引用格式：杨琳琳，贾博舒，周笑寒，金立标. 基于分数蒸馏采样的三维内容生成研究现状[J]. 中国传媒大学学报（自然科学版），2024，31（05）：41-58.
文章编号：1673-4793（2024）05-0041-18

基于分数蒸馏采样的三维内容生成研究现状

杨琳琳*，贾博舒，周笑寒，金立标*
（中国传媒大学信息与通信工程学院，北京 100024）

摘要：在当今科技赋能的文化传播过程中，三维内容生成工作已经成为了研究热点；得益于扩散模型与参数化三维表示模型的快速发展，基于二维扩散生成模型优化的三维内容生成技术已经出现众多令人惊叹的结果。其中，分数蒸馏采样技术则是利用二维扩散损失优化三维内容生成的重要环节。本文首先介绍扩散模型、三维表示方法和分数蒸馏采样等技术，然后汇总介绍基于分数蒸馏采样技术的三维内容生成代表性工作以及针对分数蒸馏采样技术不足的改进工作，此外进一步阐述领域相关的工作成果，并在最后对相关工作可能的改进方法和其他领域的应用作出展望。

关键词：三维内容生成；扩散模型；分数蒸馏采样
中图分类号：TP391.4　**文献标识码：**A

An overview of 3D generation based on score distillation sampling

YANG Linlin*, JIA Boshu , ZHOU Xiaohan, JIN Libiao*
（School of Information and Communication Engineering, Communication University of China, Beijing 100024, China）

Abstract: In the technology-enabled cultural communication process, 3D content generation has become a research hotpot. Thanks to the rapid development of diffusion models and parameterized 3D representation models, 3D content generation technology based on the optimization of 2D diffusion generation models has produced many amazing results, in which score distillation sampling is an important link in optimizing 3D content generation using 2D diffusion loss. In this paper at first the diffusion model, 3D representation methods and score distillation sampling were introduced. Then the representative works of 3D content generation based on score distillation sampling, and the improvement works for the shortcomings of score distillation sampling technology were summarized. Moreover, the results of the related works were further elaborated on. Finally the possible improvements of the works and prospects for applications in other fields was summarized.

Keywords: 3D content generation; diffusion models; score distillation sampling

1　引言

自2012年科学技术部发布《文化科技创新工程纲要》以来，前沿科技与文化传播融合发展就成为我国各级管理部门在文化领域的重点工作。其中，随着计算机图形学领域的快速发展，三维内容生成相关技术在文化科技领域生根发芽，并在近年来涌现出众多的应用成果。

基金项目：国家重点研发计划项目（2021YFF0900701）
作者简介（*为通讯作者）：杨琳琳（1991-），男，博士，讲师，主要从事三维姿态估计与三维生成方向研究。Email：lyang@cuc.edu.cn；贾博舒（2001-），男，硕士研究生，主要从事三维生成方面研究。Email: jbsenglish@cuc.edu.cn；周笑寒（2001-），男，硕士研究生，主要从事三维生成方向研究。Email: kevinzhou@cuc.edu.cn；金立标（1976-），男，博士，教授，主要从事智能媒体通信方向研究。Email: libiao@cuc.edu.cn

在传统文化保护与传播工作中,故宫博物院通过二维图像拍摄与三维空间扫描,并结合基于纹理模型的三维重建技术,对大量的古建筑与藏品文物进行了高精度的采集与加工,将文化遗产以立体模型方式呈现在数字世界中,加速推进了对文化遗产的保护、研究与传播展示工作[1]。而在消费级领域的文化传播中,数字游戏可以通过角色、环境与行动三个维度,并通过符号、知识和观念三个层面呈现传统文化,并带领玩家切身体验所表达的文化内容[2];今年上线的国产游戏大作《黑神话·悟空》,基于中国四大名著之一《西游记》的传统故事,配合中国各个地域的代表风景与传统文化,利用三维建模技术展现中国优秀传统文化底蕴,并在国内外展示出高品质的三维数字游戏对文化传播的有力促进作用。

腾讯研究院发布的《2022文化科技十大前沿应用趋势》[3]中提到,文化生产的"虚实共生"会成为文化科技应用的发展趋势之一,并同时指出生成式人工智能技术(AIGC, artificial intelligence generated content)与三维虚拟内容生成都会成为推动"虚实共生"文化科技生产的应用进展,并为文化内容产业带来多重创新与变革。

而在相关技术层面,AIGC领域扩散模型的出现和快速发展[4-6],带动了通过文本控制图像生成(T2I, text to image)领域的研究热潮;而基于深度学习训练优化的三维表示方法的不断更新,三维重建与生成工作也不断向更高质量、更高效率的效果进化。在这些相关工作的不断发展与影响中,结合T2I工作与三维内容生成工作的文本控制三维内容生成技术(T-3D, text to 3D)应运而生,并在近两年的三维生成相关工作中得到了迅速的发展。

在众多基于T2I扩散模型与三维表示模型的T-3D相关工作中,如何进行文本指令对三维表示模型的梯度回传优化是研究关键。分数蒸馏采样则是这样一项重要的梯度回传技术[7]。因此,本文主要研究基于分数蒸馏采样的应用方法与改进措施。本文将从代表性技术算法层面入手,首先在第二节对扩散模型、主要三维表示方法和分数蒸馏采样技术进行介绍;在第三节中,本文将分别介绍运用分数蒸馏采样实现三维内容生成的代表性相关工作,以及针对分数蒸馏采样各项不足的改进工作;在第四节中,将简要归纳相关工作中主要涉及的数据集与评价指标,以及有代表性的成果;最后在第五节中,将对全文进行总结,并分别对相关工作的可能应用领域与改进方法进行展望。

2 相关技术

本节将对扩散模型、常用三维表示方法和分数蒸馏采样进行详细介绍。

2.1 扩散模型

在生成式人工智能(AIGC)领域中,扩散模型的出现打破了对抗生成网络[8]在计算机视觉层面的垄断地位。由于扩散模型本质是在纯噪声或带噪图片的去噪过程中实现内容的生成,故也可称为去噪扩散模型(DDM, denoising diffusion models)。

2.1.1 扩散模型基本原理

扩散模型主要包含对数据的正向加噪过程q和反向去噪过程p。在正向加噪过程q中,假设理想目标数据$\boldsymbol{x}_0$符合分布$\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0)$,在此数据基础上进行逐步加噪$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_{t-1} + \sigma_t \boldsymbol{\epsilon}$,其中$\boldsymbol{\epsilon}$为采样的随机高斯噪声,$\alpha_t, \sigma_t$则为扩散过程中随时间变化的参数。在加噪过程末尾$t = T$时,数据近似为纯噪声,且整个加噪过程的数据分布可看作马尔科夫链进行计算,如式(2-1):

$$\begin{aligned} q_t(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) &= N(\boldsymbol{x}_t; \alpha_t \boldsymbol{x}_{t-1}, \sigma_t \mathbf{I}) \\ q_t(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) &= \prod_{t=1}^{T} q_t(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \end{aligned} \tag{2-1}$$

其中$\boldsymbol{x}_{1:T}$表示扩散时间步从1到T过程的数据序列。而反向去噪过程即为遵循最大后验分布逐步恢复数据的过程。在通常情况下,此过程由一个经过训练的去噪神经网络进行数据分布的拟合,常用的网络为U-Net[9]。基于正向加噪过程的马尔可夫性,反向过程即定义为一个马尔科夫链进行计算,并采用神经网络ϕ逐步拟合参数化的高斯分布,如式(2-2):

$$\begin{aligned} p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) &= N\left(\boldsymbol{x}_{t-1}; \mu_\phi(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\phi(x_t, t)\right) \\ p_\phi(\boldsymbol{x}_{0:T}) &= p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \end{aligned} \tag{2-2}$$

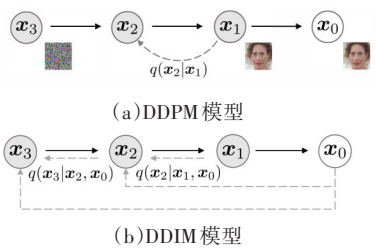
其中μ_ϕ和$\boldsymbol{\Sigma}_\phi$分别代表均值和协方差矩阵。在扩散模型的训练过程中,对数据经过加噪和去噪过程后,通过最小化去噪神经网络预测噪声$\hat{\boldsymbol{\epsilon}}_\phi$与采样噪声的区别$\min_\phi \mathrm{E}_t[\hat{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t) - \boldsymbol{\epsilon}]$进行损失优化,损失函数如式(2-3)。

$$L_{\mathrm{Diff}}(\phi, \boldsymbol{x}) = \mathrm{E}_{t\sim U(0,1),\ \boldsymbol{\epsilon}\sim N(\mathbf{0}, \boldsymbol{I})}\left[\omega(t)\left\|\hat{\boldsymbol{\epsilon}}_\phi(\alpha_t \boldsymbol{x} + \sigma_t \boldsymbol{\epsilon}; t) - \boldsymbol{\epsilon}\right\|_2^2\right] \tag{2-3}$$

当扩散去噪模型训练到最优时，预测噪声满足近似当前数据在条件 y 下的得分函数[10]，即式(2-4)所示：

$$\hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t|y,t\right)\approx -\sigma_t\nabla_x\log p_t\left(\boldsymbol{x}_t|y\right) \tag{2-4}$$

上述常用的扩散模型利用马尔科夫链性质，通过概率分布密度进行逐步计算，故此类扩散模型也称为去噪扩散概率模型(DDPM, denoising diffusion probabilistic models)。为了提升扩散模型的运算效率，去马尔可夫化的去噪扩散隐式模型(DDIM, denoising diffusion implicit models)[11]通过确定化扩散过程的噪声，使得整个加噪与去噪过程不需进行逐步计算，从而实现快速处理大规模图像数据的效果，如图 2.1 所示。



(a)DDPM 模型

(b)DDIM 模型

图 2.1 DDPM 与 DDIM 模型结构图[11]

2.1.2 计算机视觉领域代表模型

随着扩散模型在数据生成领域的爆火，以二维图像生成为代表的计算机视觉领域的扩散模型改进工作层出不穷。面对庞大的生成数据，隐式扩散模型(LDM, latent diffusion models)[12]将输入数据从像素空间中剥离出来，通过将数据压缩为潜在空间的 latent 数据并加以优化，以降低运算成本。同时，LDM 也通过引入交叉自注意力机制，处理文本或边界框等条件，初步实现了通过文本控制数据生成的效果。

而在之后改进工作中，最有代表性的扩散模型之一为稳定扩散模型(SDM, stable diffusion models)，其利用了预训练的 LDM 作为数据扩散处理部分，并引入来自 CLIP ViT-L/14[13]的文本编码器，保证对文本输入的高效处理，实现在较简短的文本指令下大幅提升图像内容生成的质量和准确性。

SDM 模型在 T2I 工作中展现的优异结果，点燃了 T2I 工作的火爆，同时也为 T-3D 工作的发展奠定了基础。

2.2 三维表示方法

在计算机图形学领域中，传统且常用的三维表示方法包括基于三维方格表示体积的体素(voxel)表示、基于三维点位置表示体积的点云(point cloud)表示和基于三角网格表示表面的网格(mesh)表示，可见图 2.2。而近年来，随着深度学习领域的发展，利用深度神经网络进行三维物体表示的工作层出不穷，在传统三维表示方法的基础上也出现了较新颖的三维表示方法。其中，使用辐射场表示三维空间内容的方式是一种保持场景精细度和运算存储效率的优秀思路。

辐射场为一种通过类似光线的射线，对环境中的物体体积点进行采样捕获，并运用特定的参数化映射对输入的特定视角数据转换为三维空间数据。由于整个映射过程即特定三维场景或物体变换，可以使用参数 $\boldsymbol{\Theta}$ 表示，故辐射场可轻松通过神经网络表示，并采用深度学习方法进行训练。

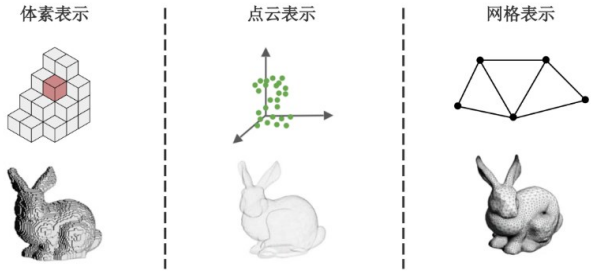


图 2.2 常见传统三维表示方法

在辐射场与深度学习优化结合进行三维表示的方法中，隐式方法神经辐射场(NeRF, neural radiance field)[14]和显式方法 3D 高斯泼溅(3DGS, 3D Gaussian splatting)[15]取得了渲染质量和应用泛化性的优异成果。

2.2.1 隐式辐射场与神经辐射场

隐式辐射场为一种无需显式地定义场景中的具体几何形状或分布的方法，通常使用神经网络学习连续的 3D 场景表示。在隐式表现三维内容时，任何点的辐射率都不会显式存储，而是通过查询神经网络参数即时计算。

而通过隐式辐射场进行三维场景表示和重建三维内容视图的众多工作中，最具里程碑意义的就是 NeRF[14]。NeRF 是一种高质量场景重建技术，它能够连续的表示场景的三维结构，从二维图像中重建出逼真高质量的三维场景，从而实现新视角的视图合成[16]。

如图 2.3(a)、(b)所示，NeRF 凭借采样点位置与观察方向 (x,y,z,θ,ϕ)，经过多层感知机(MLP, multilayer perceptron)映射生成采样点颜色 RGB 和体积密度 σ，由式(2-5)表示。而在训练过程中，NeRF 通过优化 MLP 的参数 $\boldsymbol{\Theta}$，从而实现通过训练优化存储和表示三维对象。

$$F_{\boldsymbol{\Theta}}\left(x,y,z,\theta,\phi\right)\rightarrow\left(RGB,\sigma\right) \tag{2-5}$$

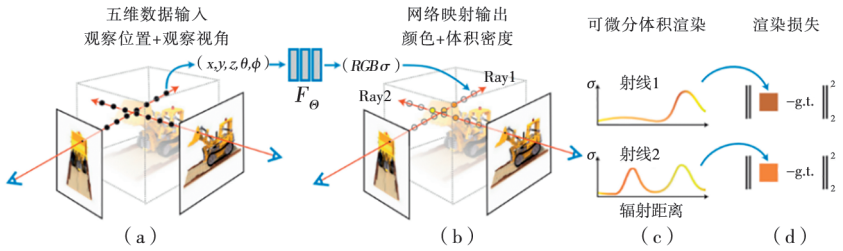


图 2.3 NeRF 场景表示和可微渲染流程图[14]

映射数据经过可微分的体积渲染合成为图像，如图2.3(c)所示，并通过最小化合成图像和真实图像之间的残差优化场景表示，如图2.3(d)所示。NeRF采用了经典体积渲染技术[17]，沿着采样光线，根据采样点的体积密度和颜色确定其累计透明度(即颜色权重)，其数学表达式如式(2-6)所示：

$$\hat{C}(r)=\sum_{i=1}^{N}T_i\left(1-\exp\left(-\sigma_i\right)\right)c_i,\ T_i=\exp\left(-\sum_{j=1}^{i-1}\sigma_j\right) \tag{2-6}$$

其中，$\hat{C}(r)$表示沿采样射线r的预测颜色，N是沿着光线的采样点数量，T_i是从光线起点到第i个点的累积透明度，σ_i与c_i分别是第i个采样点的密度和颜色。

但同时，NeRF因其全隐式表达和对MLP的深度依赖，训练和渲染需要消耗大量的算力与时间。针对此问题，InstantNGP[18]将NeRF和传统体素网格结合，并引入多分辨率哈希编码，极大提高神经辐射场训练和渲染速度，并减小空间复杂度。除此以外，Mip-NeRF[19]提出锥追踪方法实现光线追踪，改善了场景重建的精细度；Mip-NeRF 360[20]实现了无边界场景的高质量重建；而Point-NeRF[21]则使用点云进行NeRF渲染，并采用与NeRF相同的体渲染方法，储存每个采样点特征值，以实现更直观的三维表示。

2.2.2 显式辐射场与3D高斯泼溅

与隐式辐射场不同，显式辐射场则直接以可见形式表示空间离散结构中光线辐射的分布。基于此特点，显式表示不需使用神经网络对辐射信息进行编解码，故对于算力的需求较低，且能够高效快速地访问某一单位的辐射场信息。但同时，显式表达具有更大的内存需求和空间复杂度。

而显式辐射场方法中，有代表性的里程碑式的工作是3D高斯泼溅工作(3DGS)[15]。与Point-NeRF[21]相似，3DGS采用一种在三维空间中存在的各向异性的高斯分布作为高质量、非结构化的场景表达，3DGS的实现流程如图2.4所示。

在3DGS所表示的三维空间中，每一个三维高斯概率形成三维空间中的3D椭球体，分别以概率均值和协方差高效准确地表示三维几何和外观属性，并凭借分布的各向异性紧凑地表达精细几何结构。椭球体向二维空间的投影过程(泼溅过程)可以表示为式(2-7)：

$$\boldsymbol{\Sigma}'=JW\boldsymbol{\Sigma}W^{\mathrm{T}}J^{\mathrm{T}} \tag{2-7}$$

其中，$\boldsymbol{\Sigma}'$和$\boldsymbol{\Sigma}=RSST^{\mathrm{T}}R^{\mathrm{T}}$分别为3D椭球体及其在特定视角观察的二维图像上投影椭球面的协方差矩阵，椭球和椭球体的形状则通过协方差矩阵来表示；J则是投影变换的雅可比矩阵。同时，通过球面谐波函数(SH, spherical harmonics)[22]的多个参数存取三维模型在椭球法线方向的颜色信息，并以函数阶数控制颜色信息的精细度，获取可控的颜色信息，进一步实现高质量且可控的三维信息表示。

另外，在训练优化层面，3DGS提出高斯密度自适应控制机制，使用基于图块的光栅化(tile-based rasterization)[15]方法，通过混合不同高斯点的不透明度α，并无限制优化所有高斯点，以实现高效快速的训练过程。

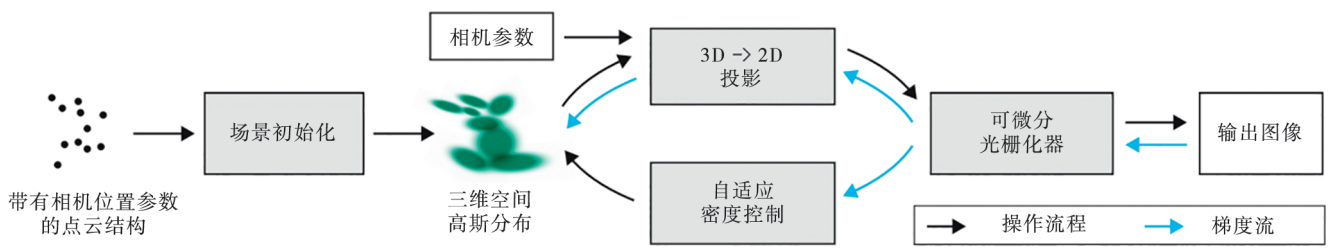


图 2.4 3DGS 实现流程[15]

2.3 基于分数蒸馏采样的三维内容处理

随着扩散模型在内容生产领域的发展，以及基于深度学习的三维内容表示方法不断成熟，在T2I工作的基础上逐渐衍生出了采用文本或其他自然指令生成或编辑三维内容的工作。其中，不少工作[23-26]通过对抗网络或更新的视觉语言模型对参数化三维表示方法进行优化，但仍没有出现较优的、应用自由度较高的代表性优化方案。

DreamFusion[7]工作则成功将扩散模型在二维生成与编辑领域的优势拓展到了三维空间，采用二维扩散模型对文本指令进行处理，并利用扩散模型的损失经过分数蒸馏采样（SDS, score distillation dampling）优化NeRF参数。

具体地说，在优化NeRF参数$\boldsymbol{\Theta}$的过程中，扩散模型部分将NeRF的渲染图片$g(\boldsymbol{\Theta})$作为输入并转换为潜在空间数据$\boldsymbol{x}_0$，对其进行采样时间步长t的一步加噪，加噪后的数据$\boldsymbol{x}_t$即送入以ϕ为参数化的去噪模型U-Net[9]，并结合文本指令y去噪。由于预测的噪声$\hat{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t|y,t)$与NeRF的输出$g(\boldsymbol{\Theta})$相关，凭借链式法则以及式(2-3)的损失函数，在此框架中扩散模型的损失梯度同时包含NeRF模型和U-Net模型的雅可比项，如式（2-8）：

$$\nabla_{\boldsymbol{\theta}} L_{\text{Diff}}\left(\phi, g(\boldsymbol{\Theta})\right) = \mathrm{E}_{t,\epsilon}\left[\omega(t)\underbrace{\left(\hat{\boldsymbol{\epsilon}}_\phi\left(\boldsymbol{x}_t; y, t\right) - \boldsymbol{\epsilon}\right)}_{\text{噪声残差项}} \underbrace{\frac{\partial \hat{\boldsymbol{\epsilon}}_\phi\left(\boldsymbol{x}_t; y, t\right)}{\partial \boldsymbol{x}_t}}_{\text{U-Net雅可比项}} \frac{\partial \boldsymbol{x}_t}{\partial g(\boldsymbol{\Theta})} \underbrace{\frac{\partial g(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}}_{\text{NeRF雅可比项}}\right] \tag{2-8}$$

其中，$\partial \boldsymbol{x}_t / \partial g(\boldsymbol{\Theta}) = \alpha_t \mathbf{I}$为扩散模型将图片数据转为潜在数据的梯度，为常数。在上述公式中，U-Net的雅可比项具有较高的计算成本，且优化效果较差，故将此项蒸馏，仅保留NeRF的雅可比项进行梯度回传，以实现对参数$\boldsymbol{\Theta}$的优化，如式（2-9）：

$$\nabla_{\boldsymbol{\theta}} L_{\text{SDS}}\left(\phi, g(\boldsymbol{\Theta})\right) \triangleq \mathrm{E}_{t,\epsilon}\left[\omega(t)\left(\hat{\boldsymbol{\epsilon}}_\phi\left(\boldsymbol{x}_t; y, t\right) - \boldsymbol{\epsilon}\right)\frac{\partial g(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}\right] \tag{2-9}$$

另外，经过数学推导，SDS损失也等价于扩散模型得分函数的加权概率密度蒸馏损失，在数学意义上相当于扩散过程与去噪过程数据分布的KL散度，如式（2-10）：

$$\nabla_{\boldsymbol{\theta}} L_{\text{SDS}}\left(\phi, g(\boldsymbol{\Theta})\right) = \nabla_{\boldsymbol{\theta}} \mathrm{E}_{t,\epsilon}\left[\frac{\sigma_t}{\alpha_t}\omega(t)\mathbf{KL}\left(q_t\left(\boldsymbol{x}_t|g(\boldsymbol{\Theta}); y, t\right)\|p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t; y, t\right)\right)\right] \tag{2-10}$$

同时，为了缓解扩散损失的模式寻求（Mode-Seeking）行为，SDS损失也加入了受参数ω控制的机制。在优化过程中，基于特定的扩散去噪条件y，引入无分类器指导（CFG, classifier-free guidance）机制调整对生成质量和条件忠实度的倾向，如式（2-11）：

$$\hat{\boldsymbol{\epsilon}}_\phi\left(\boldsymbol{x}_t; y, t\right) = (1+\omega)\boldsymbol{\epsilon}_\phi\left(\boldsymbol{x}_t; y, t\right) - \omega\boldsymbol{\epsilon}_\phi\left(\boldsymbol{x}_t; t\right) \tag{2-11}$$

其中，$\omega > 0$时，扩散去噪模型会倾向于牺牲数据多样性以保证质量。

至此，SDS通过对扩散模型部分的梯度蒸馏，将文本指令所优化的对象变换为三维表示网络的参数$\boldsymbol{\Theta}$，同时通过CFG机制和其他优化方法进行细节改进，以达到可用的T-3D效果。

3 基于SDS的三维内容生成工作研究现状

自从主流的三维表示方法和扩散模型通过SDS机制进行联合之后，受此启发，出现了很多通过扩散模型、SDS和三维表示模型实现三维生成与编辑的相关工作，如表3.1所示。

这些工作在应用领域、三维表示方法和扩散模型的运用方面具有很大差异，但均大体采用类似的优化流程，即：将可学习参数定义的三维表示网络的输出作为扩散模型部分的输入，扩散模型结合文本指令对数据进行扩散与去噪，过程中的损失梯度经过蒸馏后回传至三维表示网络进行优化，从而达到根据文本指令优化三维表示模型的目的。SDS及类似T-3D方法简略框图如图3.1所示，其中$g(\boldsymbol{\Theta}_\tau, \boldsymbol{c})$表示在模型优化轮次为$\tau$、三维表示模型参数为$\boldsymbol{\Theta}_\tau$且当前图像采样的相机参数为$\boldsymbol{c}$时，渲染出的单张图片。

通过对目前主要工作调研发现，当前在SDS机制应用领域的工作主要挑战在于如何有效地提高模型结果的质量与文本指令对齐程度，这对三维表示网络的参数合理性和渲染质量、对扩散模型部分的梯度引导能力和蒸馏损失的优化机制都有很大关系，因此SDS机制的相关应用工作的重点就是寻求更好的三维表示-扩散-蒸馏损失方案。本节则会对当前基于SDS三维内容生成的代表性工作以及针对各种问题的改进工作进行汇总。

3.1 基于SDS在三维内容生成工作中的应用

三维内容生成工作主要目标是受输入文本条件控制的三维内容建立，在优化进程之初，对三维表示

表3.1 基于SDS的三维内容生成主要工作汇总

主要工作	控制条件	三维表示模型	三维模型精细化处理	噪声预测网络	蒸馏损失法	其他损失	其他模块
Magic3D[27]	文本指令	InstantNGP	网格模型精细化	U-Net[9]	SDS	无	可变形四面体网格精细三维形状
Make-It-3D[28]	文本指令+单图像	InstantNGP	点云优化	U-Net[9]	SDS	扩散CLIP损失	无
LatentNeRF[29]	文本指令	InstantNGP	纹理网格优化渲染	U-Net[9]	SDS	无	Latent-Paint精细化纹理
Fantasia3D[30]	文本指令	DMTet[38] BRDF[39]	无	U-Net[9]	SDS	SDF损失	无
GaussianDreamer[31]	文本指令	3DGS	采用网格生成先验	U-Net[9]	SDS	无	3D扩散模型初始化点云
IPDreamer[32]	文本指令+图片风格	NeRF	几何纹理解耦	U-Net[9]	SDS+IPSD	无	Zero-1-to-3初始化；交叉注意力
MVDream[33]	文本指令	NeRF	无	多视图扩散模型	SDS	多视图扩散损失	二维和三维注意力机制训练扩散
DreamBooth3D[34]	文本指令+图片风格	NeRF	无	Imagen T2I[37]	SDS	NeRF正则化损失[20]	分步微调DreamBooth
DiverseDream[35]	文本指令+图片风格	NeRF	无	U-Net[9]	TSD	无	Hiper文本反转
InstructHumans[36]	文本指令（编辑）	NeRF	视角采样策略	U-Net[9]	SDS-E	拉普拉斯平滑损失	SMPL混合三维人体表示

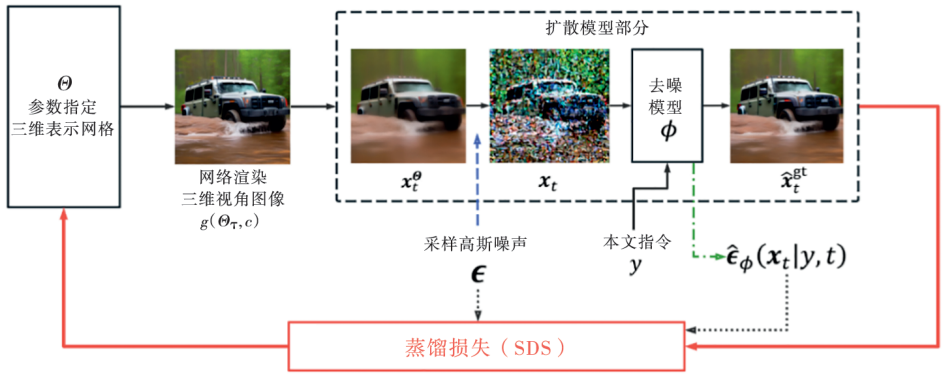


图3.1 SDS及类似T-3D方法简略框图

网络进行初始化，扩散模型则根据文本指令对噪声进行去噪，去噪结果则通过蒸馏损失对三维表示网络进行优化。而控制生成的条件则为影响整个三维生成过程的重要因素，基于当前主要工作的生成控制条件，可将此工作划分为纯文本控制生成与文本图像条件联合控制两大类，如图3.2所示。

3.1.1 纯文本条件控制相关工作

纯文本控制条件中，由于控制三维内容生成的因素仅有文本输入，无额外的输入进行细节控制。但此类工作较为简单，且在工作机制上与DreamFusion[7]原方法最为接近，主要在于优化模块和加入更多优化措施提升生成质量。

受到InstantNGP[18]的启发，针对NeRF的优化速度慢且低分辨率三维模型质量低下的问题，Magic3D[27]工作提出了采用两阶段框架优化生成三维内容，如图3.3所示。具体来说，Magic3D工作指出，在DreamFusion[7]的SDS机制中，当需要产生高分辨率的三维内容时，优化速度的低下主要来自于链式法则中对$\partial \boldsymbol{x}_t/\partial g(\boldsymbol{\Theta})$和$\partial g(\boldsymbol{\Theta})/\partial \boldsymbol{\Theta}$两项梯度的计算。也基于此，Magic3D通过采用不同的三维表示方法模型，同时利用了InstantNGP[18]模型的快速生成特点和参数易于初始化特点，以及3D网格模型的数据直观性，提升计算效率。

在第一阶段，此工作采用低分辨率的扩散模型通过SDS初始化并优化InstantNGP[18]表示的低分辨率三维内

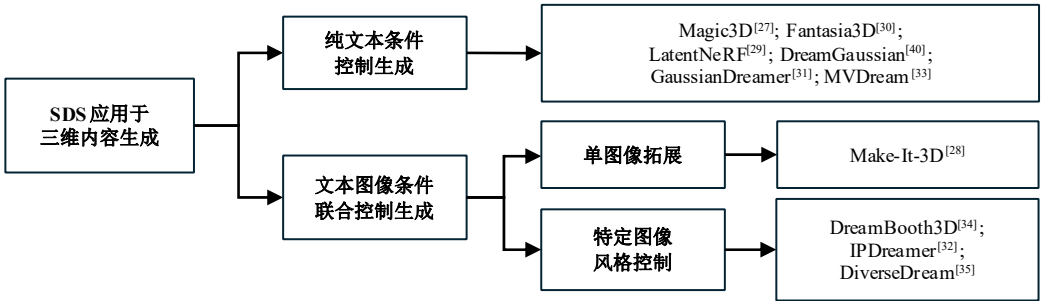


图 3.2 SDS 应用于三维内容生成主要工作归纳

容。当达到低分辨率要求时，第一阶段优化所得的三维内容信息会通过三维网格变换方法 DMTet[38]变换为三维网格模型，而网格则在第二阶段进一步被采用高分辨率部分的扩散模型进行细粒度优化。由于 3D 网格模型本身属于较为直观的三维模型，故在高分辨率下不会产生大量的计算复杂度，故可以在较高效率的计算中得到高分辨率三维内容，但由于其中带有三维模型之间的转化，无法有效保证生成质量。

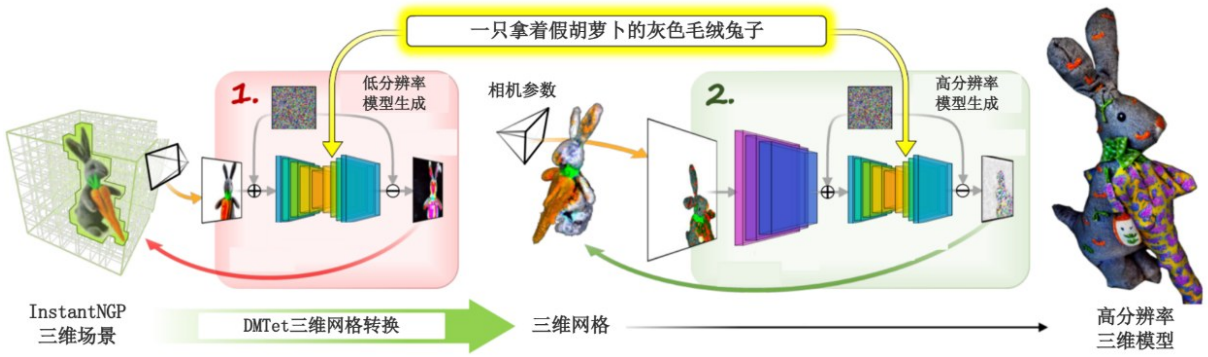


图 3.3 Magic3D 技术流程图[27]

然而，通过优化框架中的三维表示方法提升生成内容质量的工作并不在少数。由于三维模型的质量同时受到几何形状与外观纹理的影响，Fantasia3D[30]工作将几何建模与纹理建模进行分开优化，如图 3.4 所示。

其中几何建模与纹理建模两部分均为一个完整的三维模型-扩散-蒸馏损失结构。几何建模部分由参数化为 Ψ 的 DMTet 模型作为三维几何表示方法，从而预测三维对象顶点的 SDF 值 $s(\cdot)$ 和形变偏移量，并对三维对象采样点 p_i 加入了 SDF 损失，以保证几何建模的有效性，SDF 损失如式(3-1)：

$$L_{\text{SDF}} = \sum_{p_i \in \boldsymbol{P}} \left\| s\left(p_i; \boldsymbol{\Psi}\right) - SDF\left(p_i\right) \right\|_2^2 \qquad (3\text{-}1)$$

而对于几何内容，在 DMTet 所得到的几何模型的基础上，本工作采用较优的物理材质渲染模型 PBR 材质[39]，并将外观参量分为采用双向反射分布函数(BRDF)建模，从而完成在几何模型上的高质量纹理着色。但 Fantasia3D[30]与 Magic3D[27]类似，其模型中包含多个三维表示模型和扩散回传机制，不利于模型的转移与运用。

除此之外，将 NeRF 及其衍生模型作为三维表示模型的 T-3D 工作有很多，其中 LatentNeRF[29]工作通过引入 Latent-Paint 模块，将 InstantNGP 转为与扩散模型数据处理过程相同的 latent 空间中，同时采用纹理网格优化渲染，实现更高效率与质量的 T-3D 结果。但 NeRF 模型本身存在的分辨率与渲染效率上限均限制了此类工作的发展。

在优化三维表示部分的方法中，具有里程碑意义的工作之一为 GaussianDreamer[31]。受到 DreamGaussian[40]的启发，此项工作成功将二维扩散模型通过 SDS 与 3DGS[15]方法结合到一起，并保证了模型体量与生成质量，如图 3.5。

此项工作通过利用 3D 扩散模型生成初始化点云，并将点云进行加噪与颜色扰动，处理后的点云对 3DGS 模型进行初始化，并进一步利用扩散模型与 SDS 对 3DGS 模型进行优化。从实验结果也可知，3DGS 的运用，相较 NeRF 有了更高的生成质量，在效

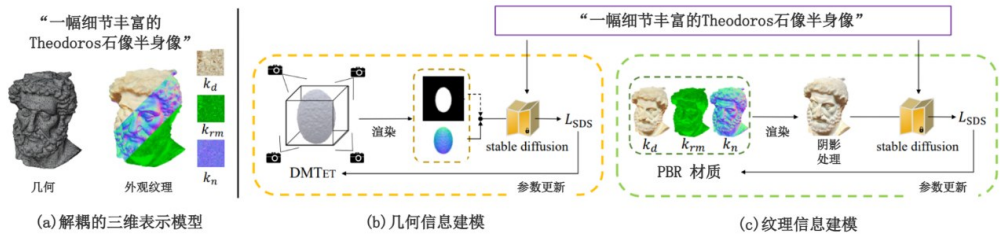


图3.4 Fantasia3D技术流程图[30]

率层面获得了提升,同时也证明了3DGS这一当前SOTA的参数化三维表示方法在三维表示-扩散-蒸馏损失框架中的可用性与优越性,为基于3DGS模型的T-3D工作打下了基础。

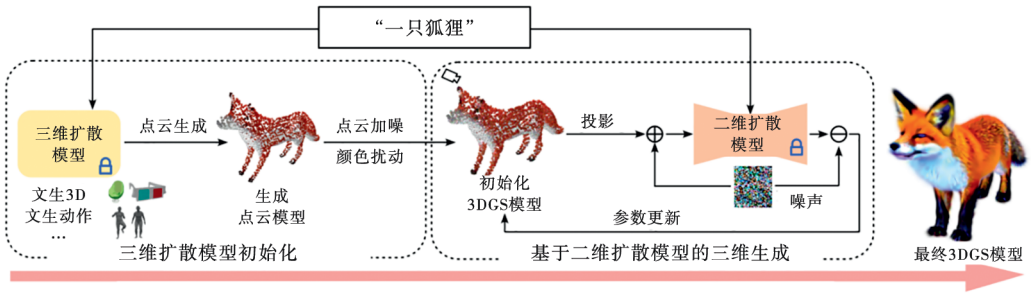


图3.5 GaussianDreamer技术流程图[31]

另外,改进二维扩散模型并令其更适合三维多视角合成的方法也备受关注。MVDream[33]则凭借更优的二维扩散模型DreamBooth[41],利用注意力机制对扩散模型进行多视角处理过程的优化,如图3.6。具体来说,针对多视角扩散模型的训练,同时采用基于二维注意力机制的图像训练损失模式和基于三维注意力和相机信息嵌入的多视角损失模式,对扩散模型进行训练;而针对DreamBooth的训练,可凭借训练出的多视角扩散模型指导优化,并同时对模型进行微调。采用这种方法,即使得扩散模型有了三维感知能力,在生成过程中扩散模型即对当前渲染的多视角结果有了三维感知能力。

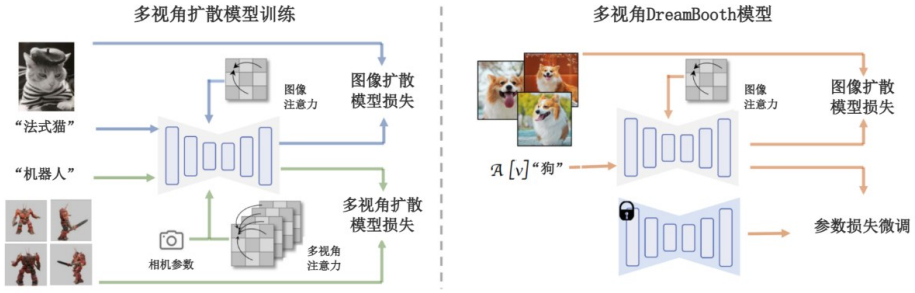


图3.6 MVDream技术流程图[33]

3.1.2 文本图像条件联合控制相关工作

由于三维模型本身为一种可视化模型,仅依靠文本指令难免会在主观上带来可控程度不够灵活的感受,同时由于三维表示-扩散-蒸馏损失方法中,根据文本指令进行生成梯度的是扩散模型部分,故仅依靠文本指令进行生成的质量容易受到扩散模型预训练的影响。故加入可视化的控制条件也是SDS机制应用在三维生成工作的一个重要方面。

虽然之前已经出现部分单视图合成的相关工作,但大部分均没有令人满意的泛化性和3D一致性,而Make-It-3D[28]则实现了高质量的单图像三维生成工作。

如图3.7所示,Make-It-3D采用一种两阶段的由粗至细的生成框架。与Magic3D[27]类似,在粗略生成阶段通过SDS优化NeRF,根据输入单视角图像生成三维对象的大致几何轮廓,并引入输入视角的参考视图像素损失L_{ref}和深度损失L_{depth}。其中,对于扩散模

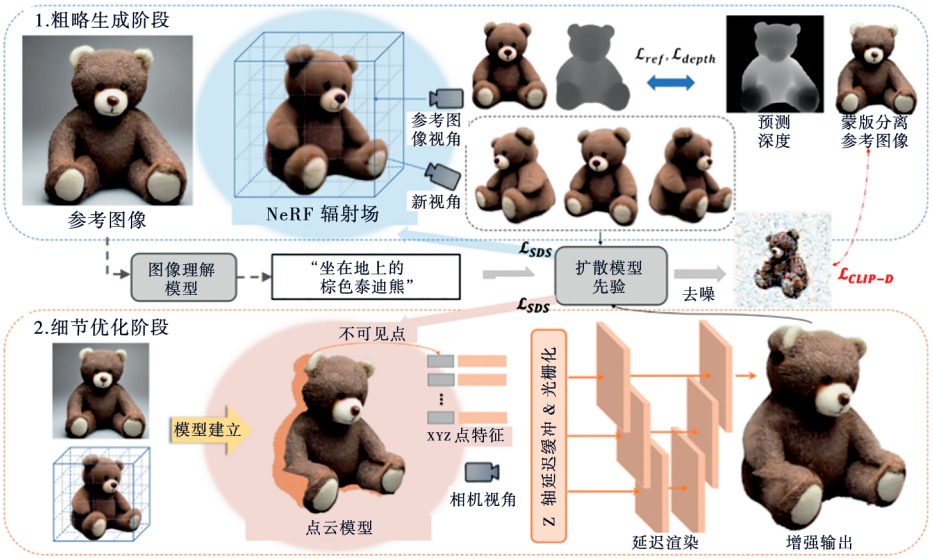


图 3.7 Make-It-3D 技术流程图[28]

型的去噪结果 $\hat{\boldsymbol{x}}_0^{\mathrm{gt}}$，利用 CLIP 图像编码器 $\varepsilon_{\mathrm{CLIP}}$ 计算其与参考图像 $\boldsymbol{x}_{\mathrm{ref}}$ 的匹配程度，以保证生成模型与参考图匹配，并采用对其 $L_{\mathrm{CLIP-D}}$ 进行监督，如式(3-2)所示：

$$L_{\mathrm{CLIP-D}}\left(\hat{\boldsymbol{x}}_0^{\mathrm{gt}}, g(\boldsymbol{\Theta})\right) = -\varepsilon_{\mathrm{CLIP}}\left(\boldsymbol{x}_{\mathrm{ref}}\right)\varepsilon_{\mathrm{CLIP}}\left(\hat{\boldsymbol{x}}_0^{\mathrm{gt}}\right) \quad (3\text{-}2)$$

而在细节阶段采用 NeRF 辐射场数据和参考图像构建纹理点云，并继续通过 SDS 进行优化。优化过程中的点云数据则通过带有缓冲区和光栅化的延迟渲染机制进行渲染，从而达到使用单视图和文本指令生成新三维对象及其新视图图片的效果。

除利用单视图进行三维扩展之外，很多利用文本和图像条件联合控制三维生成的工作也从通过图片控制生成对象的风格层面入手。受到个性化 T2I 工作的 ImagenT2I[37]的启发，DreamBooth3D[34]通过更新的 T2I 扩散模型 DreamBooth[41]结合 NeRF 与 SDS 实现了基于指定图像风格和文本的三维对象生成方法，如图 3.8 所示。

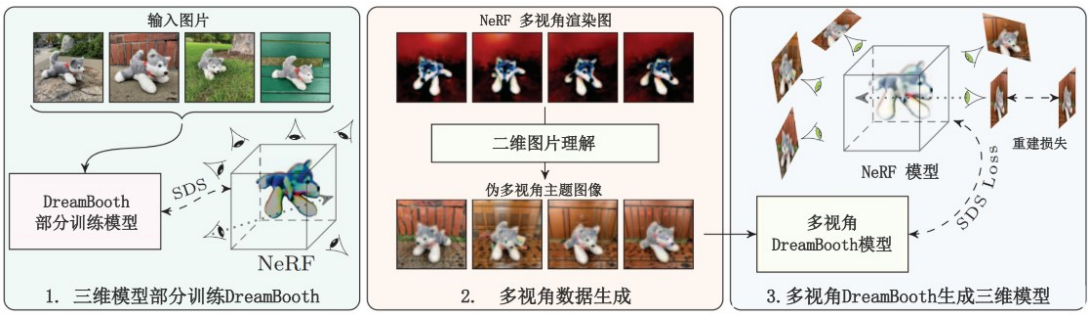


图 3.8 DreamBooth3D 技术流程图[34]

本工作分为了三个阶段，其中第一阶段利用输入参考风格图像对扩散模型 DreamBooth 进行部分训练，从语义上将特定风格的图片与文本指令建立隐性的联系，以保证扩散模型在去噪过程中向图像中蕴含的风格逼近，同时使用扩散模型通过 SDS 初始化 NeRF；第二阶段采用初始化的 NeRF 输出图像结合 DreamBooth 进行图对图转化过程，从而生成伪多视角主题图像，即生成了符合指定主题的三维对象的多视角图片，以对后续的生成过程提供进一步参照；而在第三阶段，利用伪多视角主题图像对 DreamBooth 进行微调，以确保其完全结合输入图片所指定风格，并使用 SDS 损失优化 NeRF 网络。另外，此工作同时引入重建损失和 NeRF 正则化损失[20]以避免 NeRF 导致的 3D 不一致问题，成功实现了高质量的基于图片风格和文本联合控制三维生成的效果。

其他工作也凭借类似方法实现了联合图片风格

与文本控制进行三维生成的效果,例如IPDreamer[32],通过Zero-1-to-3[42]对三维表示与扩散网络进行初始化,并采用交叉注意力机制处理图像指令,从而实现精确的风格生成;而在DiverseDream[35]中,为了解决三维表示-扩散-蒸馏损失的多样性低下问题,采用Hiper文本反转标签以增加更多信息,从而对于各种输入参考图片均可生成对应风格的三维对象,也相当于指定了当前生成对象的风格。

3.2 针对SDS在三维内容生成工作中的改进

虽然SDS通过搭起二维扩散模型与三维表示模型之间的桥梁,实现了文本控制或编辑三维内容的效果,但SDS本身具有很多不足与缺陷。针对三维内容处理过程,SDS所导致的问题主要聚焦于质量低下与多样性不足问题,而当前针对两问题也有众多论文进行改进,如图3.9所示。

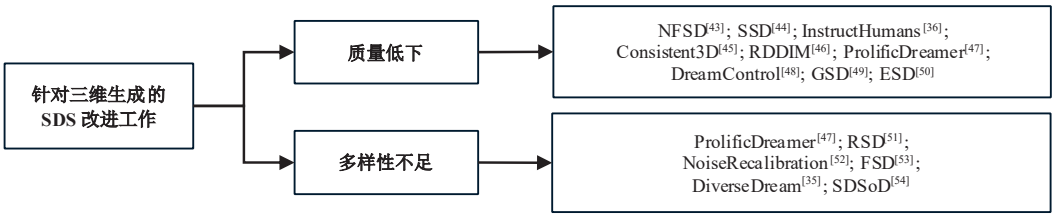


图3.9 针对三维生成的SDS改进相关工作归纳

3.2.1 针对质量低下问题的改进工作

基于三维表示模型和二维扩散模型在单独运用时都可产生高质量结果,基于三维表示-扩散-蒸馏损失框架的T-3D生成质量下降问题即定位至扩散损失与SDS对三维表示模型的梯度指导。

针对蒸馏损失所造成的平滑与纹理过度饱和,NFSD[43]方法则将问题的原因定位到SDS在梯度回传过程中对数据加入噪声部分的蒸馏作用,并通过阻止对噪声的蒸馏方式缓解纹理过度饱和问题。在此工作中,将SDS与CFG公式(见式(2-11))结合,将SDS分解为条件梯度δ_{C}、域矫正梯度δ_{D}和噪声消除梯度δ_{N},并通过去除δ_{N}并调整δ_{C}和δ_{D}的方式实现优化三维模型和生成图像的改善,如式(3-3)、式(3-4)和图3.10所示。

$$\nabla_{\boldsymbol{\theta}} L_{\mathrm{SDS}} = \omega(t)\left(\delta_{\mathrm{N}} + \delta_{\mathrm{D}} + s\delta_{\mathrm{C}} - \epsilon\right)\frac{\partial g(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \tag{3-3}$$

$$\nabla_{\boldsymbol{\theta}} L_{\mathrm{NFSD}} = \omega(t)\left(\delta_{\mathrm{D}} + s\delta_{\mathrm{C}}\right)\frac{\partial g(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \tag{3-4}$$

类似地,SSD[44]同样结合CFG,根据优化过程对最优数据分布的搜索过程将SDS损失分解为模式脱离项、模式搜索项与方差削弱项三部分,如式(3-5)所示,并对每一项进行梯度分析,指出过度平滑与过度饱和问题主要源于后两项的内在缺陷,在扩散与去噪过程中根据步长对三项进行调整,起到相较其他方法削弱过饱和现象、提升质量的效果,如图3.10所示。

$$\hat{\boldsymbol{\epsilon}}_{\phi}^{\mathrm{CFG}}\left(\boldsymbol{x}_t; y, t\right) - \boldsymbol{\epsilon} = \underbrace{\omega\left(\hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t; y, t\right) - \hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t; \varnothing, t\right)\right)}_{\text{模式脱离项}} + \underbrace{\hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t; y, t\right)}_{\text{模式搜索项}} - \underbrace{\boldsymbol{\epsilon}}_{\text{方差削弱项}} \tag{3-5}$$

另外,针对人体编辑的InstructHumans[36],针对SDS对文本条件y和原图条件I的处理,分离出baseline偏移项m_1、条件散度项m_2和完整约束项m_3,如式(3-6);同时本文中模拟了在去噪步长逐渐增加时三项对数据向目标分布驱动效果,如图3.11所示。InstructHumans工作指出,baseline偏移项仅会将参数优化过程向无条件状态的反方向驱动,而条件散度项和完整约束项则需要根据扩散去噪时间步长进行调控。通过分析,根据时间步长对编辑过程数据的影响,选择性调控编辑过程中三项的权重与调用方式,并配合更优的视角采样策略和平滑损失,避免人体编辑的质量低下现象。

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_{\phi}^{\mathrm{CFG}}\left(\boldsymbol{x}_t; y, t\right) - \boldsymbol{\epsilon} = {} & \underbrace{\left(\omega_I - 1\right)\left(\hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t; t, \varnothing, I\right) - \hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t; t, \varnothing, \varnothing\right)\right)}_{\text{baseline偏移项}} \\ & + \underbrace{\left(\omega_t - 1\right)\left(\hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t; t, y, I\right) - \hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t; t, \varnothing, I\right)\right)}_{\text{条件散度项}} \\ & + \underbrace{\hat{\boldsymbol{\epsilon}}_{\phi}\left(\boldsymbol{x}_t; t, y, I\right) - \boldsymbol{\epsilon}}_{\text{完整约束项}}\end{aligned} \tag{3-6}$$

图3.11中灰星代表无任何条件下数据分布点,黄星代表未编辑原图,蓝星代表文本条件所指定的数据分布点,红星代表满足原图和文本的目标分布点,而蓝线、橙线、绿线和红线分别代表baseline偏移项、条件散度项和完整约束项联合、条件散度项和完整约束项对数据的优化驱动方向。

以上凭借CFG对SDS进行分离建模的方法,通过数学推导对采用CFG模式的SDS进行深入分析,

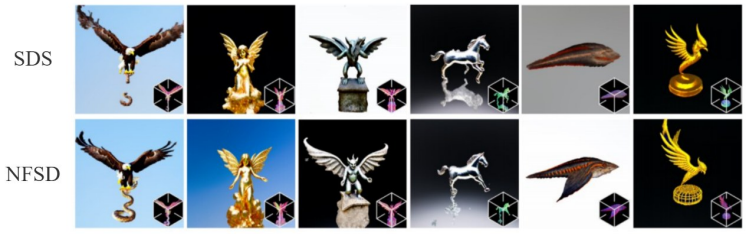


图3.10 NFSD（下排）对NeRF生成模型的提升

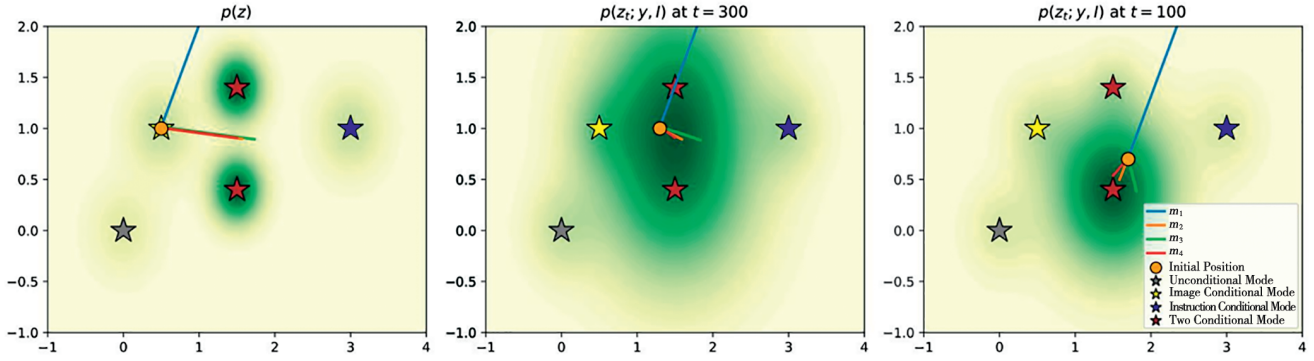


图3.11 InstructHumans模拟各项对数据优化驱动效果[36]

并在实现过程中通过调参的方式快捷有效地提升了SDS的优化效果。同时，基于SDS每步均采用随机高斯噪声进行加噪、去噪的准则，部分工作也结合微分方程推导对SDS进行优化。Consistent3D[45]工作将SDS的梯度优化采用随机微分方程（SDE, stochastic differential equations）进行表示，并指出方程中的随机项即为随机噪声的引入结果，并通过调整噪声采样的方式将梯度优化过程转化为常微分方程（ODE, ordinary differential equations），从而实现对SDS的改进；而RDDIM[46]则指出SDS的随机噪声策略导致在优化过程中引入了高方差数据，并受到DDIM[11]思想的启发，通过调整优化过程中的控制权重对三维模型进行由粗至细粒度的优化。

除对SDS进行数学层面的优化之外，ProlificDreamer[47]则对三维表示-扩散-蒸馏损失整体框架改进。基于T-3D过程中满足文本指令的三维表示模型参数可能不止一个，受到基于粒子的变分推理方法[55]的启发，ProlificDreamer将符合文本控制条件的NeRF参数建模为符合特定分布$\boldsymbol{\mu}$的粒子$\{\boldsymbol{\Theta}_i\}_{i=1}^{n}$，并在优化过程中通过采样$\boldsymbol{\Theta}_i$的方式对分布$\boldsymbol{\mu}$进行优化。同时，此工作也针对扩散去噪过程进行了优化，对整个去噪过程进行向带噪真实图像的噪声进行逼近优化，以实现扩散过程中分布的最优化逼近，此工作中对SDS的改进损失及梯度如式（3-7）和式（3-8）：

$$D_{\mathrm{KL}}\left(q_t^{\mu}\left(\boldsymbol{x}_t|\boldsymbol{c},y\right)\|\,p_t\left(\boldsymbol{x}_t|\boldsymbol{c},y\right)\right)=0\Leftrightarrow q_0^{\mu}\left(\boldsymbol{x}_0|\boldsymbol{c},y\right)=p_0\left(\boldsymbol{x}_0|\boldsymbol{c},y\right) \tag{3-7}$$

$$\frac{d\boldsymbol{\Theta}_\tau}{d\tau}=-\mathrm{E}_{t,\epsilon,\boldsymbol{c}}\left[\omega(t)\left(\underbrace{-\sigma_t\nabla_{\boldsymbol{x}_t}\log p_t\left(\boldsymbol{x}_t|\boldsymbol{c},y\right)}_{\text{带噪真实图像噪声分数}}-\underbrace{\left(-\sigma_t\nabla_{\boldsymbol{x}_t}\log q_t^{\mu_\tau}\left(\boldsymbol{x}_t|\boldsymbol{c},y\right)\right)}_{\text{带噪渲染图像噪声分数}}\right)\frac{\partial g\left(\boldsymbol{\Theta}_\tau\right)}{\partial\boldsymbol{\Theta}_\tau}\right] \tag{3-8}$$

进一步针对采用特定的预训练模型近似带噪真实图像加噪过程的噪声$\boldsymbol{\epsilon}_{pretrain}$，从而定义工作所提出的变分分数蒸馏（VSD, variational score distillation），如式（3-9）：

$$\nabla_{\boldsymbol{\Theta}}L_{\mathrm{VSD}}(\boldsymbol{\Theta})\triangleq\mathrm{E}_{t,\epsilon,\boldsymbol{c}}\left[\omega(t)\left(\boldsymbol{\epsilon}_{pretrain}\left(\boldsymbol{x}_t,t,\boldsymbol{c},y\right)-\boldsymbol{\epsilon}_{\phi}\left(\boldsymbol{x}_t,t,\boldsymbol{c},y\right)\right)\frac{\partial g(\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}}\right] \tag{3-9}$$

ProlificDreamer[47]同时利用粒子变分优化和改进的蒸馏机制VSD，实现了更高质量和更细节的生成，如图3.12所示，同时也通过建立分布使得多样性问题得以解决。此外VSD也被更多后续相关工作采用，例如DreamControl[48]工作就在VSD的基础上采用由粗至细的参数优化策略进行T-3D高质量生成。但后续相关工作也指出，VSD的T-3D效果仍会出现三维不一致的Janus伪影问题。

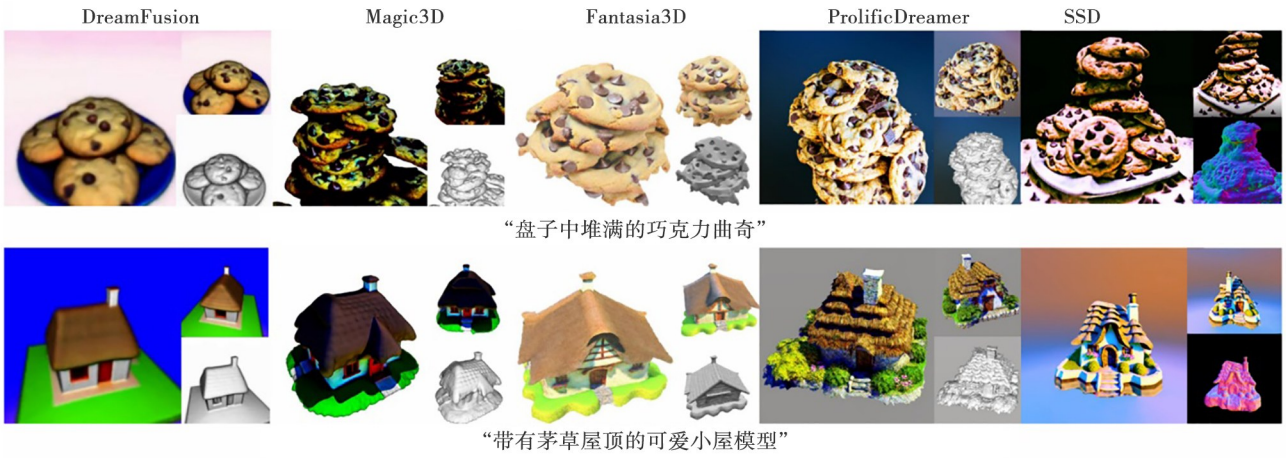


图3.12 SSD与ProlificDreamer相较其他方法的改进效果[44]

针对Janus伪影问题，GSD[49]工作指出其根本原因是多视图表示三维内容过程中，各个视角的二维图像预测分数不一致，故在其工作中对扩散过程进行优化，针对三维编辑工作采用具有三维一致性的噪声，以保证编辑过程各个视角图像与三维内容映射的一致性。另外，ESD[50]工作也将问题定位在VSD仍然存在的Janus伪影上，指出现有的蒸馏损失机制会对每个输入视图独立优化，并退化为寻求最大似然分布。因此，ESD方法在生成过程中对优化对象分布μ建立熵项，并利用熵项监督三维模型的渲染图像分布，通过最大化熵函数的方式促进不同视角渲染图像之间的差异化，从而缓解VSD方法中普遍存在的Janus伪影问题。

3.2.2 针对多样性低下问题的改进工作

基于SDS的T-3D生成工作，除质量低下以外，生成三维内容的多样性差也是目前探讨的主要问题之一。关键在于，单独的二维扩散模型进行生成任务时，不同生成轮次所生成的内容具有多样性，但基于SDS的T-3D生成工作则在每一次优化过程都倾向于同一种结果，故生成多样性低下的原因也聚焦于SDS。

从InstructHumans[36]中的模拟图3.11可看出，一般情况下可满足条件的数据分布不止一种；而ProlificDreamer[47]也因此将符合条件的参数化三维表示建模为多个模型构成的分布$\boldsymbol{\mu}$，并将其作为蒸馏损失的优化对象，每轮训练均从$\{\boldsymbol{\Theta}_i\}_{i=1}^{n}$中采样$\boldsymbol{\Theta}_i$进行优化，因此ProlificDreamer工作所优化出的三维内容可以具有多样性，如图3.13所示。

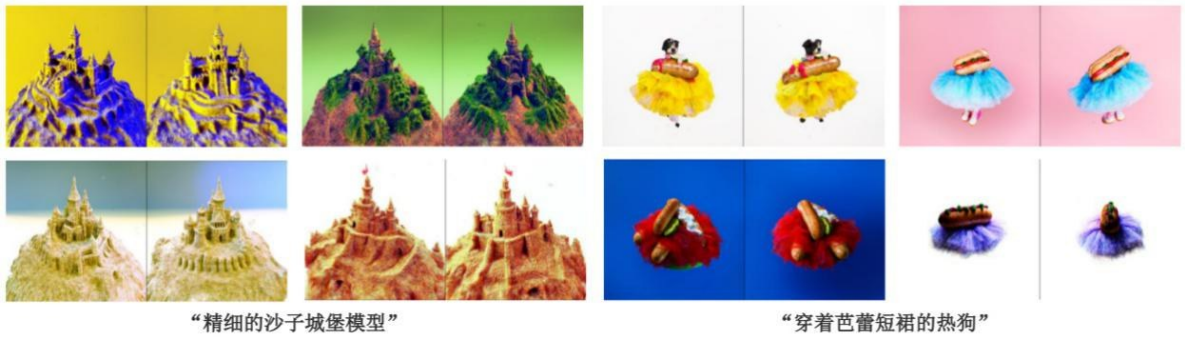


图3.13 ProlificDreamer工作的多样性结果[47]

类似地，RSD[51]工作也利用了基于粒子分布的变分过程，通过多尺度正则化扩散过程和多样性排斥两种正则化变分算法，使用允许粒子之间相互作用的排斥正则化过程以促进优化结果的多样性，实现可即插即用的排斥分数蒸馏损失(RSD, repulsive score distillation)。ProlificDreamer与RSD工作均证明了满足文本指令的三维模型参数并不是唯一的，并在SDS生成多样性低下的问题方面提供了解决思路。

而另一方面，也有部分论文将改进重点聚焦于优化过程中每轮重新随机采样噪声的机制。NFSD[43]工作中就指出，低多样性也是随机噪声经过蒸馏的结

果，采用具有规律性、原则性的噪声调度会增加多样性。NoiseRecalibration[52]工作通过限制生成过程为采样单个高斯噪声，以产生多样性结果，但在单个高斯噪声的控制下无法保证生成质量。

FSD[53]工作也将问题聚焦于噪声上，并通过实验证明扩散模型中数据的改变方向受到扩散过程加入噪声的影响，也推断导致多样性问题的原因在于每轮不同的噪声导致优化方向平均化。因此，此工作在扩散过程采用了一种针对多视图的噪声指定方式，针对不同视角规定了一种随视角变化的噪声采样规则，但保证同样视角图片采用相同噪声进行扩散；同时将图片中的前景与背景噪声分割，保证针对三维内容本体的处理质量。另外，在本文的分析过程中，与Consistent3D[45]和RDDIM[46]相似，通过分析噪声引入的随机量，将SDS更新机制与DDIM[11]相类比，并将其近似为扩散概率流常微分方程（PF-ODE, probability flow ODE），从而减少随机噪声带来的随机量，以提高多样性，如图3.14所示。

图3.14 FSD工作多样性结果[53]

除噪声控制之外，DiverseDream[35]工作也通过引入额外模块提升生成多样性。通过文本反转HiperInversion模块[56]，通过对多样的输入参考风格图片加入文本反转标签，从而控制扩散模型根据参考图片风格生成三维内容，实现通过图片控制风格的同时提升模型可生成内容的多样性。

除质量与多样性的问题之外，基于SDS的T-3D方法仍有很多其他问题和改进工作，如针对高计算复杂度和低效率问题的Instant3D[57]与SDSoD[54]工作。

4 相关工作成果

在本节，将阐述基于SDS的三维生成工作的主要成果，包括当前基于SDS的T-3D工作的相关数据集、评价指标以及主要工作的成果。

4.1 相关数据集

在基于SDS进行三维内容重建与编辑工作中，三维表示模型和扩散模型均采用预训练模型，故当前此类工作的数据集运用主要在测试模型生成效果。同时，由于T-3D工作是近一两年较为新兴的工作类型，且由于完整的三维场景与物体数据难以捕捉，因此目前仍未出现较完整、较全面的文本与三维内容的数据集。故本小节将对当前T-3D工作中主要涉及的文本-图像数据集和三维内容数据集进行简要概括。

文本-图像数据集得益于以CLIP和DALL-E为首的视觉语言架构的快速发展与广泛应用，文本-图像工作对大规模数据集有很大需求。在这种背景下，众多较大规模的文本-图像数据集推动了视觉语言模型的发展，而其中LAION-5B数据集[26]凭借其大规模和高全面度成为视觉语言模型训练和评价普遍使用的数据集。

LAION-5B数据集由约58亿个图像文本组成，其中23亿为图像-英文文本对，22亿为图像数据，其余10亿为不限语言的图像与文本对。此数据集在此类型数据集数据规模最大的同时，也具有很高的数据多样性，包含各种领域的图像内容，同时也包含水印图片和异常图片，在众多相关的研究方向均具有可用性。同时，由于数据集规模较大，LAION数据集还具有很多可用的子集，如LAION-400M，保证数据多样性与质量的同时保证了在图像生成、文本生成领域的训练效率。

而在DreamFusion[7]的工作中，扩散模型部分的StableDiffusion预训练模型即为LAION数据集的子集进行预训练，采用此数据集预训练的扩散模型部分也运用到了Magic3D[27]、Fantasia3D[30]、GaussianDreamer[31]等工作中。

三维内容数据集因T-3D工作训练流程主要是特

定文本指令的三维表示模型生成过程,模型效果也采用定性与定量结果综合评价,故很多工作针对三维表示模型并没有特定的带有标签的训练用数据集。但部分工作也运用了特定的三维内容数据集进行模型的微调或评价。

其中DTU MVS数据集[58]为丹麦技术大学提供的数据集,主要目的为针对以NeRF为首的多视图立体工作的评估数据集。此数据集主要用于评估三维内容多视图生成的无偏见评估,数据由80个不同场景组成,其中59个场景包含49个摄像机位置,21个场景包含64个摄像机位置,并以1200×1600分辨率进行存储。此数据集针对基于多视图的三维内容合成质量进行评估,通过三维内容重建点的准确度和完整性评估三维表示质量。此数据集在Make-It-3D[28]工作中充当了评估生成与编辑质量的数据指标。

目前也存在具有代表性的文本-三维内容数据集,如Objaverse[59]数据集,主要由华盛顿大学提出,截止论文发表时已经包含超过80万个带有标题、标签描述和动画的三维模型,如图4.1所示。MV-Dream[33]与基于Zero-1-to-3[42]的IPDreamer[32]也采用了此数据集或其预训练模型实现基于文本的三维数据合成。但此数据集因三维模型权限和NoAI标签的缺失遇到问题,可能是此数据集目前未被大规模使用的原因。

另外,更多的工作也采用了自建数据集的方式,Mip-NeRF 360[20]则通过自建数据集实现了模型训练与评估,同时数据集也运用在了多个工作中。

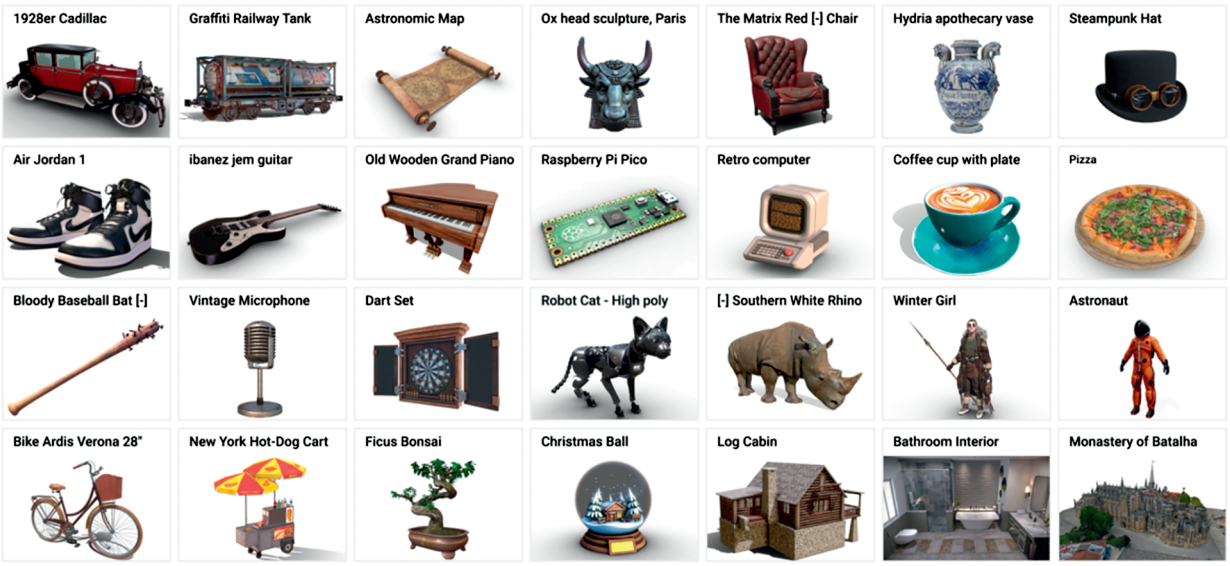


图4.1 Objaverse数据集[59]

4.2 评价指标

针对当前现有的T-3D工作,大部分对三维内容的表示方法仍是新视角下的图片输出。因此针对模型输出成果的质量评估,当前相关工作的评价指标基本与文本-图像相关工作的评价指标相同,例如衡量数据之间相似度的CLIP分数[60]、评价生成图像质量的IS指标[61]和FID指标[62],但大部分相关工作所涉及评估指标并不统一。

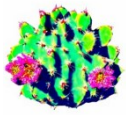
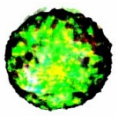
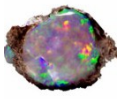
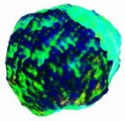
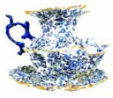
近期出现了针对T-3D工作的评估标准T^{3}Bench[63]。T^{3}Bench为首个全面的文本生成三维内容的评价基准,通过同时评估结果质量和文本对齐程度全面评估T-3D工作。其中质量指标结合多视图的文本图像分数和对结果的区域卷积结果,对生成结果质量进行评估;而对齐指标则凭借多视图抓取和大预言模型GPT-4评估文本-三维内容一致性。在T^{3}Bench原文中[63],就对当前有代表性的T-3D工作进行了数据归纳;而在工作GaussianDreamer[31]中,作者就采取了T^{3}Bench对模型进行评估。

4.3 代表性工作成果

由于目前的T-3D工作中所运用的数据集与评价指标均不相同,此部分仅整理主要相关工作的实验结果。

定性结果 在三维内容生成方面,由于各个工作进行评估时采用指令不同,故采用指标T^{3}Bench中的定性结果进行列举,如表4.1所示。

表 4.1 主要三维生成工作定性结果对比

	DreamFusion[7]	Magic3D[27]	LatentNeRF[29]	Fantasia3D[30]	MVDream[33]	ProlificDreamer[47]
“开着粉花的仙人掌”						
“光滑的圆形鹅卵石”						
“蓝白相间的陶瓷茶杯”						
“绿色田野中的红色仓库”						

定量结果 由于三维内容生成方面的各个工作采用的数据指标不同，且应用领域也不同，因此针对主要工作采用 T^{3}Bench 指标评估[63]，结果如表 4.2 所示，表中的数据越大表明模型生成效果越佳。

根据结果，不难看出当前基于 SDS 的三维内容生成方法普遍存在颜色过饱和现象。虽然采用更新的或解耦的三维表示方法可以略微改善，但仍然存在几何形状与表面纹理不合理的普遍现象。然而，利用多视图化扩散模型的方法 MVDreamer[33]和改进蒸馏损失优化机制方法 ProlificDreamer[47]获得了相对较好的定性结果与定量结果，也印证了改进扩散模型和蒸馏损失机制对提升质量具有较大的作用。

表 4.2 主要三维生成工作 T^{3}Bench 指标[63]

	单物体指标	带有背景单物体指标	多物体指标	指标平均值
DreamFusion[7]	24.4	24.6	16.1	21.7
Magic3D[27]	37.0	35.4	25.7	32.7
LatentNeRF[29]	33.1	30.6	20.6	28.1
Fantasia3D[30]	26.4	27.0	18.5	24.0
ProlificDreamer[47]	49.4	44.8	35.8	43.3
MVDream[33]	47.8	42.4	33.8	41.3

5 总结与展望

随着扩散模型与三维表示模型的发展，文本控制三维内容生成的方法已经成为计算机图形学领域的重要组成部分，同时也对文化内容生产与传播产生了重大影响。本文则从技术方法归纳入手，对基于 SDS 的三维内容生成工作进行了简要汇总，在介绍了扩散模型、三维表示方法与分数蒸馏损失之后，归纳总结了具有代表性的相关应用与改进工作。同时，本文也整理了目前此研究领域主要的数据集与评价指标，并对主要工作成果进行了简单汇总。

随着更多基于 SDS 的 T-3D 工作出现和改进，通过条件控制生成的三维内容必定会朝着更高精度和更高质量方向进行。而高精准度、高质量与高效率的三维生成工作可能会对众多领域产生影响。

近年来,文生图像、文生视频技术在各个娱乐媒体平台已经掀起了巨大波澜,而由于文生三维内容技术需要较大的计算资源,目前还没有出现代表性的应用消费级产品。然而,部分内容生成工具如有言等产品已经推出了文生三维动画的功能,结合云演艺平台与虚实结合等相关领域的蓬勃发展,高效率和高质量的场景迁移、演员生成和演出控制等必定是虚拟演艺领域的有效催化剂。另外,从2023年泰勒·斯威夫特全球巡回演唱会中所运用的三维投影技术也可看出,三维技术在当前实体演出中也得到了广泛应用,因此高自由度的文生三维内容技术必定会令舞台复制、硬件模拟等过程变得更加高效。

除娱乐领域之外,文生三维内容的发展也将对包括元宇宙在内的其他领域产生重大影响。自2020年元宇宙概念爆火以来,国家已出台相关政策,对元宇宙概念与相关技术在基建、商业和农业等领域的应用进行引导。随着高质量与高效率的条件控制三维生成的普及,一定会进一步提高元宇宙技术应用的普及性,并快速推动虚拟现实(virtual reality)和增强现实(augmented reality)等技术在教育、医疗等领域工作的普及。更进一步讨论,更高精准度和更高质量的三维内容快速生成,也有望应用在军事和安防等领域的区域模拟、危险监测等功能中,以提升相关工作的安全性和效率。

然而基于当前此领域的已有成果与应用设想,对扩散模型与分数蒸馏采样损失的进一步改进是必经之路。同时,CDM工作[64]与DiverseDreamer[35]工作也提供了采用文本嵌入(text embedding)方法,对文本指令进行处理以得到更优化效果;而针对此工作的通用生成工具界面也是推动此工作走向应用的重要方法。因此在对模型本身进行改进的同时,结合更优的指令处理方法和使用门槛更低的开发应用,定可大幅推动文本控制三维内容生成相关工作的发展和应用。相信更富想象力的、更实用的高效、高质量、高精度三维内容生成方案将在不久的将来成功实现。

参考文献(References):

[1] 欧阳宏.故宫院藏文物的三维数据采集与应用[J].数字图书馆论坛,2019(07):48-53.

[2] 何威,牛雪莹.数字游戏开展中华传统文化国际传播的趋势、方式与特点[J].对外传播,2022(09):21-25.

[3] 陈孟,曹建峰,易镁金.文化科技十大前沿应用趋势[R].北京:腾讯研究院,2023.

[4] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684-10695.

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

[6] Song J, Meng C, Ermon S. Denoising diffusion implicit models[DB/OL]. arXiv:2010.02502, 2020.

[7] Poole B, Jain A, Barron J T, et al. DreamFusion: text-to-3D using 2D diffusion[DB/OL]. arXiv:2209.14988, 2022.

[8] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[9] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[C]// Proceedings of 18th International Conference on Medical Image Computing and Computer-assisted Intervention, 2015: 234-241.

[10] Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic differential equations [DB/OL]. arXiv:2011.13456, 2020.

[11] Song J, Meng C, Ermon S. Denoising diffusion implicit models[DB/OL]. arXiv:2010.02502, 2020.

[12] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684-10695.

[13] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of International Conference on Machine Learning, 2021: 8748-8763.

[14] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: representing scenes as neural radiance fields for view synthesis [J]. Communications of the ACM, 2021, 65(1): 99-106.

[15] Kerbl B, Kopanas G, Leimkühler T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Transactions on Graphics, 2023, 42(4), 139:114.

[16] 韩开,徐娟.3D场景渲染技术-神经辐射场的研究综述[J].计算机应用研究,2024,41(08):2252-2260.

[17] Max N. Optical models for direct volume rendering[J]. IEEE Transactions on Visualization and Computer Graphics, 1995, 1(2): 99-108.

[18] Müller T, Evans A, Schied C, et al. Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM Transactions on Graphics (TOG), 2022, 41(4): 1-15.

[19] Barron J T, Mildenhall B, Tancik M, et al. Mip-nerf: a multiscale representation for anti-aliasing neural radiance fields [C]// Proceedinds of the IEEE/CVF International Conference on Computer Vision, 2021: 5855-5864.

［20］ Barron J T，Mildenhall B，Verbin D，et al. Mip-nerf 360：unbounded anti-aliased neural radiance fields［C］// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2022：5470-5479.

［21］ Xu Q，Xu Z，Philip J，et al. Point-nerf：point-based neural radiance fields［C］// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2022：5438-5448.

［22］ Green R. Spherical harmonic lighting：the gritty details［C］// Proceedings of the Game Developers Conference，2003，56：4.

［23］ Jain A，Mildenhall B，Barron J T，et al. Zero-shot text-guided object generation with dream fields［C］// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2022：867-876.

［24］ Cai R，Yang G，Averbuch-Elor H，et al. Learning gradient fields for shape generation［C］// Proceedings of 16th European Conference on Computer Vision，2020：23-28.

［25］ Chandraker K G M. Neural mesh flow：3D manifold mesh generation via diffeomorphic flows［DB/OL］. arXiv：2007.10973，2020.

［26］ Schuhmann C，Beaumont R，Vencu R，et al. Laion-5b：an open large-scale dataset for training next generation image-text models［J］. Advances in Neural Information Processing Systems，2022，35：25278-25294.

［27］ Lin C H，Gao J，Tang L，et al. Magic3D：high-resolution text-to-3D content creation［C］// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2023：300-309.

［28］ Tang J，Wang T，Zhang B，et al. Make-it-3D：high-fidelity 3D creation from a single image with diffusion prior［C］// Proceedings of the IEEE/CVF International Conference on Computer Vision，2023：22819-22829.

［29］ Metzer G，Richardson E，Patashnik O，et al. Latent-nerf for shape-guided generation of 3D shapes and textures［C］// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2023：12663-12673.

［30］ Chen R，Chen Y，Jiao N，et al. Fantasia3D：disentangling geometry and appearance for high-quality text-to-3D content creation［C］// Proceedings of the IEEE/CVF International Conference on Computer Vision，2023：22246-22256.

［31］ Yi T，Fang J，Wang J，et al. Gaussiandreamer：fast generation from text to 3D gaussians by bridging 2D and 3D diffusion models［C］// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2024：6796-6807.

［32］ Zeng B，Li S，Feng Y，et al. Ipdreamer：appearance-controllable 3D object generation with image prompts［DB/OL］. arXiv：2310.05375，2023.

［33］ Shi Y，Wang P，Ye J，et al. Mvdream：multi-view diffusion for 3D generation［DB/OL］. arXiv：2308.16512，2023.

［34］ Raj A，Kaza S，Poole B，et al. Dreambooth3D：subject-driven text-to-3D generation［C］// Proceedings of the IEEE/CVF International Conference on Computer Vision，2023：2349-2359.

［35］ Tran U D，Luu M，Nguyen P，et al. DiverseDream：diverse text-to-3D synthesis with augmented text embedding［DB/OL］. arXiv：2312.02192，2023.

［36］ Zhu J，Yang L，Yao A. InstructHumans：editing animated 3D human textures with instructions［DB/OL］. arXiv：2404.04037，2024.

［37］ Saharia C，Chan W，Saxena S，et al. Photorealistic text-to-image diffusion models with deep language understanding［J］. Advances In Neural Information Processing Systems，2022，35：36479-36494.

［38］ Shen T，Gao J，Yin K，et al. Deep marching tetrahedra：a hybrid representation for high-resolution 3D shape synthesis［J］. Advances in Neural Information Processing Systems，2021，34：6087-6101.

［39］ McAuley S，Hill S，Hoffman N，et al. Practical physically-based shading in film and game production［C］// ACM SIGGRApH 2012 Courses，2012：1-7.

［40］ Tang J，Ren J，Zhou H，et al. Dreamgaussian：generative gaussian splatting for efficient 3D content creation［DB/OL］. arXiv：2309.16653，2023.

［41］ Ruiz N，Li Y，Jampani V，et al. Dreambooth：fine tuning text-to-image diffusion models for subject-driven generation［C］// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2023：22500-22510.

［42］ Liu R，Wu R，Van Hoorick B，et al. Zero-1-to-3：zero-shot one image to 3D object［C］// Proceedings of the IEEE/CVF International Conference on Computer Vision，2023：9298-9309.

［43］ Katzir O，Patashnik O，Cohen-Or D，et al. Noise-free score distillation［DB/OL］. arXiv：2310.17590，2023.

［44］ Tang B，Wang J，Wu Z，et al. Stable score distillation for high-quality 3D generation［DB/OL］. arXiv：2312.09305，2023.

［45］ Wu Z，Zhou P，Yi X，et al. Consistent3D：towards consistent high-fidelity text-to-3D generation with deterministic sampling prior［C］// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2024：9892-9902.

［46］ Lukoianov A，Borde H S O，Greenewald K，et al. Score distillation via reparametrized DDIM［DB/OL］. arXiv：2405.15891，2024.

［47］ Wang Z，Lu C，Wang Y，et al. Prolificdreamer：high-fidelity and diverse text-to-3D generation with variational score distillation［J］. Advances in Neural Information Processing Systems，2024，36.

［48］ Huang T，Zeng Y，Zhang Z，et al. Dreamcontrol：control-based text-to-3D generation with 3D self-prior［C］// Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition, 2024: 5364-5373.

[49] Kwak M S, Ahn D, Kim I H, et al. Geometry-aware score distillation via 3D consistent noising and gradient consistency Modeling[DB/OL]. arXiv:2406.16695, 2024.

[50] Wang P, Xu D, Fan Z, et al. Taming mode collapse in score distillation for text-to-3D generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 9037-9047.

[51] Zilberstein N, Mardani M, Segarra S. Repulsive score distillation for diverse sampling of diffusion models[DB/OL]. arXiv:2406.16683, 2024.

[52] Yang X, Liu F, Xu Y, et al. Diverse and stable 2D diffusion guided text to 3D generation with noise recalibration [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(7): 6549-6557.

[53] Yan R, Wu K, Ma K. Flow score distillation for diverse text-to-3D generation[DB/OL]. arXiv:2405.10988, 2024.

[54] Cheng Y, Yin F, Huang X, et al. Efficient text-guided 3D-aware portrait generation with sore distillation sampling on distribution[DB/OL]. arXiv:2306.02083, 2023.

[55] Liu Q, Wang D. Stein variational gradient descent: a general purpose bayesian inference algorithm[J]. Advances in Neural Information Processing Systems, 2016, 29.

[56] Han I, Yang S, Kwon T, et al. Highly personalized text embedding for image manipulation by stable diffusion[DB/OL]. arXiv:2303.08767, 2023.

[57] Li M, Zhou P, Liu J W, et al. Instant3D: instant text-to-3D generation[J]. International Journal of Computer Vision, 2024: 1-17.

[58] AanæsH, Jensen R R, Vogiatzis G, et al. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 2: 1-16.

[59] Deitke M, Schwenk D, Salvador J, et al. Objaverse: a universe of annotated 3D objects[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 13142-13153.

[60] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of International Conference on Machine Learning, 2021: 8748-8763.

[61] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans[J]. Advances in Neural Information Processing Systems, 2016, 29.

[62] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in Neural Information Processing Systems, 2017, 30.

[63] He Y, Bai Y, Lin M, et al. T $^ 3$ Bench: benchmarking current progress in text-to-3D generation[DB/OL]. arXiv: 2310.02977, 2023.

[64] Liu N, Li S, Du Y, et al. Compositional visual generation with composable diffusion models[C]// Proceedings of European Conference on Computer Vision, 2022: 423-439.

编辑:赵志军