. 2023. Realfusion: 360° reconstruction of any object from a single image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 8446-8455 [DOI: 10.1109/CVPR52733.2023.00816]

Michel O, Bar-On R, Liu R, Benaim S and Hanocka R. 2022. Text2mesh: Text-driven neural stylization for meshes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13492-13502 [DOI: 10.1109/CVPR52688.2022.01314]

Mittal P, Cheng Y C, Singh M and Tulsiani S. 2022. Autosdf: Shape priors for 3d completion, reconstruction and generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 306-315 [DOI: 10.1109/CVPR52688.2022.00038]

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): 1-15 [DOI: 10.1145/3528223.3530127]

Nash C, Ganin Y, Eslami S M A and Battaglia P. 2020. Polygen: An autoregressive generative model of 3d meshes//International Conference on Machine Learning. Vienna, Austria: PMLR: 7220-7229

Nguyen-Phuoc T H, Richardt C, Mai L, Yang Y and Mitra N. 2020. Blockgan: Learning 3d object-aware scene representations from unlabelled images. Advances in Neural Information Processing Systems, 33: 6767-6778

Nguyen-Phuoc T, Li C, Theis L, Richardt C and Yang Y L. 2019. Hologan: Unsupervised learning of 3d representations from natural images//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE: 7588-7597 [DOI: 10.1109/ICCV.2019.00768]

Nichol A and Dhariwal P. 2021. Improved denoising diffusion probabilistic models//International Conference on Machine Learning. Virtual: PMLR: 8162-8171

Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I and Chen M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741

Nichol A, Jun H, Dhariwal P, Mishkin P and Chen M. 2022. Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751

Niemeyer M and Geiger A. 2021. Giraffe: Representing scenes as compositional generative neural feature fields//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11453-11464 [DOI: 10.1109/CVPR46437.2021.01129]

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, et al. 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193

Pang Y, Tay E H F, Yuan L and Liu M. 2023. Masked autoencoders for 3d point cloud self-supervised learning. World Scientific Annual Review of Artificial Intelligence, 1: 2440001 [DOI: 10.1142/S2811002924400015]

Parish Y I and Müller P. 2001. Procedural modeling of cities//Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA: 301-308 [DOI: 10.1145/383259.383292]

Park J J, Florence P, Straub J, Newcombe R and Lovegrove S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 165-174 [DOI: 10.1109/CVPR.2019.00025]

Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, Penna J and Rombach R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv: 2307.01952

Poole B, Jain A, Barron J T and Levinson D. 2022. Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988

Pumarola A, Corona E, Pons-Moll G and Moreno-Noguer F. 2021. D-nerf: Neural radiance fields for dynamic scenes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10318-10327 [DOI: 10.1109/CVPR46437.2021.01018]

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 the IEEE Conference on Computer Vision and Pattern Recognition.

© 中国图象图形学报版权所有

Honolulu, USA: IEEE: 652-660 [DOI: 10.1109/CVPR.2017.16]

Qi C R, Yi L, Su H and Guibas L J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems, 30: 5099-5108

Qi Z, Fang Y, Sun Z, Wu Z, Wang J, Zhang Y and Lu H. 2024. Gpt4point: A unified framework for point-language understanding and generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26417-26427 [DOI: 10.1109/CVPR52733.2024.02532]

Qian G, Mai J, Hamdi A, Ren J, Siarohin A, Lee B, Li H, Skorokhodov I, Wonka P, Tulyakov S and Ghanem B. 2023. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//International Conference on Machine Learning. Virtual: PMLR: 8748-8763

Ran H, Liu J and Wang C. 2022. Surface representation for point clouds//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18942-18952 [DOI: 10.1109/CVPR52688.2022.01835]

Reiser C, Peng S, Liao Y and Geiger A. 2021. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14335-14345 [DOI: 10.1109/ICCV48922.2021.01408]

Rempe D, Birdal T, Hertzmann A, Yang J, Sridhar S and Guibas L J. 2021. Humor: 3d human motion model for robust pose estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 11488-11499 [DOI: 10.1109/ICCV48922.2021.01131]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10684-10695 [DOI: 10.1109/CVPR52688.2022.01042]

Rombach R, Esser P and Ommer B. 2021. Geometry-free view synthesis: Transformers and no 3d priors//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14356-14366 [DOI: 10.1109/ICCV48922.2021.01411]

Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M and Aberman K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22500-22510 [DOI: 10.1109/CVPR52733.2023.02155]

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E L, Ghasemipour S K S, Ayan B K, Mahdavi S S, Lopes R G, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35: 36479-36494

Sanghi A, Chu H, Lambourne J G, Wang Y, Cheng C Y, Fumero M and Malekshan K R. 2022. Clip-forge: Towards zero-shot text-to-shape generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18603-18613 [DOI: 10.1109/CVPR52688.2022.01806]

Sanghi A, Fu R, Liu V, Willis K, Ritchie D and Tu S. 2023. Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18339-18348 [DOI: 10.1109/CVPR52733.2023.01758]

Sbrolli C, Cudrano P, Frosi M and Matteucci M. 2022. Ic3d: Image-conditioned 3d diffusion for shape generation. arXiv preprint arXiv: 2211.10865

Sbrolli C and Matteucci M. 2025. SCENEFORGE: Enhancing 3D-text alignment with Structured Scene Compositions. arXiv preprint arXiv:2509.15693

Schwarz K, Liao Y, Niemeyer M and Geiger A. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems, 33: 20154-20166

Shen T, Munkberg J, Hasselgren J, Gao J, Chen Z, Gojcic Z, Fidler S and Mitra N J. 2023. FlexiCubes: Flexible Isosurface Extraction for Gradient-Based Mesh Optimization. ACM Transactions on Graphics, 42(4)

Shi R, Chen H, Zhang Z, Liu M, Xu C, Wei X, Zhang L and Ye J. 2023. Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110

Shi Y, Wang P, Ye J, Mai B, Li Y and Yao J. 2023. Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512

Shu D W, Park S W and Kwon J. 2019. 3d point cloud generative adversarial network based on tree structured graph convolutions//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE: 3859-3868 [DOI: 10.1109/ICCV.2019.00396]

Siddiqui Y, Alliegro A, Artemov A, Tommasi T, Sirigatti D, Rosov V, Dai A and Nießner M. 2024. Meshgpt: Generating triangle meshes with decoder-only transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 19615-19625 [DOI: 10.1109/CVPR52733.2024.01852]

Song A P, Di X Y, Xu X K and Liu Y. 2020. MeshGraphNet: An-effective-3D-polygon-mesh-recognition-with-topology-reconstruction. IEEE Access, 8: 205181-205189 [DOI: 10.1109/ACCESS.2020.3037523]

Su J, Ahmed M, Lu Y, Pan S, Bo W and Liu Y. 2024. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568: 127063 [DOI: 10.1016/j.neucom.2023.127063]

© 中国图象图形学报版权所有

Sun C, Han J, Deng W, Liu Y, Zhang Y, Li H andet al. 2025. 3d-gpt: Procedural 3d modeling with large language models//International Conference on 3D Vision. Tokyo, Japan: IEEE: 1253-1263

Sun J, Zhang B, Shao R, Wang L, Liu W, Xie Z and Liu Y. 2023. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. arXiv preprint arXiv:2310.16818

Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, et al. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 2446 - 2454. [DOI:10.1109/CVPR42600.2020.00252]

Sun W, Chen S, Liu F, Yin K, Liu Y and Wang B. 2024. DimensionX: Create any 3d and 4d scenes from a single image with controllable video diffusion. arXiv preprint arXiv:2411.04928

Tancik M, Casser V, Yan X, Pradhan S, Mildenhall B, Srinivasan P P, Barron J T and Kretzschmar H. 2022. Block-nerf: Scalable large scene neural view synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8248-8258 [DOI: 10.1109/CVPR52688.2022.00810]

Tang J, Ren J, Zhou H, Liu Z and Zeng G. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653

Tang J, Wang T, Zhang B, Zhang Y, Chen T, Bao H, Zhang D and Chen D. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22819-22829 [DOI: 10.1109/ICCV51070.2023.02093]

Tang L, Jia M, Wang Q, Phoo C P, Hariharan B, Snavely N, Yang S and Chao W L. 2023. Mvdiffusion: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems, 36: 1363-1389

Tseng H Y, Li Q, Kim C, Alsisan S, Huang J B and Kopf J. 2023. Pose-guided diffusion models for consistent view synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 16773-16783 [DOI: 10.1109/CVPR52733.2023.01608]

Tsalicoglou C, Manhardt F, Tonioni A, Niemeyer M and Tombari F. 2023. Textmesh: Generation of realistic 3d meshes from text prompts. arXiv preprint arXiv:2304.12439

Tudosiu P D, Pinaya W H L, Ferreira Da Costa P, Graham M S, Varsavsky T, Nachev P, Ourselin S and Cardoso M J. 2024. Realistic morphology-preserving generative modelling of the brain. Nature Machine Intelligence, 6(7): 811-819 [DOI: 10.1038/s42256-024-00856-1]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30: 5998-6008

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//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5481-5490 [DOI: 10.1109/CVPR52688.2022.00542]

Voleti V, Yao C H, Boss M, Letts A, Jampani V, Rombach R, Vahdat A, Kautz J, Kreis K andet al. 2024. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion//European Conference on Computer Vision. Milan, Italy: Springer: 439-457

Wang C, Chai M, He M, Chen D and Liao J. 2022. Clip-nerf: Text-and-image driven manipulation of neural radiance fields//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 3835-3844 [DOI: 10.1109/CVPR52688.2022.00382]

Wang H, Tang J, Ji J, Sun Z, Ma L and Zhang Y. 2023. JM3D: Beyond first impressions: Integrating joint multi-modal cues for comprehensive 3d representation//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: 3403-3414 [DOI: 10.1145/3581783.3612273]

Wang J, Chen K, Xu R, Liu Z, Loy C C, and Lin D. 2023. DL3DV-10K: A Large-Scale Scene Dataset for Deep Learning-based 3D Vision. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 12885 - 12895. [DOI:10.1109/ICCV51070.2023.01183]

Wang J, Chen Z, Ling J, Wang X, Yang Z, Ma C and Liu Z. 2023. Panodiff: 360-degree panorama generation from few unregistered nfov images. arXiv preprint arXiv:2308.14686

Wang L, Zhang H and Xiao J G. 2022. PointCLIP: Hierarchical contrastive learning for point cloud and text cross-modal pre-training. Journal of Software, 33(11): 4235-4250 (王力,张华,肖建国. 2022. PointCLIP:基于层次化对比学习的点云与文本跨模态预训练. 软件学报, 33(11): 4235-4250). [DOI: 10.13328/j. cnki. jos. 006475]

Wang N, Zhang Y, Li Z, Fu Y, Liu W and Jiang Y G. 2018. Pixel2mesh: Generating 3d mesh models from single rgb images//Proceedings of the European Conference on Computer Vision. Munich, Germany: 52-67 [DOI: 10.1007/978-3-030-01252-6_4]

Wang P, Liu L, Liu Y, Theobalt C, Komura T and Wang W. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689

Wang P and Shi Y. 2023. ImageDream: Image-prompt multi-view diffusion for 3d generation. arXiv preprint arXiv:2312.02201

Wang Q, Li W, Mou C, Wang X, Jin X, Shan Y andet al. 2024. 360dvd: Controllable panorama video generation with 360-degree video diffusion model//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6913-6923 [DOI: 10.1109/CVPR52733.2024.00664]

Wang Z, Lu C, Wang Y, Bao F, Li C, Su H and Zhu J. 2023. Prolific-

© 中国图象图形学报版权所有

dreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems, 36: 8406-8441

Wei S T, Wang R H, Zhou C Z, Chen Z, Liu Y and Hua B S. 2025. Octgpt: Octree-based multiscale autoregressive models for 3d shape generation//Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. Los Angeles, USA: 1-11 [DOI: 10.1145/3651229.3651256]

Wu J, Li R, Zhu Y, Zhang Y and Liu Y. 2025. Sparse2DGS: Geometry-Prioritized Gaussian Splatting for Surface Reconstruction from Sparse Views//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Wu J, Wang Y, Xue T, Sun X, Freeman W T and Tenenbaum J B. 2017. Marrnet: 3d shape reconstruction via 2.5 d sketches. Advances in Neural Information Processing Systems, 30: 540-550

Wu J, Zhang C, Xue T, Freeman W T and Tenenbaum J B. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. Advances in Neural Information Processing Systems, 29: 82-90

Wu Qiling, Xu Kun. 3DGS Generation and Color Editing with Rich Text [J]. Journal of Computer-Aided Design & Computer Graphics. DOI: 10.3724/SP.J.1089.2025-00261(吴启凌，徐昆.基于富文本的三维高斯生成与色彩编辑[J].计算机辅助设计与图形学学报. DOI: 10.3724/SP.J.1089.2025-00261)

Wu R, Gao R, Poole B, Kautz J, Holynski A and Kanazawa A. 2025. Cat4d: Create anything in 4d with multi-view video diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26057-26068 [DOI: 10.1109/CVPR52733.2025.02489]

Wu T., Zhang J., Fu X., Wang Y., Ren J., Pan L., Wu W., Yang L., Wang J., Qian C. and Lin, D., 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 803-814).

Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X and Xiao J. 2015. 3d shapenets: A deep representation for volumetric shapes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1912-1920 [DOI: 10.1109/CVPR.2015.7298801]

Wu Z, Wang Y, Feng M, Zhang S, Xu Y and He X. 2023. Sketch and text guided diffusion model for colored point cloud generation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 8929-8939 [DOI: 10.1109/ICCV51070.2023.00822]

Xiang J, Lv Z, Xu S, Wang Y, Lin D and Qiao Y. 2024. Trellis: Structured 3d latents for scalable and versatile 3d generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 21469-21480 [DOI: 10.1109/CVPR52733.2024.02026]

Xie H, Yao H, Sun X, Zhou S and Zhang S. 2019. Pix2vox: Context-aware 3d reconstruction from single and multi-view images//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE: 2690-2698 [DOI: 10.1109/ICCV.2019.00278]

Xu D, Jiang Y, Wang P, Fan Z, Shi H and Wang Z. 2023. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360° views//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4479-4489 [DOI: 10.1109/CVPR52733.2023.00435]

Xu J, Cheng W, Gao Y, Li X, Zhang J, Huang Y and Liu Y. 2024. InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models. arXiv preprint arXiv:2403.19652

Xu J, Cheng W, Gao Y, Li X, Zhang J, Huang Y and Liu Y. 2024. Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 21669-21678 [DOI: 10.1109/CVPR52733.2024.02049]

Xu Q, Wang W, Ceylan D, Mech R and Neumann U. 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. Advances in Neural Information Processing Systems, 32: 492-502

Xu Y, Tan H, Luan F, Bi S, Wang Z, Li J, Shi Z, Sunkavalli K, Wetzstein G, Xu Z, et al. 2023. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. arXiv preprint arXiv: 2311.09217

Xue L, Gao M, Xing C, Martín-Martín R, Fei-Fei J, Wu J, Xiong C, Xu H, Savarese S and Gweon H. 2023. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1179-1189 [DOI: 10.1109/CVPR52733.2023.00120]

Xue L, Yu N, Zhang S, Darrell T, Gonzalez J E and Garrity-Albiz D. 2024. Ulip-2: Towards scalable multimodal pre-training for 3d understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 27091-27101 [DOI: 10.1109/CVPR52733.2024.02586]

Yang H, Chen Y, Pan Y, Wang Z, Liu Z and Lin D. 2024. Dreammesh: Jointly manipulating and texturing triangle meshes for text-to-3d generation//European Conference on Computer Vision. Milan, Italy: Springer: 162-178

Yang H, Chen Y, Pan Y, Wang Z, Liu Z, Lin D and Qiao Y. 2024. Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: 6870-6879 [DOI: 10.1145/3641519.3657491]

© 中国图象图形学报版权所有

Yang S, Tan J, Zhang M, Liu Y, Wang Y and Jin X. 2025. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation//Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. Los Angeles, USA: 1-10 [DOI: 10.1145/3651229.3651267]

Yeshwanth C, Liu Y-C, Niessner M, and Dai A. 2023. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 12896 - 12907. [DOI:10.1109/ICCV51070.2023.01184]

Yi T, Fang J, Wang J, Wu L, Zhang Y, Zhao X, Liu Y and Wang R. 2024. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6796-6807 [DOI: 10.1109/CVPR52733.2024.00662]

Yin F, Chen X, Zhang C, Wang Y, Jin X andet al. 2025. Shapegpt: 3d shape generation with a unified multi-modal language model. IEEE Transactions on Multimedia, 4107-4120

Yu A, Ye V, Tancik M, Kanazawa A. 2021. pixelnerf: Neural radiance fields from one or few images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE: 4578-4587 [DOI: 10.1109/CVPR46437.2021.00455]

Yu H X, Duan H, Hur J, Lee G W, Zhang K, Shakhnarovich G and Park J J. 2024. Wonderjourney: Going from anywhere to everywhere//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6658-6667 [DOI: 10.1109/CVPR52733.2024.00639]

Yu H, Wang C, Zhuang P, Zhang J, Wang Y, Zhang Y, Zhang Y, Wang Z, Wetzstein G andet al. 2024. 4real: Towards photorealistic 4d scene generation via video diffusion models. Advances in Neural Information Processing Systems, 37: 45256-45280

Yu X, Guo Y C, Li Y, Liang D, Liu Y J, Qiao Y and Liu M. 2024. Text-to-3d with classifier score distillation. International Conference on Learning Representations (ICLR)

Yu X, Tang L, Rao Y, Huang T, Zhou J and Lu J. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 19313-19322 [DOI: 10.1109/CVPR52688.2022.01872]

Yu X, Xu M, Zhang Y, Liu H, Ye C, Huang S, Jin X andet al. 2023. Mvimgnet: A large-scale dataset of multi-view images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9150-9161 [DOI: 10.1109/CVPR52733.2023.00883]

Yuan L, Chen D, Chen Y L, Codella N, Dai X, Gao J, Hu H, Huang X, Li B, Liu C, Liu M, Liu Z, Lu Y, Shi Y, Wang L, Wang J, Xiao B, Xiao Z, Yang J, Zeng M, Zhou L and Zhang P. 2021. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432

Zeng Y, Jiang C, Mao J, Liu Y, Qi X, Ye X, Wang Y, Liu Q, Ma Y, Wang J and Jia J. 2023. Clip2: Contrastive language-image-point pretraining from real-world point cloud data//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 15244-15253 [DOI: 10.1109/CVPR52733.2023.01464]

Zhan Q Q and Huang H. 2023. A survey on intelligent generation techniques for large-scale 3D scenes in digital twin cities. Journal of Image and Graphics, 28(5): 1289-1308 (詹庆庆,黄惠. 2023. 数字孪生城市中大规模三维场景的智能生成技术综述. 中国图象图形学报,28(5):1289-1308). [DOI:10.11834/jig.220305]

Zhang J, Dong R and Ma K. 2023. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2048-2059 [DOI: 10.1109/ICCV51070.2023.00191]

Zhang J, Xiong F and Xu M. 2024. G3pt: Unleash the power of autoregressive modeling in 3d generation via cross-scale querying transformer. arXiv preprint arXiv:2409.06322

Zhang J, Zhang Y, Cun X, Huang Y, Zhang Y, Zhao H, Lu H and Shen C. 2023. T2M-GPT: Generating human motion from textual descriptions with discrete representations//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14730-14740 [DOI: 10.1109/CVPR52733.2023.01415]

Zhang L, Wang Z, Zhang Q, Liu Z, Zhang Y, Zhou J, Lin D and Qiao Y. 2024. Clay: A controllable large-scale generative model for creating high-quality 3d assets. ACM Transactions on Graphics, 43(4): 1-20 [DOI: 10.1145/3658367]

Zhang M, Liu R and Zhao W. 2023. Disentangled generative adversarial network for controllable 3D textured mesh generation. Journal of Computer-Aided Design & Computer Graphics, 35(7): 1012-1022 (张明,刘瑞,赵伟. 2023. 解耦的生成对抗网络用于可控三维纹理网格生成. 计算机辅助设计与图形学学报,35(7):1012-1022). [DOI:10.3724/SP.J.1089.2023.19765]

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]

Zhang X, Liu Y, Li Y, Wang Y, Lin D and Qiao Y. 2025. Tar3D: Creating high-quality 3d assets via next-part prediction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 5134-5145 [DOI: 10.1109/ICCV51070.2025.00489]

Zhao H, Jiang L, Jia J, Torr P H S and Koltun V. 2021. Point transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 16259-16268 [DOI: 10.1109/ICCV48922.2021.01598]

© 中国图象图形学报版权所有

Zhao Y, Lin C C, Lin K, Wang Z, Zhang Y, Wang Y, Zhang Y, Wang Z, Wetzstein G andet al. 2024. Genxd: Generating any 3d and 4d scenes. arXiv preprint arXiv:2411.02319

Zhou L, Du Y and Wu J. 2021. 3d shape generation and completion through point-voxel diffusion//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 5826-5835 [DOI: 10.1109/ICCV48922.2021.00578]

Zhou T and Lin J. 2024. Fine-grained text-to-3D shape generation for Chinese descriptions. Journal of Computer Research and Development, 61(4): 889-902 (周涛,林军. 2024. 面向中文描述的细粒度文本到三维形状生成. 计算机研究与发展,61(4):889-902). [DOI:10.7544/issn1000-1239.202440123]

Zhou T, Tucker R, Flynn J, Fyffe G and Snavely N. 2018. Stereo magnification: Learning view synthesis using multiplane images. ACM Transactions on Graphics, 37(4), 1-12 [DOI: 10.1145/3197517.3201323]

Zhu Z, Fan Z, Jiang Y, Wang Z and Liu X. 2024. FSGS: Real-Time Few-Shot View Synthesis Using Gaussian Splatting//Proceedings of the European Conference on Computer Vision. Milan, Italy: Springer: 245-261 [DOI: 10.1007/978-3-031-72343-2_15]

作者简介

陈智能,男,教授,博士生导师,主要研究方向为计算机视觉、具身感知。E-mail: zhinchen@fudan.edu.cn

袁召全,男,副教授,硕士生导师,论文通信作者,主要研究方向为计算机视觉、多模态语义分析、生成式人工智能。E-mail: zqyuan@swjtu.edu.cn

杨小汕,男,研究员,博士生导师,主要研究方向为多媒体内容分析、计算机视觉、模式识别。E-mail: xiaoshan.yang@nlpr.ia.ac.cn

曹艺馨,男,研究员,博士生导师,主要研究方向为自然语言处理、知识工程、多模态信息处理。E-mail: yxcao@fudan.edu.cn

李亮,男,研究员,博士生导师,主要研究方向为生成式人工智能、计算机视觉、多模态学习。E-mail: liang.li@ict.ac.cn

吴晓,男,教授,博士生导师,主要研究方向为计算机视觉、智能交通、图像/视频处理。E-mail: wuxiaohk@gmail.com

鲍秉坤,女,教授,博士生导师,主要研究方向为多媒体分析、跨媒体生成。E-mail: bingkunbao@njupt.edu.cn

© 中国图象图形学报版权所有