

对论文SyncDreamer的复现与实验

摘要

近年来，AIGC技术的发展迅速，GPT的智能问答、基于扩散模型的文生图、图生图等应用使得人们可以更快地将想法转变为现实。在游戏开发、动画制作等领域中，通过文字或图像直接生成三维模型也有着大量应用场景。然而在三维生成领域，由于三维数据集的规模限制，直接使用海量三维数据对三维生成模型进行训练的方法在资源消耗与生成效果上均不尽如人意，因此使用扩散模型先生成图片，再通过可微的三维表示方法对模型进行三维重建成为了一种可选的方法。但此类方法由于扩散模型生成的随机性导致存在多视角不一致的问题，本实验所复现的文章即旨在解决该问题，该文章提出的方法可以以单视角图像为条件生成具有多视角一致性的三维模型。

关键词：三维生成；扩散模型；多视角一致性

1 引言

AIGC (AI Generated Content) 技术，即AI自动创作生成内容的技术，在文字生成、图片生成、视频生成等领域均展现出了绝佳的效果，例如文字生成领域基于Transformer架构的GPT模型与图片生成领域基于扩散模型的Stable Diffusion [15]等模型。这些基于AI的生成模型只需要一句话或者一张图作为提示词，就可以生成出用户需求的结果。AIGC技术的飞速发展，依靠的不仅仅是性能更佳的各类生成模型，更是愈发庞大且包罗万象的数据集。GPT等大语言模型模型使用的训练数据集大小约为2000B，其中包含大约1.5万亿个单词，Stable Diffusion模型的训练数据集也包含百亿个图像标题对。

然而在三维生成领域，数据集的规模却一直限制着三维模型生成技术的发展，过去通常使用的三维数据集ShapeNet仅包含4000余个类别的约300万个模型 [1]，而最新的三维数据集Objaverse也仅包含约1000万个三维物体模型 [3]。由于三维数据集类别与规模的限制，过去的绝大多数三维生成工作都是基于模板进行生成，仅能应用在训练集所包含的某些固定类别上，例如桌椅、汽车、飞机等。[4]

得益于扩散模型在图片生成领域的杰出表现，三维生成相关的研究者们也从中得到启发，产生了两种新方法。第一种方法是使用扩散模型的思路，通过对三维模型进行加噪与去噪的训练，从而预测噪音生成新的三维模型；第二种方法是使用扩散模型所拥有的图像理解与生成能力，通过对扩散模型生成的图片进行升维，进而产生三维模型。前一种方法依然没能摆脱三维数据集规模较小的限制，而后一种方法则借由SDF [12]、DMTet [16]、NeRF [10]等可微分的三维表示方法产生了相对良好的效果。

然而，此方法依然存在部分缺陷，其中之一是生成的三维模型可能存在多视角不一致的问题（Multi-Face Janus Problem）[17]，该问题产生的某一原因是扩散模型每次根据提示词生成的图像并非完全可控，因此使用相同的提示词生成得到的同一物体的不同视角图片在细节上均有差别，这些差别导致了重建出的模型存在多视角不一致问题。本实验复现的文章提出了一种新的对扩散模型进行改进的方法，可以以单张图片作为输入，得到具有多视角一致性的多张图片。以该扩散模型为基础，可以一定程度上缓解使用扩散模型进行三生成中多视角不一致的问题。

2 相关工作

2.1 扩散模型

扩散模型在二维图像生成方面取得了令人印象深刻的成果。扩散模型或扩散概率模型，是利用变分推理训练的参数化马尔科夫链，它可以在有限时间内产生与数据匹配的样本。扩散模型中包括一个正向过程与一个反向过程，在正向过程中，模型逐步给原始数据添加高斯噪音，直到原始数据被噪音破坏，而在反向过程或者生成过程中，模型对加噪过的数据逐步进行去噪，通过对每步添加的噪音进行学习训练，最终将原始数据从噪音中恢复。[5]

随着生成模型的进步和数据集规模的增加，扩散模型可以从文本描述（如名词、形容词或艺术风格等）中组成复杂的语义概念，从而生成高质量的物体和场景图片。由于运行在像素空间的扩散模型在使用时会耗费大量算力与时间，Stable Diffusion [15]提出使用VAE将扩散模型应用在隐空间中，从而使扩散模型能够以更快的时间内生成质量更高的图像。其主要包含三个模型，UNet是扩散模型的主体，用于预测对应提示词的图像噪音；Autoencoder即VAE模型，负责将图像编码至隐空间，并将结果解码为图像；CLIP Text Encoder负责提取输入提示词的文本编码，通过交叉注意力方式放入UNet中作为条件引导模型生成。

2.2 基于二维扩散模型的从文字生成三维模型

使用二维扩散模型进行三维物体生成的最早方法由DreamFusion [13]和SJC [18]两篇文章先后分别提出，该方法为对二维文本到图像的扩散模型进行提取，从而使用文本生成三维模型。在DreamFusion中，作者提出了分数蒸馏采样（SDS），通过使用扩散模型生成的过程来监督NeRF的训练。在扩散模型中，损失函数可以写为预测噪音与真实噪音的均方差损失形式

$$L_{diff} = \mathbb{E}[w(t)||\epsilon_{pred} - \epsilon||^2] = \mathbb{E}[w(t)||UNet(\alpha_t x_{render} + \sigma_t \epsilon | t) - \epsilon||^2]$$

其中 $w(t)$ 为关于时间步 t 的加权函数。使用扩散模型监督一个NeRF模型的生成时，则对NeRF模型渲染出的图片 x_{render} 也加上噪声 $\alpha_t x_{render} + \sigma_t \epsilon$ ，此时如果该NeRF模型渲染出的某角度图像与目标物体的该角度图像接近，则扩散模型预测的噪音应与加上的噪音 ϵ 接近，因此可以使用损失函数 L_{NeRF} 来度量噪音的接近程度

$$L_{NeRF} = \mathbb{E}[w(t)||UNet(\alpha_t x_{render} + \sigma_t \epsilon | t) - \epsilon||^2]$$

由于该损失函数在训练时需要对UNet的损失函数进行梯度的反向传播，导致优化过程缓慢且复杂，因此DreamFusion提出对损失函数的梯度进行分解的方法，从而有

$$\nabla_\theta L_{SDS}(\phi, g(\theta)) = \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_\phi(x_t; y, t) - \epsilon) \frac{\partial x_{render}}{\partial \theta}]$$

其中 $\epsilon_\phi(x_t; y, t)$ 即为噪声预测器UNet， ϕ 为其参数， $g(\theta)$ 为所生成的NeRF模型。

后续工作如Magic3d [6]、Fantasia3d [2]、Prolificdreamer [19]等在不同方面对此方法进行改进。Magic3d提出了两段式的生成流程，先使用InstantNGP [11]生成粗模型，从中提取出包含三角面片的DMTet模型，再经过高分辨率的优化过程，最终生成出分辨率更高且质量更佳的三维模型。Fantasia3d提出将三维模型的几何属性与外观属性进行解耦，使用DMTet作为三维表示，通过扩散模型得到三维模型的几何特征，再通过扩散模型对已有模型的外观进行训练，最终得到目标模型。Prolificdreamer对原SDS损失函数进行了优化，提出了VSD损失函数

$$\nabla_\theta L_{VSD}(\theta) \triangleq \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_{pretrain}(x_t; y, t) - \epsilon_\phi(x_t; y, t)) \frac{\partial g(\theta)}{\partial \theta}]$$

其中 $\epsilon_{pretrain}(x_t; y, t)$ 为预训练的扩散模型去噪器， $\epsilon_\phi(x_t; y, t)$ 为基于变分分布产生的噪音。

2.3 基于二维扩散模型的从单视角图片生成三维模型

基于SDS的由文字生成三维模型的方法对算力与时间的消耗极大，因此使用扩散模型由单视角图片生成三维模型成为了一种可行的方法。Zero123 [8]提出了一种以单视角图片为输入，使用2D扩散模型预测指定视角图片的方法，该文提出使用额外加入的图像及其对应的相机位姿信息作为条件对扩散模型中预训练的UNet进行训练。

One2345 [7]使用预训练的Zero123模型，对输入图像的不同角度新图像进行预测，同时对其相机位姿进行估计，使用若干角度的图像与对应的相机位姿进行三维重建。但由于扩散模型的预测具有随机性，因此当物体在不同视角下几何形状或纹理信息具有显著差别时，该方法得到的重建效果较差，只适用于形状与纹理信息较简单的物体。

3 本文方法

3.1 本文方法概述

本实验所复现的工作SyncDreamer [9]即旨在解决单视角图片生成三维模型中多视角不一致的问题。给定一个物体的输入视角图像 y ，SyncDreamer的目标是生成该物体的多视角图像，为了提升其一致性，该文将生成过程表示为一个多视角扩散模型，从而将不同角度图像的生成进行关联。

3.2 扩散模型

扩散模型的目标是通过学习得到一个概率模型 $p_\theta(x_0) = \int p_\theta(x_{0:T}) dx_{1:T}$ ，其中 x_0 是原始数据， $x_{1:T} := x_1, \dots, x_T$ 是隐变量，扩散模型的联合分布由一个马尔科夫链构成

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t)$$

其中 $p(x_T) = \mathcal{N}(x_T; 0, I)$ ， $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$ ，而其中的 $\mu_\theta(x_t, t)$ 是一个可训练的分量，方差 σ_t^2 是与时间相关的常数。为了从噪音中获得原始数据，模型需要对 μ_θ 进行预

测，因此前向过程由一个马尔科夫链构成

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1})$$

其中 $q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ 且 $\beta_t \in (0, 1), t = 1, 2, \dots, T$ 。在DDPM中， μ_θ 被定义为

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t))$$

其中 α_t 和 $\bar{\alpha}_t$ 是由 β_t 得到的常数， ϵ_θ 是噪音预测器，因此可以通过如下损失函数对 ϵ_θ 进行学习

$$L = \mathbb{E}_{t, x_0, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|_2]$$

其中 ϵ 为随机采样的高斯分布。

3.3 多视角扩散模型

在将原始的扩散模型应用在多视角图像生成任务中难以保持不同视角图像的信息一致性，为了解决该问题，该文将生成过程构造成一个不同视角互相关联的多视角扩散模型。假设输入的某视角图像为 y ，需要生成 N 张图像 $\{x_0^{(1)}, \dots, x_0^{(N)}\}$ ，其中 x 的下角标为时间步 t ， x 的上角标为第 i 个视角的图像， $i \in 1, 2, \dots, N$ ，需要学习的联合分布为

$$p_\theta(x_0^{(1:N)}|y) := p_\theta(x_0^{(1)}, \dots, x_0^{(N)}|y)$$

多视角扩散模型的前向过程是原始扩散模型前向过程的直接扩展，将噪声添加在每个视角的图像上

$$q(x_{1:T}^{(1:N)}|x_0^{(1:N)}) = \prod_{t=1}^T q(x_t^{(1:N)}|x_{t-1}^{(1:N)}) = \prod_{t=1}^T \prod_{n=1}^N q(x_t^{(n)}|x_{t-1}^{(n)})$$

其中 $q(x_t^{(n)}|x_{t-1}^{(n)}) = \mathcal{N}(x_t^{(n)}; \sqrt{1-\beta_t}x_{t-1}^{(n)}, \beta_t I)$ 。逆向过程也使用相同方式进行构建

$$p_\theta(x_{0:T}^{(1:N)}) = p(x_T^{(1:N)}) \prod_{t=1}^T p_\theta(x_{t-1}^{(1:N)}|x_t^{(1:N)}) = p(x_T^{(1:N)}) \prod_{t=1}^T \prod_{n=1}^N p_\theta(x_{t-1}^{(n)}|x_t^{(1:N)})$$

其中 $p_\theta(x_{t-1}^{(n)}|x_t^{(1:N)}) = \mathcal{N}(x_{t-1}^{(n)}; \mu_\theta^{(n)}(x_{t-1}^{(1:N)}, t), \sigma_t^2 I)$ ，上式的后半部分成立是因为该文假设有一个对角方差矩阵，第 n 个视图 $x_{t-1}^{(n)}$ 的均值 $\mu_\theta^{(n)}$ 取决于所有视角图像 $x_t^{(1:N)}$ 的状态，因此 $\mu_\theta^{(n)}$ 被定义为

$$\mu_\theta^{(n)}(x_t^{(1:N)}, t) = \frac{1}{\sqrt{\alpha_t}}(x_t^{(n)} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta^{(n)}(x_t^{(1:N)}, t))$$

训练的损失函数为

$$L = \mathbb{E}_{t, x_0^{(1:N)}, n, \epsilon^{(1:N)}} [\|\epsilon^{(n)} - \epsilon_\theta^{(n)}(x_t^{(1:N)}, t)\|_2]$$

其中 $\epsilon^{(1:N)}$ 是添加到 N 张图像中的形状为 $N * H * W$ 的标准高斯噪音， $\epsilon_\theta^{(n)}$ 是对第 n 个图像的噪音预测器。

在训练中，首先从数据集中获得同一物体的 N 个角度的图像，再采样获得时间步 t 和对应噪音 $\epsilon^{(1:N)}$ ，从而由原始数据 $x_0^{(1:N)}$ 得到加噪后的数据 $x_t^{(1:N)}$ ，之后随机选择一个视角 n 并使用

对应的噪音预测器 $\epsilon_{\theta}^{(n)}$ 预测噪音，最后计算采样噪音 $\epsilon^{(1:N)}$ 和预测噪音之间的L2距离作为训练的损失。

该文提出的多视角扩散模型可以视作 N 个同步的噪音预测器 $\{\epsilon_{\theta}^{(n)} | n = 1, \dots, N\}$ ，在每个时间步 t ，每个噪音预测器 $\epsilon_{\theta}^{(n)}$ 负责预测从 $x_t^{(n)}$ 到 $x_{t-1}^{(n)}$ 的噪音，同时这些噪音预测器是同步的，因为在每个去噪步骤中每个噪音预测器都通过关联其他视角的带噪图像 $x_t^{(1:N)}$ 来互相交换信息。在实际实现中，所有的 N 个噪声预测器使用一个共享的UNet，以输入视角和第 n 个目标视角间的差值 Δv^n 以及所有视角的带噪图像 $x_t^{(1:N)}$ 作为条件，即

$$\epsilon_{\theta}^{(n)}(x_t^{(1:N)}, t) = \epsilon_{\theta}(x_t^{(n)}; t, \Delta v^n, x_t^{(1:N)})$$

3.4 基于3D特征的去噪

与DDPM和Stable Diffusion类似，该文使用的噪音预测器 ϵ_{θ} 包含一个以带噪的图像作为输入并对图像进行去噪的UNet。为了保证模型的泛化能力，该文使用Zero123中预训练的UNet，将输入视角图像和带噪的目标视角图像拼接后作为UNet的输入，然后计算输入视角与目标视角之间的视角差，再与通过CLIP [14]模型产生的图像的文本编码信息进行拼接作为UNet的条件。在训练SyncDreamer模型的过程中，对UNet与CLIP模型的参数进行固定。

为了增强生成的多个视角间的一致性，模型在生成当前图像时，需要在三维空间中感知相应的特征，为了实现这一点，该文构造了一个带有 V^3 个顶点的三维体，然后将这些顶点投影到所有目标视角上，从而获得这些特征，每个目标视角的特征被拼接形成一个空间特征体，然后对特征体应用一个3D CNN来获取并处理空间关系。为了对第 n 个目标视角进行去噪，该文构造了一个与该视角图像按像素级对其的特征平截锥体，其特征通过对空间特征体插值获得。最后，在UNet中当前视图的每个中间特征图上，该文应用了一个新的关于深度的注意力层，沿着深度维度从像素级对其的特征平截锥体中提取特征。

该文为以上设计提出了两点理由。首先是特征空间体由所有目标视角构建，所有目标视角的图像都共享相同的特征空间体进行去噪，这意味着所有目标视角都有一个全局约束，即关注同一个物体。其次是新的注意力层直沿着深度维度进行注意力计算，这加强了一个局部的对极线约束，即特定位置的特征应与其他视角上的对极线上的相应特征相一致。

4 复现细节

4.1 与已有开源代码对比

SyncDreamer发布了开源代码，因此本实验主要在源代码上进行测试与修改。SyncDreamer的原结构如下图所示，主要包含四个模块，分别是AutoEncoderKL、UNet、CLIP以及SpatialVolumeNet，其中AutoEncoder与UNet模块使用预训练的Zero123模型，CLIP模块使用预训练的CLIP模型。

由于该文的核心部分均为预训练的大模型，因此本实验未对其核心部分进行修改，主要在输入与输出部分进行补充。在输入上，源代码仅支持单视角图像作为输入，因此额外添加了一个文转图模块，通过扩散模型接收用户的文字提示词，从而生成对应图片，同时由于扩散模型生成的图片可能包含影响三维生成的背景或其它物体，因此使用SAM提取核心部分，

最后作为模型的输入。在输出部分，源代码仅能生成多视角的图片，因此添加了一个三维重建模块，通过NeRF将图片转化为三维模型。

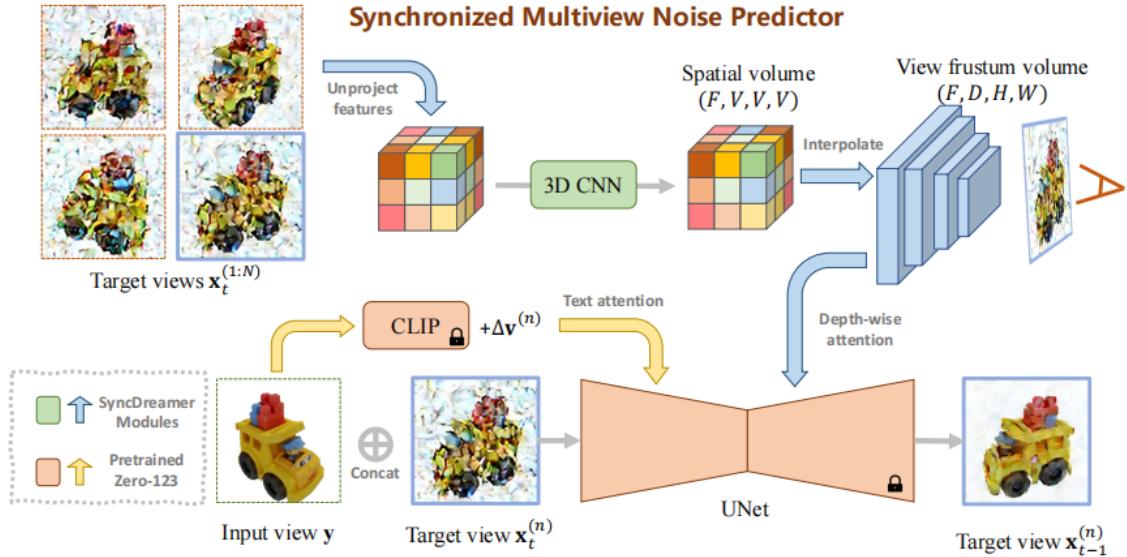


图 1. SyncDreamer原结构

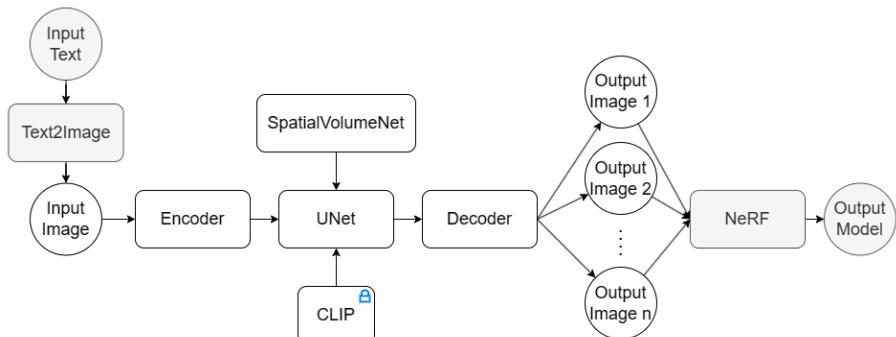


图 2. 本实验结构

5 实验结果展示

5.1 使用单张图片进行生成



图 3. 盔甲图片



图 4. 猫玩具图片



图 5. 路灯图片



图 6. 悟空图片



图 7. 沙发图片



图 8. 茶壶图片

5.2 使用文本进行生成



图 9. "a zoomed out DSLR photo of a beautifully carved wooden knight chess piece"



图 10. "a DSLR photo of an ice cream sundae"



图 11. "a teddy toy"



图 12. "a zoomed out DSLR photo of a baby bunny sitting on top of a stack of pancakes"

6 总结与展望

本实验对SyncDreamer一文进行了复现与改进。SyncDreamer提出了多视角扩散模型，通过构建不同视角共享的特征空间体使不同视角的噪音预测器可以互相关联，从而达到提高多视角一致性的目的。但在实际实验中发现，此方法对部分输入图片依然存在不同视角信息一致性较低的情况。未来的研究可以尝试采用对几何结构与纹理信息进行解耦并逐步生成的方式，并在UNet的去噪过程中添加具有更多三维先验的信息如深度图等用于监督三维模型生成的训练。

参考文献

- [1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, and Hao Su. Shapenet: An information-rich 3d model repository. 2015.
- [2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023.
- [3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.
- [4] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Advances In Neural Information Processing Systems*, 2022.
- [5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.
- [6] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [7] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023.
- [8] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023.
- [9] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. 2023.

- [10] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [11] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022.
- [12] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [13] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.
- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2021.
- [16] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [17] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. 2023.
- [18] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation, 2022.
- [19] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Proflificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation, 2023.