

硕 士 研 究 生 读 书 报 告



题目 ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation读书报告

作者姓名 赵成

作者学号 22351273

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 2024年1月

# 摘要

分数蒸馏采样（SDS）通过蒸馏预训练的大规模文本到图像扩散模型，在文本到3D生成中获得了很好的结果，但其存在过饱和、过平滑和低多样性问题。在这篇论文中，作者将3D参数建模为随机变量，而不是SDS中的常数，并提出变分分数蒸馏（VSD），这是一个基于粒子的变分框架，用于解释和解决文本到3D生成中的上述问题。论文中提出的方法被称为ProlificDreamer，可以生成高渲染分辨率（512 × 512）和具有丰富结构和复杂效果的高保真NeRF（例如，烟和水滴）。

**关键词：**文本生成3D，分数蒸馏采样，NeRF，生成模型

Abstract

Score distillation sampling (SDS) has shown great promise in text-to-3D generation by distilling pretrained large-scale text-to-image diffusion models, but suffers from over-saturation, over-smoothing, and low-diversity problems. In this paper, the authors model 3D parameters as random variables instead of constants in SDS and propose Variational Score Distillation (VSD), a particle-based variational framework for interpreting and solving the aforementioned problems in text-to-3D generation. The method in the paper, called ProlificDreamer, generates high rendering resolution (512 × 512) and high-fidelity NeRFs with rich structure and complex effects (e.g., smoke and water droplets).

**Keywords：**Text to 3D, Score Distillation Sampling, NeRF, Generative Model

1 引言

3D内容和技术使我们能够可视化，理解并与反映我们现实生活体验的复杂对象和环境进行交互。它们的关键作用延伸到广泛的领域，包括建筑，动画，游戏以及快速发展的虚拟和增强现实领域。尽管其有广泛的应用，优质3D内容的制作往往仍然是一项艰巨的任务。它需要大量的时间和精力投入，即使是由专业设计师进行。这一挑战促使了文本到3D方法的发展。通过基于文本描述自动生成3D内容，这些创新方法为简化3D内容创建过程提供了一种有前途的方法。此外，它们还将使这一进程更容易实现，有可能鼓励上述领域的重大范式转变。

本篇论文首先介绍了变分分数蒸馏（VSD），它将相应的3D场景作为一个随机变量，而不是像SDS中那样将文本提示作为一个单一的点。VSD优化了3D场景的分布，使得从所有视图渲染的图像上诱导的分布在KL散度方面尽可能接近地与预训练的2D扩散模型定义的分布对齐。在这种变分公式下，VSD自然地表征了多个3D场景可能与一个提示对齐的现象。为了有效地解决这个问题，VSD采用了基于粒子的变分推理，并将一组3D参数作为粒子来表示3D分布。论文通过Wasserstein梯度流推导出一种新的基于梯度的粒子更新规则，并且保证当优化收敛时，粒子将是来自期望分布的样本。论文的更新需要估计扩散渲染图像上分布的得分函数，这可以通过预训练扩散模型的低秩自适应（LoRA）有效地实现。最终算法交替更新粒子和得分函数。

论文进一步系统地研究了与文本到3D的算法正交的其他元素，并提出了一个清晰的设计方法。具体来说，论文提出了一个渲染分辨率为512 × 512的NeRF训练和退火时间表，以提高生成的视觉质量。论文的整体方法可以生成高保真和多样化的3D结果，称之为ProlificDreamer。ProlificDreamer可以生成512 × 512渲染分辨率和高保真神经辐射场，具有丰富的结构和复杂的效果（例如，烟和水滴）。此外，ProlificDreamer还首次在文本提示下，在360度视图中成功构建了多个对象的复杂场景。此外，从生成的NeRF来进行初始化，ProlificDreamer可以生成细致入微和照片般逼真的3D纹理网格。

# 2方法

## 2.1方法概述

ProlificDreamer的整体结构如下所示。

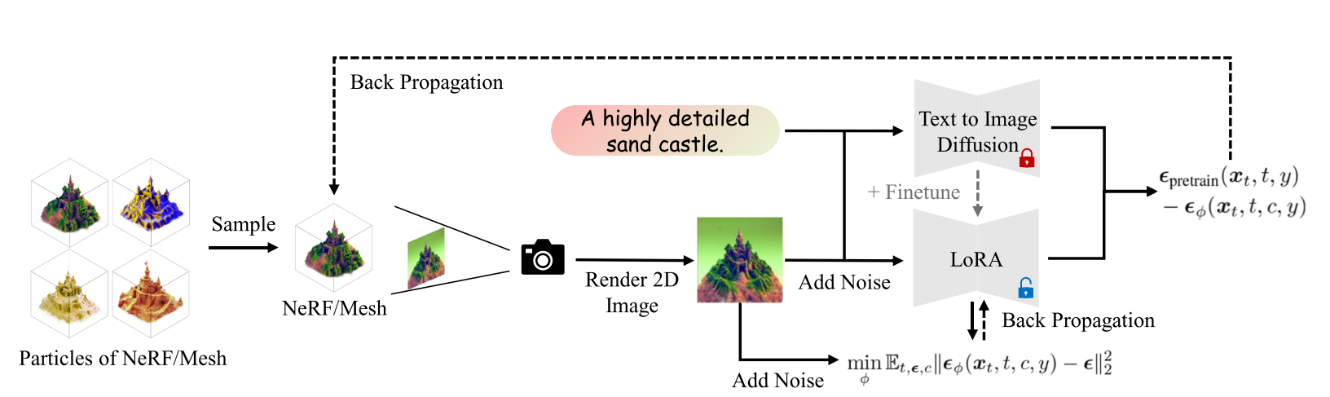
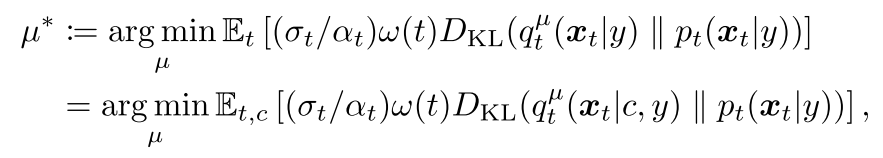


图1 ProlificDreamer架构图

ProlificDreamer的整体训练流程与DreamFusion类似。首先通过对初始化的NeRF进行采样得到渲染图像，对图像进行加噪，并通过预训练的Diffusion模型预测噪声进行反向传播来更新NeRF。但与DreamFusion的SDS不同的是，VSD的优化目标是让场景的渲染图像分布靠近预训练扩散模型的图像分布，SDS是VSD只有一个粒子的特例。因此，ProlificDreamer使用了多个粒子来建模图像分布，其中对每个NeRF的训练视为从分布中采样，这样做提高了生成结果的多样性。为了估计扩散渲染图像上分布的得分函数，ProlificDreamer使用LoRA来实现。

VSD的3D场景被建模为概率密度分布，是渲染图像在相机和渲染函数下的分布。要使渲染图像与扩散模型图像分布相同，可以使用KL散度：

直接求解是很困难的，VSD受扩散模型的启发，构建了一系列由索引的具有不同扩散分布的优化问题：

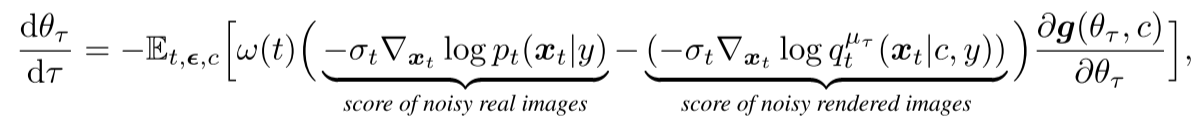


对于模型的训练，VSD使用个参数，每个参数称为一个粒子，

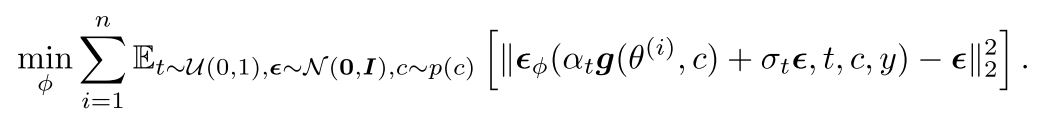
表示当前分布，如果优化收敛，则将是来自最优分布的样本。VSD通过迭代优化粒子直至收敛。

## 2.2 VSD（变分分数蒸馏）更新规则

VSD通过模拟下列常微分方程来模拟粒子的更新：



真实图像噪声的分数函数可以用预训练的扩散模型来近似，渲染图像噪声则需要由另一个噪声预测网络来估计，其标准扩散目标为



实践上，论文使用一个小型U-Net或一个LoRA来参数化预训练，并将额外的相机参数c添加到网络中的条件嵌入。在大多数情况下，论文发现使用LoRA可以大大提高所获得样本的保真度。论文认为，这是因为LoRA是为高效的few-shot微调而设计的，并且可以利用图像预训练中的先验信息、。

## 2.3 与分数蒸馏采样（SDS）的比较

本篇论文主要与SDS进行了以下几个方面的比较：

**SDS是VSD的一个特例**。SDS是VSD的一种特殊情况，它采用单点Dirac分布作为变分分布。特别地，VSD不仅潜在地采用多个粒子，而且即使对于单个粒子也学习参数评分函数，提供了优于SDS的上级推广。此外，通过使用LoRA，VSD还可以在估计参数中利用文本提示y，而SDS中使用的高斯噪声参数不能利用来自y的信息。

**VSD对CFG是友好的**。由于VSD的目标是从预训练模型定义的最佳分布中采样粒子，因此通过VSD在3D样本的预训练模型中调整CFG的效果与传统对2D样本采样的效果非常相似。因此，VSD可以像经典的文本到图像方法一样灵活地调整CFG，并且论文使用与常见的文本到图像生成任务相同的CFG设置（例如7.5）以获得最佳性能。

**VSD与SDS在隔离3D表示的2D实验中的比较**。为了直接比较SDS和VSD，论文考虑了渲染函数的特殊情况，以将优化算法与3D表示解耦。在这种情况下，优化参数等同于在2D空间中生成图像，从而独立于3D表示。论文显示了不同采样方法的结果。SDS在小和大的CFG权重下都表现出失效。特别是对于默认的CFG权重（即100)，SDS中使用的2D样本具有之前在文本到3D生成中观察到的相同问题，例如过饱和和过平滑。相比之下，VSD展示了适应各种CFG权重的灵活性，并且使用正常的CFG权重（即7.5），其表现类似于2D的扩散模型。

## 2.4 3D表征与训练

**用于NeRF训练的高分辨率渲染**。论文选择Instant NGP 进行高效的高分辨率渲染，并使用VSD优化NeRF高达512的训练分辨率。具体来说，论文发现SDS是在先前的工作中优化NeRF的主要瓶颈之一。相反，通过应用VSD，论文可以获得分辨率从64到512的高保真NeRF。

**NeRF训练的场景初始化**。对于以对象为中心的场景，论文遵循Magic 3D 中使用的以对象为中心的初始化；对于复杂场景，论文建议通过设置参数使密度“中空”，通过包围相机来初始化场景。场景初始化可以帮助生成高保真的复杂场景，而无需对现有算法进行其他修改。

**分数蒸馏退火时间表**。论文在分数蒸馏目标中利用时间步长t的简单两阶段退火，同时适用于SDS和VSD。对于前几个步骤，论文对时间步长（0.02，0.98）进行采样，然后退火到（0.02，0.50）。本质上，论文的目标是匹配原始分布和目标分布。大的时间步的KL散度可以在训练的早期阶段提供合理的优化方向。在训练期间，当x接近p0（x）时，较小的t可以缩小分布之间的差距，并提供更加一致的详细信息。

**网格表示和微调**。论文采用了一个基于坐标的哈希网格编码器继承NeRF阶段表示的网格纹理。论文根据Fantasia3D来首先使用法线贴图优化几何形状，然后优化纹理来优化几何形状和纹理。在最初的实验中，论文发现用VSD优化几何形状并不比使用SDS提供更多的细节。这可能是因为网格分辨率不足以表示高频细节。因此，论文用SDS优化几何结构以提高效率。但与Fantasia3D不同的是，论文的纹理优化由VSD监督，设置CFG = 7.5，可以提供比SDS更多的细节。

# 3 实验与结果

图2显示了这篇文章使用ProlificDreamer方法生成的结果。

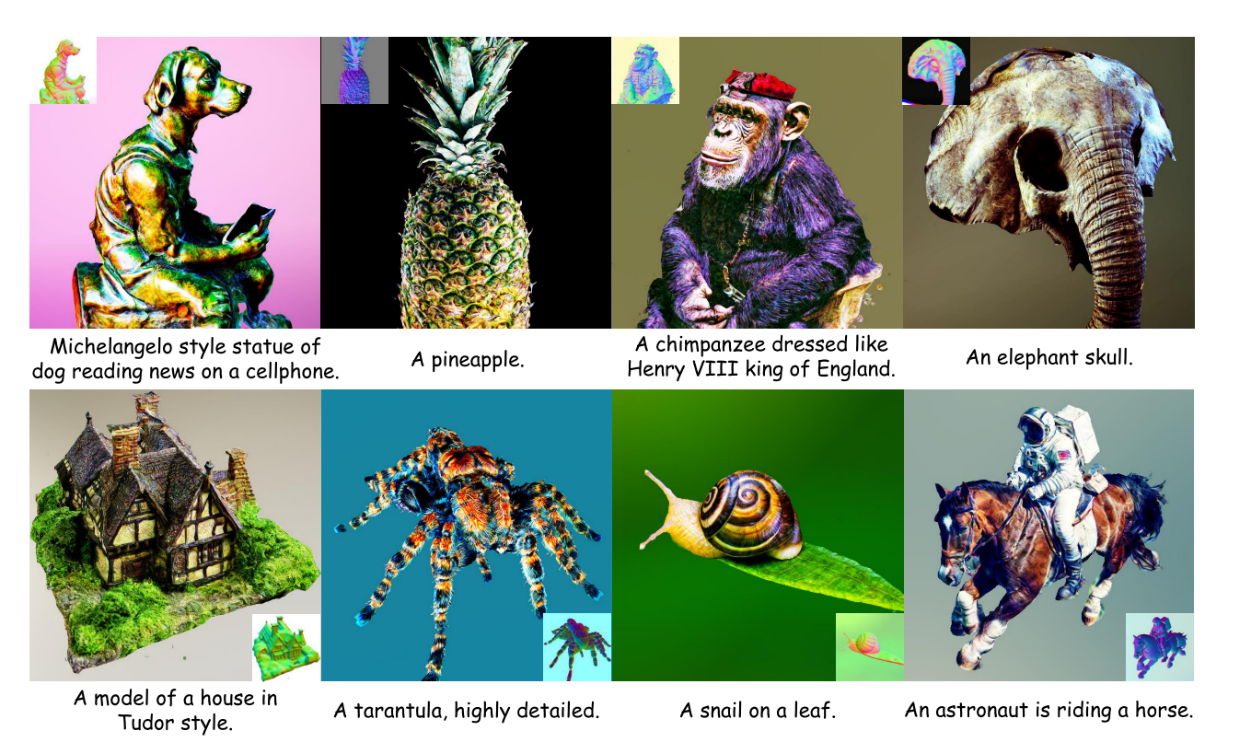


图2 ProlificDreamer生成结果

上图是ProlificDreamer从头开始生成的文本到3D样本。论文使用的基础模型是Stable Diffusion，不使用任何其他辅助模型或用户提供的形状指导。所有结果均由VSD生成。VSD使用n = 1个粒子与SDS进行公平的比较。

图3显示了ProlificDreamer以对象为中心的生成。论文将论文中的方法与三个SOTA方法DreamFusion，Magic3D和Fantasia3D进行了比较。如图3所示，ProlificDreamer生成的3D对象具有更高的保真度和更多的细节，这证明了论文方法的有效性。

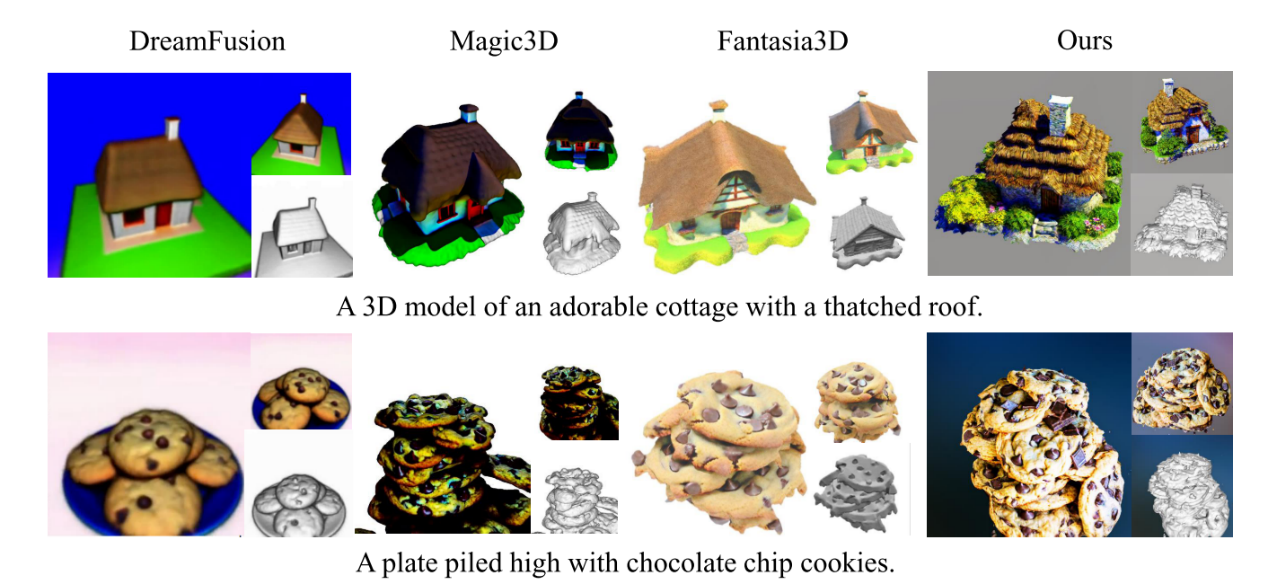


图3 ProlificDreamer与baseline的比较

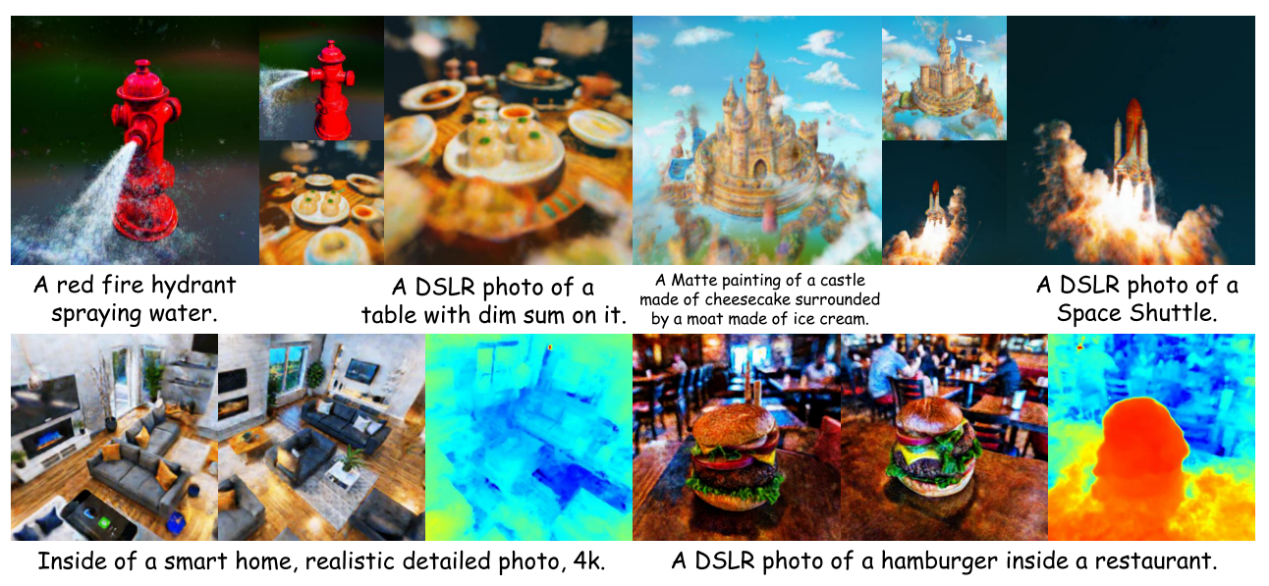
图4显示了ProlificDreamer进行大场景生成的结果。论文的方法可以生成具有高保真度和精细细节的360度全景场景。深度图显示了场景具有几何形状，而不仅仅是360度纹理球体，验证了仅使用场景初始化，论文提出的方法可以生成高保真的大型场景，而无需对现有组件进行太多修改。

图4 ProlificDreamer进行大场景生成的结果

# 3 总结

在这篇论文中，作者系统地研究了文本到3D生成的问题。在算法制定方面，论文提出了变分蒸馏（VSD），一个基于粒子的变分框架，将3D参数作为一个随机变量，并推断其分布。VSD在变分公式中自然地概括了SDS，并解释和解决了之前观察到的SDS的实际问题。通过对3D表示的其他正交改进，论文的整体方法ProlificDreamer可以生成高保真NeRF和照片般逼真的纹理网格。

虽然ProlificDreamer实现了显著的文本到3D的结果，目前的生成仍然需要几个小时的时间，远远高于图像生成的扩散模型。虽然论文的场景初始化可以实现大场景生成，但训练期间的相机姿态与场景结构无关，这可以通过根据场景结构设计自适应相机姿态范围来改善，以获得更好的生成细节。此外，与其他生成模型一样，论文的方法可能会被用来生成虚假和恶意的内容。

最后，将论文生成的3D内容和text-to-image的Diffusion模型生成的2D图像进行对比可以看到，论文生成的3D物体的质量已经和2D图像的生成质量很接近了。论文证明了，只要蒸馏算法设计的比较好，基于2D蒸馏的3D生成是可以达到和2D生成模型相近的纹理质量的。

# 4 参考文献

[1] Wang Z, Lu C, Wang Y, et al. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation[J]. arXiv preprint arXiv:2305.16213, 2023.