

One-2-3-45++：通过一致的多视角生成和3D扩散实现快速单图像到3D对象转换

刘明华^{1*}史若曦^{1*}陈凌浩^{1,2†}张卓扬^{3†}徐超^{4*}魏欣悦¹

陈汉生^{5†}曾崇^{2†}顾家源¹苏昊¹

1 加州大学圣地亚哥分校2 浙江大学3 清华大学4 加州大学洛杉矶分校5 斯坦福大学

arX
[cs.
14
No.
202

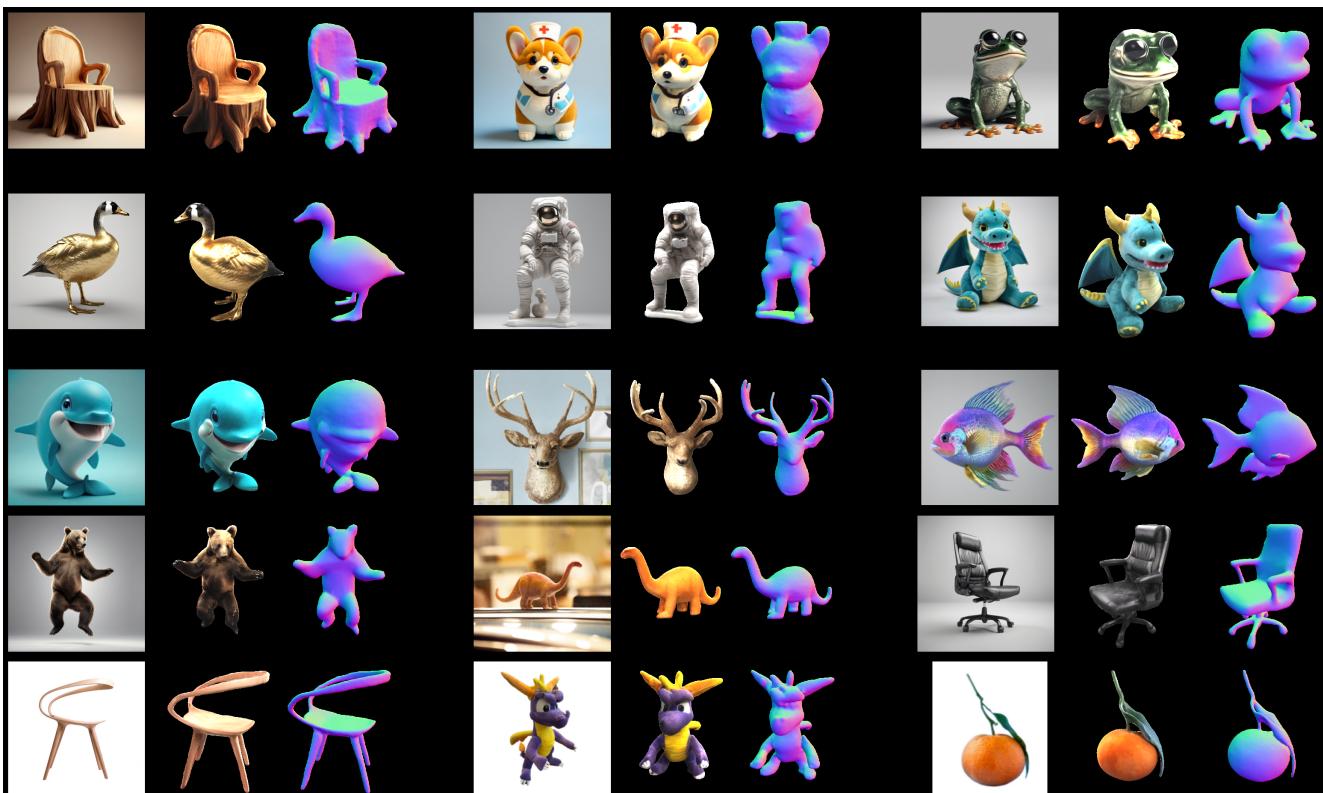


图1. One-2-3-45++ 能够在不到一分钟的时间内将任意物体的单张RGB图像转换为高保真的纹理网格。生成的网格与原始输入图像高度一致。图中展示了输入图像（及文本提示）、纹理网格和法线贴图。

摘要

最近，开放世界3D物体生成的进展显著，其中图像到3D的方法相比文本到3D的方法提供了更精细的控制。然而，大多数现有模型在同时提供快速生成速度和高保真度方面表现不足，这两个特性对于实际应用至关重要。本文中，我们提出了One-2-3-45++，一种创新方法，能够在约一分钟内将单张图像转化为详细的3D纹理网格。我们的方法旨在充分利用广泛的知识。

嵌入在2D扩散模型和来自珍贵但有限的3D数据的先验知识中。这一过程首先通过微调2D扩散模型以实现一致的多视角图像生成，随后借助多视角条件下的3D原生扩散模型将这些图像提升至3D。广泛的实验评估表明，我们的方法能够生成高质量、多样化的3D资产，这些资产与原始输入图像高度吻合。我们的项目网页：<https://sudoai-3d.github.io/One2345plus>。

1. 引言

从单张图像或文本提示生成3D形状是计算机视觉领域的一个长期难题，对于众多应用而言至关重要。尽管取得了显著进展，

*同等贡献。†在加州大学圣地亚哥分校实习期间完成的工作。

在2D图像生成领域，由于先进的生成方法和大规模的图文数据集，已经取得了显著的进展。然而，将这一成功转移到3D领域却受到3D训练数据有限的阻碍。尽管许多研究引入了复杂的3D生成模型[8, 16, 38, 87]，但大多数仅依赖于3D形状数据集进行训练。鉴于公开可用的3D数据集规模有限，这些方法在开放世界场景中往往难以泛化到未见过的类别。

另一类工作，如DreamFusion [50]、Magic3D [31]，利用了像CLIP [52]和Stable Diffusion [57]这样的2D先验模型的广泛知识或强大的生成潜力。它们通常从零开始为每个输入的文本或图像优化一个3D表示（例如，NeRF或网格）。在优化过程中，3D表示被渲染成2D图像，并利用2D先验模型来计算这些图像的梯度。尽管这些方法取得了令人印象深刻的结果，但每个形状的优化过程可能非常耗时，生成每个输入的单个3D形状可能需要数十分钟甚至数小时。此外，它们经常遇到“多面”或Janus问题，生成的结果往往色彩过饱和，并且带有从NeRF或三平面表示继承的伪影，同时在不同随机种子下生成多样结果方面也面临挑战。

最近的一项工作One-2-3-45 [34]提出了一种创新的方法，利用2D扩散模型的丰富先验知识进行3D内容生成。它首先通过视图条件化的2D扩散模型Zero123 [35]预测多视角图像。随后，这些预测的图像通过一种可泛化的NeRF方法 [39] 进行3D重建。尽管One-2-3-45能够通过一次前馈传递生成3D形状，但其效果往往受到Zero123多视角预测不一致的限制，导致3D重建结果不尽如人意。

在本文中，我们介绍了一项名为One-2-3-45++的创新方法，该方法有效克服了One-2-3-45的不足，显著提升了鲁棒性和质量。One-2-3-45++以单张图像作为输入，同样包含两个主要阶段：2D多视角生成和3D重建。在初始阶段，One-2-3-45++不是单独使用Zero123来预测每个视角，而是同时预测一致的多视角图像。这一过程通过将一组简洁的六视角图像拼接成单张图像，然后微调2D扩散模型，以输入参考图像为条件生成这张组合图像来实现。这样，2D扩散网络在生成过程中能够关注每个视角，确保各视角间结果更加一致。在第二阶段，One-2-3-45++采用了一个多视角条件下的基于3D扩散的模块，以由粗到细的方式预测带纹理的网格。一致的多视角条件图像作为3D重建的蓝图，

建设，促进零样本幻觉能力。同时，3D扩散网络在提升多视角图像方面表现出色，得益于其能够利用从3D数据集中提取的广泛先验知识。最终，One-2-3-45++采用轻量级优化技术，利用一致的多视角图像进行监督，从而有效提升纹理质量。

如图1所示，One-2-3-45++能够在不到一分钟的时间内高效生成具有逼真纹理的3D网格，并提供精确的细粒度控制。我们的全面评估，包括在广泛测试集上的用户研究和客观指标，突显了One-2-3-45++在鲁棒性、视觉质量以及最重要的是对输入图像的忠实度方面的优越性。

2. 相关工作

2.1. 三维生成

3D生成近年来引起了广泛关注。在大规模预训练2D模型出现之前，研究人员通常深入研究3D原生成模型，这些模型直接从3D合成数据或真实扫描中学习，并生成各种3D表示，如点云[1, 15, 42, 48, 83]、3D体素[9, 60, 74, 75]、多边形网格[16, 17, 26, 33, 38, 47, 68]、参数化模型[21]以及隐式场[8, 14, 19, 25, 30, 43, 49, 73, 78, 82, 84, 86, 87]。然而，由于3D数据的有限可用性，这些模型往往集中于少数几个类别（如椅子、汽车、飞机、人类等），难以在开放世界中推广到未见过的类别。

近年来二维生成模型（如DALL-E [54]、Imagen [59] 和 Stable Diffusion [58]）以及视觉语言模型（如CLIP [52]）的出现，为我们提供了关于三维世界的强大先验知识，从而推动了三维生成研究的蓬勃发展。值得注意的是，像DreamFusion [50]、Magic3D [31] 和 ProlificDreamer [71] 这样的模型开创了一种逐形状优化的方法 [6, 7, 12, 23, 29, 41, 44–46, 51, 53, 61, 63, 64, 67, 76, 77, 81]。这些模型旨在为每个独特的输入文本或图像优化三维表示，利用二维先验模型进行梯度引导。尽管这些方法取得了令人印象深刻的结果，但它们往往面临优化时间过长、“多面问题”、颜色过饱和以及结果缺乏多样性等问题。一些工作还专注于为输入网格创建纹理或材质，利用二维模型的先验知识 [5, 56]。

新一轮研究浪潮，以Zero123 [35]等作品为代表，展示了利用预训练的2D扩散模型从单张图像或文本中合成新视角的潜力，为3D生成打开了新的大门。例如，One-2-3-45 [34]利用Zero123预测的多视角图像，仅需45秒即可生成带纹理的3D网格。然而，

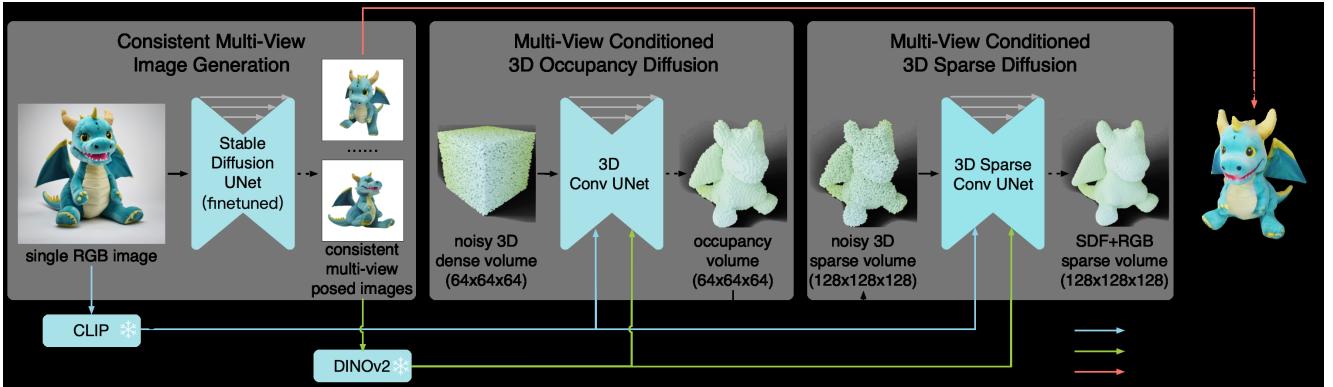


图2. 从一张RGB图像作为输入开始，我们首先通过微调2D扩散模型生成一致的多视角图像。这些多视角图像随后通过一对3D原生扩散网络提升为3D。在整个3D扩散过程中，生成的多视角图像作为关键的引导条件。从去噪后的体素中提取3D网格后，我们进一步通过使用多视角图像作为监督的轻量级优化来增强纹理。我们的One-2-3-45+能够在20秒内生成初始的纹理网格，并在大约一分钟内输出精细化的结果。

Zero123生成的多视角图像缺乏三维一致性。我们的研究，连同几项同时期的研究[32, 37, 40, 62, 72, 80]，致力于提升这些多视角图像的一致性——这是后续三维重建应用的关键步骤。

2.2. 稀疏视角重建

尽管传统的三维重建方法，如多视图立体或基于NeRF的技术，通常需要密集的输入图像集合以进行精确的几何推断，但许多最新的通用NeRF解决方案[3, 24, 28, 36, 39, 55, 66, 69, 70, 79]力求在场景间学习先验知识。这使得它们能够从稀疏的图像集合中推断出NeRF，并推广到新的场景。这些方法通常摄取少量源视图作为输入，利用二维网络提取二维特征。这些像素特征随后被反投影并聚合到三维空间中，从而促进密度（或SDF）和颜色的推断。然而，这些方法可能依赖于具有精确对应关系的一致多视图图像，或者在超越训练数据集的泛化方面拥有有限的先验知识。

最近，一些方法[2, 27, 65, 88]采用了扩散模型来辅助稀疏视角重建任务。然而，它们通常将问题框架为新颖视角合成，需要额外的处理，如使用3D表示进行蒸馏，以生成3D内容。我们的工作利用多视角条件下的3D扩散模型进行3D生成。该模型直接从3D数据中学习先验知识，并消除了对额外后处理的需求。此外，一些同期工作[37, 40, 62]采用基于NeRF的每场景优化进行重建，利用了专门的损失函数。

3. 方法

在传统游戏工作室中，3D内容的创作涵盖了一系列阶段，包括概念艺术、3D建模等。

建模和纹理处理等。每个阶段都需要不同的专业技能，并且这些技能是互补的。例如，概念艺术家需要具备创造力、丰富的想象力以及将3D资产可视化的能力。相比之下，3D建模师必须熟练掌握3D建模工具，并能够将多视角的概念图转化为逼真的模型，即使这些图纸存在不一致或错误。

One-2-3-45+旨在利用丰富的2D先验知识和有限但宝贵的3D数据，遵循类似的哲学理念。如图2所示，对于一个物体的单张输入图像，One-2-3-45+首先生成该物体的连贯多视角图像。这是通过微调预训练的2D扩散模型实现的，其作用类似于概念艺术家的角色。这些生成的图像随后输入到多视角条件下的3D扩散模型中进行3D建模。该3D扩散模块经过大量多视角和3D配对的训练，擅长将多视角图像转换为3D网格。最后，生成的网格通过一个轻量级的优化模块进行进一步的纹理质量提升，该模块受多视角图像的引导。

3.1. 一致性多视图生成

最近，Zero123展示了通过微调预训练的2D扩散网络来整合相机视角控制的能力，从而能够从单一参考图像中合成物体的新视角。尽管之前的研究已经利用Zero123生成了多视角图像，但这些图像在不同视角间往往存在不一致性。这种不一致性产生的原因在于Zero123模型在生成多视角图像时，对每个视角的条件边际分布进行独立建模，而没有考虑多视角生成过程中的视角间通信。在本研究中，我们提出了一种创新方法，以生成一致的多视角图像，显著提升了下游3D重建的效果。

多视图平铺

为了在单一扩散过程中生成多个视图，我们采用了一种简单的策略，即将

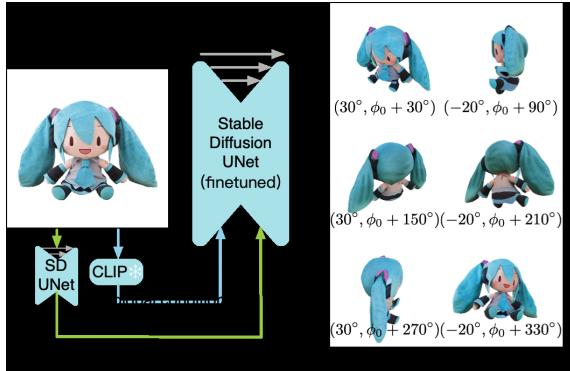


图3. 一致的多视角生成：我们将多视角图像拼接成单帧，并微调Stable Diffusion模型以生成该合成图像，使用输入的参考图像作为条件。我们利用预定的绝对仰角和相对方位角。在3D重建过程中，我们不需要推断输入图像的仰角。

将稀疏的6个视图合成一张具有 3×2 布局的图像，如图3所示。随后，我们微调一个预训练的2D扩散网络，以生成合成图像，条件是单张输入图像。这种策略使得多个视图在扩散过程中能够相互作用。

定义多视角图像的相机姿态并非易事。考虑到训练数据集中的三维形状缺乏对齐的标准姿态，使用绝对相机姿态来处理多视角图像可能会导致生成模型出现歧义。或者，如果我们像Zero123那样将相机姿态设置为相对于输入视图，那么下游应用将需要推断输入图像的仰角以推导出多视角图像的相机姿态。这一额外步骤可能会引入误差。为了解决这些问题，我们选择固定绝对仰角并结合相对方位角来定义多视角图像的姿态，从而有效地消除方向模糊性，而无需进一步的仰角估计。更具体地说，六个姿态是通过交替使用 30° 和 -20° 的仰角，以及从 30° 开始并以 60° 为增量递增的方位角来确定的，如图3所示。

网络与训练细节

为了微调Stable Diffusion，使其能够添加图像条件并生成连贯的多视角合成图像，我们采用了三种关键的网络或训练设计：

(a) 局部条件：我们采用参考注意力技术[85]来引入局部条件。具体来说，我们使用去噪UNet模型处理参考输入图像，并将条件参考图像的自注意力键和值矩阵附加到去噪多视角图像的相应注意力层中。

(b) 全局条件：我们利用CLIP图像嵌入作为全局条件，替代Stable Diffusion中原本使用的文本标记特征。

融合。这些全局图像嵌入通过一组可学习的权重进行乘法运算，为网络提供了对物体整体的语义理解。(c) 噪声调度：原始的Stable Diffusion模型使用的是缩放线性噪声调度。我们发现在微调过程中有必要切换到线性噪声方案。我们使用Objaverse数据集[11]中的3D形状对Stable Diffusion2 v-mode进行微调。对于每个形状，我们通过从指定范围内随机采样输入图像的相机姿态，并从提供均匀照明的精选HDRI环境光照集中随机选择一个，生成三个数据点。最初，我们仅对自注意力层以及交叉注意力层的关键矩阵和值矩阵使用LoRA[22]进行微调。随后，我们使用保守的学习率对整个UNet进行了微调。微调过程使用了16个GPU，耗时约10天。

3.2. 多视图条件下的三维扩散

尽管先前的工作利用可泛化的NeRF方法进行3D重建，但主要依赖于多视角图像的精确局部对应关系，并且对3D生成的先验信息有限。这限制了它们在处理由2D扩散网络生成的复杂且不一致的多视角图像时的有效性。相反，我们提出了一种创新的方法，通过利用多视角条件下的3D生成模型，将生成的多视角图像提升到3D。该方法旨在通过在大量3D数据上训练表达性的3D原生扩散网络，学习一个以多视角图像为条件的合理3D形状流形。

****3D体积表示**** 如图2所示，我们将一个带纹理的3D形状表示为两个离散的3D体积：有符号距离函数(SDF)体积和颜色体积。SDF体积测量从每个网格单元中心到最近形状表面的有符号距离，而颜色体积捕捉相对于网格单元中心最近表面点的颜色。此外，SDF体积可以转换为离散的占用体积，其中每个网格单元根据其SDF的绝对值是否低于预定义阈值来存储二进制占用信息。

****两阶段扩散**** 捕捉3D形状的精细细节需要使用高分辨率的3D网格，这不可避免地带来了巨大的内存和计算成本。因此，我们遵循LAS-Diffusion [87]，采用从粗到细的两阶段方式生成高分辨率体积。具体来说，初始阶段生成一个低分辨率（例如， $n = 64$ ）的全3D占用体积 $F \in \mathbb{R}^{n \times n \times n \times 1}$ ，以近似3D形状的外壳，而第二阶段则生成一个高分辨率（例如， $N = 128$ ）的稀疏体积 $S \in \mathbb{R}^{N \times N \times N \times 4}$ ，该体积预测占用区域内的精细SDF值和颜色。我们为每个阶段使用单独的扩散网络。

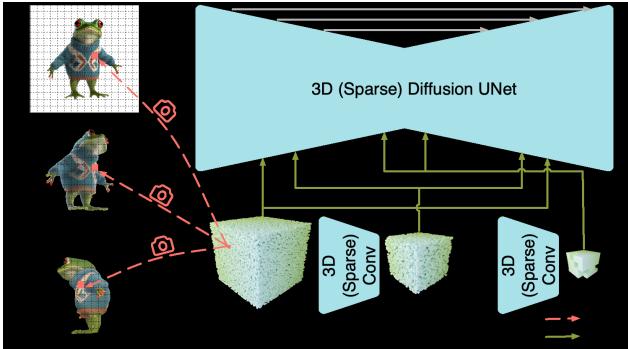


图4. 多视角局部条件：我们采用预训练的2D骨干网络为每个视图提取2D块特征。然后，利用已知的投影矩阵将这些特征聚合，构建一个3D特征体。该特征体通过3D卷积神经网络进一步处理，生成不同分辨率的特征体。随后，这些特征体与扩散UNet中相应的特征体连接，以指导3D扩散过程。

在第一阶段，我们在UNet中使用普通的3D卷积来生成完整的3D占据体积F，而在第二阶段，我们在UNet中引入了3D稀疏卷积以产生3D稀疏体积S。这两个扩散网络均使用去噪损失[20]进行训练：

$$\mathcal{L}_{x_0} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), t \sim \mathcal{U}(0, 1)} \|f(x_t, t, c) - x_0\|_2^2$$

其中， ϵ 和 t 分别代表采样的噪声和时间步， x_0 是数据点（F 或 S）， x_t 是其噪声版本， c 是多视角条件， f 是UNet。N 和 U 分别表示高斯分布和均匀分布。多视角条件训练传统的3D原生扩散网络由于3D数据的有限可用性，可能难以泛化。然而，使用生成的多视角图像可以提供全面的指导，大大简化3D生成的想象难度。我们通过首先提取局部图像特征，然后构建一个条件3D特征体积，记作C，来整合多视角图像以指导扩散过程。这一策略遵循局部先验有助于更容易泛化的原理 [87]。

如图4所示，给定m个多视角图像，我们首先采用预训练的2D骨干网络DINOv2，为每张图像提取一组局部块特征。然后，我们通过将体积内的每个3D体素投影到m个多视角图像上，利用已知的相机姿态构建一个3D特征体积C。对于每个3D体素，我们通过共享权重的MLP聚合m个关联的2D块特征，随后进行最大池化。这些聚合的特征共同形成了特征体积C。

在扩散网络中，UNet由几个层次组成。例如，初始阶段的占用UNet有五个层次： 64^3 、 32^3 、 16^3 、 8^3 和 4^3 。首先，我们构建一个与起始分辨率匹配的条件特征体积C，如前所述。随后进行3D卷积操作。

网络随后应用于C，生成后续分辨率的体素。生成的条件体素随后与UNet内部的体素连接，以指导扩散过程。在第二阶段，我们构建稀疏条件体素并利用3D稀疏卷积。为了促进颜色体素的扩散，我们还将2D像素级投影的颜色连接到扩散UNet的最后一层。此外，我们将输入图像的CLIP特征作为全局条件进行整合。详细解释请参阅补充材料。

训练与推理细节

我们使用Obajverse数据集[11]中的3D形状来训练两个扩散网络。对于每个3D形状，我们首先将其转换为水密流形，然后提取其SDF体积。我们将形状的多视图渲染反投影以获取3D彩色点云，用于构建颜色体积。在训练过程中，我们利用地面真值渲染作为多视图条件。由于两个扩散网络分别训练，我们引入了相机姿态的随机扰动，并对第二阶段的初始占用率注入随机噪声以增强鲁棒性。我们使用8个A100 GPU对两个扩散网络进行训练，每个阶段大约需要10天。更多细节请参阅补充材料。

在推理过程中，首先用高斯噪声初始化一个 64^3 的网格，然后由第一个扩散网络进行去噪。每个预测的占用体素进一步细分为8个更小的体素，用于构建高分辨率稀疏体积。稀疏体积用高斯噪声初始化，然后由第二个扩散网络去噪，得到每个体素的SDF和颜色预测。最后，应用Marching Cubes算法提取纹理网格。

3.3. 纹理细化

鉴于多视图图像具有比3D颜色更高的分辨率，我们可以通过一个轻量级的优化过程来细化生成的网格的纹理。为此，我们固定生成的网格的几何结构，同时优化由TensoRF[4]表示的颜色场。在每次迭代中，通过光栅化将网格渲染成2D，并查询颜色网络。我们利用生成的多视图一致图像，使用L2损失来指导纹理优化。最后，我们将优化后的颜色场烘焙到网格上，表面法线作为观察方向。

4. 实验

4.1. 图像到三维的对比

基线方法：我们评估了One-2-3-45++在基于优化和前馈方法中的表现。在基于优化的方法中，我们的基线包括使用Zero123 XL [35]作为其骨干的DreamFusion [50]，以及SyncDreamer [37]和DreamGaussian [63]。

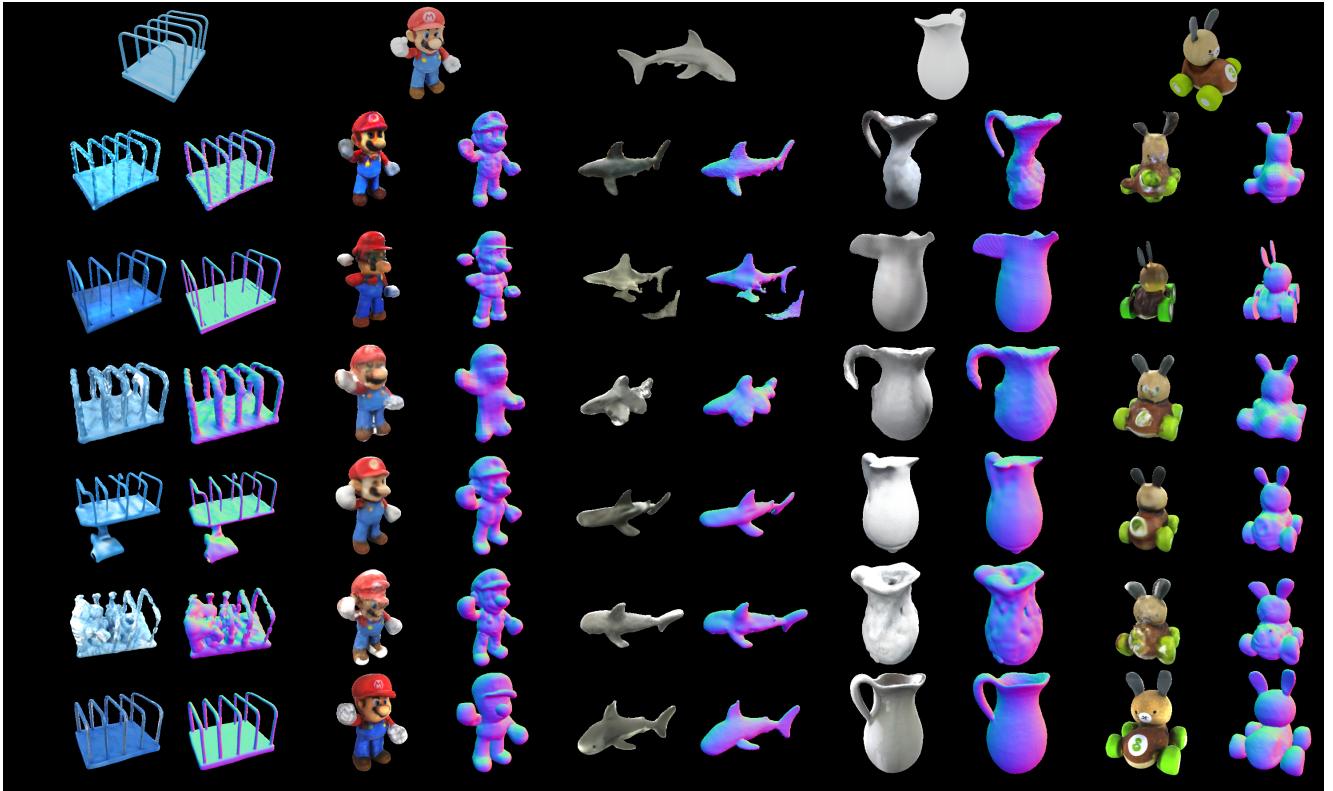


图5. 各种单图像到3D方法的定性结果。展示了输入图像、纹理网格和法线图。

表1. 单张图像到3D的比较。评估于
GSO [13] 数据集，包含 1,030 个 3D 物体。

| Method | F-Sco. (%) | CLIP-Sim | User-Pref. (%) | Time |
|----------------------|------------|----------|----------------|-------|
| Zero123 XL [10] | 91.6 | 73.1 | 58.6 | 30min |
| One-2-3-45 [34] | 90.4 | 70.8 | 52.7 | 45s |
| SyncDreamer [37] | 84.8 | 68.9 | 28.4 | 6min |
| DreamGaussian [63] | 81.0 | 68.4 | 31.5 | 2min |
| Shap-E [25] | 91.8 | 73.1 | 40.8 | 27s |
| Ours | 93.6 | 81.0 | 87.6 | 60s |

对于前馈方法，我们与One-2-3-45 [34]和Shap-E [25]进行比较。我们采用了ThreeStudio [18]的实现来处理Zero123 XL [18]，并使用其他方法的原始官方实现。

数据集与评估指标：我们使用GSO数据集[13]中的全部1,030个形状来评估各方法的性能，据我们所知，这些形状在训练过程中未被任何方法接触过。对于每个形状，我们生成一张正面视图图像作为输入。遵循One-2-3-45[34]的方法，我们采用F-Score和CLIP相似度作为评估指标。F-Score评估预测网格与真实网格之间的几何相似性。对于CLIP相似度指标，我们对每个预测和真实网格渲染24个不同视角，计算每对对应图像的CLIP相似度，然后对所有视角的值进行平均。在计算指标之前，我们进行对齐操作。

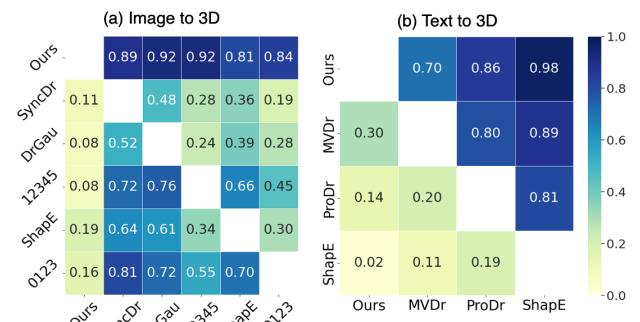


图6. 一项涉及53名参与者的用户研究结果。

每个单元格显示一种方法（行）优于另一种方法（列）的概率或偏好率。

使用线性搜索和ICP算法相结合的方法，将预测的网格与地面真值网格进行比较。

用户研究：我们还进行了一项用户研究。对于每位参与者，从整个GSO数据集中随机选择了45个形状，并为每个形状随机抽取了两种方法。参与者被要求从每对比较结果中选择质量更优且更符合输入图像的结果。然后根据这些选择统计所有方法的偏好率。总共从53名参与者中收集了2,385对评估结果。

结果：如表1所示，One-2-3-45++ 超越

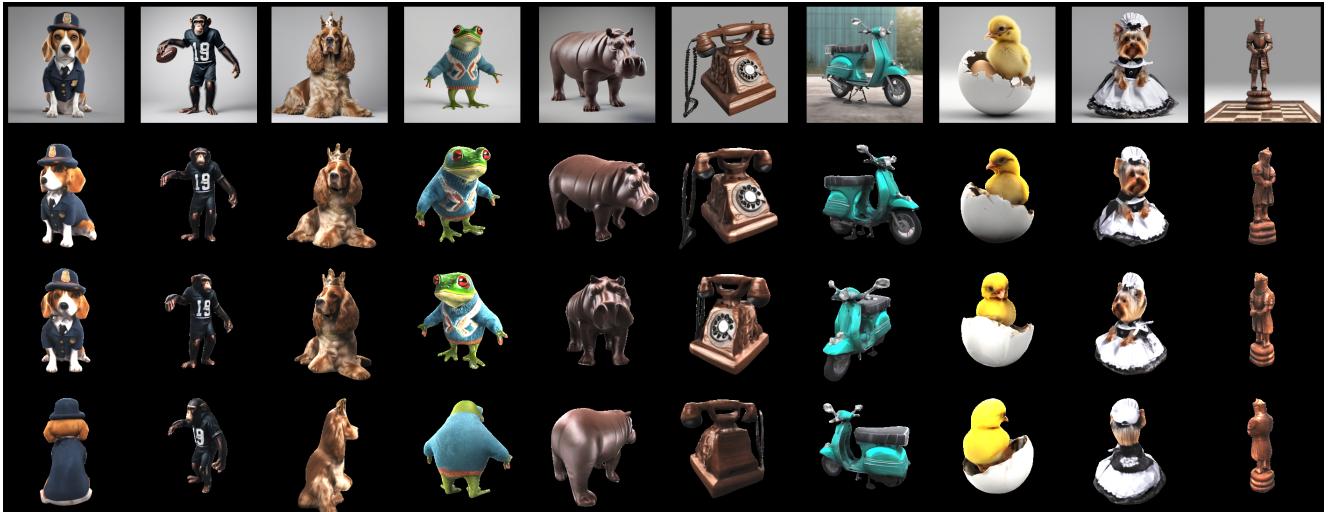


图7. 我们的定性结果：顶行显示输入图像；后续行展示生成的网格的多视角渲染。

表2. 与各种文本到3D方法的定量比较。

在DreamFusion [50]中的50个文本提示上进行了评估。

| Method | CLIP-Sim | User-Pref. | Runtime |
|----------------------|----------|------------|---------|
| ProlificDreamer [71] | 25.7 | 39.5 | 10h + |
| MVDream [62] | 24.8 | 66.2 | 2h |
| Shap-E [25] | 22.3 | 11.1 | 27s |
| Ours | 26.8 | 84.1 | 60s |

所有基线方法在F-Score和CLIP相似性方面的表现。用户偏好分数进一步突显了显著的性能差异，我们的方法大幅超越了竞争方法。请参阅图6的深入混淆矩阵，其中展示了One-2-3-45++在92%的情况下优于One-2-3-45。此外，与基于优化的方法相比，我们的方法在运行时间上表现出显著优势。图5和图7展示了定性结果。

4.2. 文本到3D的对比

基线方法：我们将One-2-3-45++与基于优化的方法进行了比较，具体包括ProlificDreamer [71]和MVDream [62]，以及前馈方法Shap-E [25]。对于ProlificDreamer，我们使用了ThreeStudio实现[18]，而对于其他方法，我们采用了各自官方的实现。

数据集与评估指标：鉴于许多基线方法需要数小时才能生成一个3D形状，我们的评估基于从DreamFusion [50]中抽样的50个文本提示进行。我们采用CLIP相似度，通过比较预测网格的24个渲染视图与输入文本提示，然后对所有视图的相似度分数进行平均计算。

用户研究：类似于图像到3D的评估，用户研究涉及每位参与者随机选择的30对结果。总共从53名参与者中收集了1,590对评估结果。

表3. 不同模块的消融研究。评估基于

complete GSO [13] 数据集。“多视图”，“重建”，以及“纹理”表示多视角生成，稀疏视角重建tion，以及纹理细化模块，分别。

| MultiView | Reconstruction | Texture | ↓-Sc. | CLIP-Sim | Time |
|-----------------|-----------------|---------|-------|----------|------|
| Zero123 XL [10] | Ours | w/o | 92.9 | 71.9 | 14s |
| Ours | SparseNeuS [39] | w/o | 81.2 | 67.2 | 15s |
| Ours | Ours | w/o | 93.6 | 73.4 | 20s |
| Ours | Ours | w, | 93.6 | 81.0 | 60s |

结果：如表2所示，One-2-3-45++在CLIP相似性方面优于所有基线方法。这一点通过用户偏好评分进一步得到证实，我们的方法显著超越了竞争对手的技术。详见图6的深入分析。在直接比较One-2-3-45++与第二佳方法MVDream [62]时，我们的方法获得了70%的用户偏好率。此外，尽管我们的方法能迅速得出结果，MVDream [62]生成单个形状需要大约2小时。图8展示了定性结果。

4.3. 分析

整体流水线One-2-3-45++由三个关键模块组成：一致的多视角生成、多视角条件下的3D扩散和纹理细化。我们在完整的GSO数据集[13]上对这些模块进行了消融研究，结果详见表3。将我们的一致多视角生成模块替换为Zero123XL[10]后，性能显著下降。此外，将我们的3D扩散模块替换为One-2-3-45[34]中使用的通用NeRF，性能下降更为显著。然而，加入我们的纹理细化模块显著提升了纹理质量，获得了更高的CLIP相似度评分。

消融研究 表4展示了3D扩散的

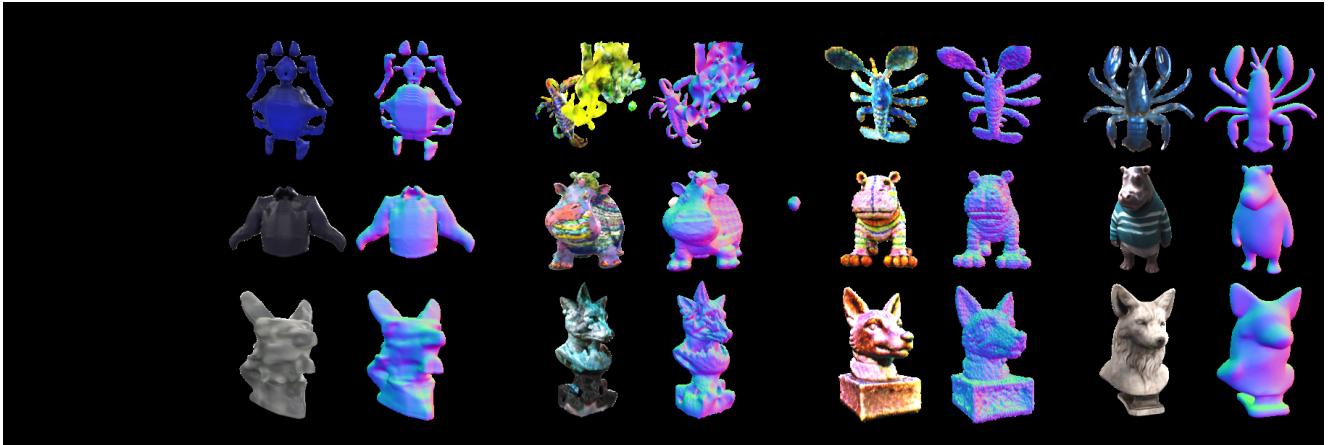


图8. 各种文本到3D方法的定性结果。展示了输入图像、纹理网格和法线贴图。

表4. 3D扩散模块的消融研究。3D IoU值

初始阶段入住率预测已报告。请注意，
3D IoU 是针对3D壳体计算的，不包括实体内部。

| id | multi-view cond. | global cond. | image source | proj. | perturb. | 3D IoU |
|--------|------------------|--------------|--------------|-------|----------|--------|
| w/o | w/ | rendering | N/A | | | 18.3 |
| global | w/ | rendering | N/A | | | 24.4 |
| local | w/o | rendering | w/o | | | 41.4 |
| local | w/ | prediction | w/o | | | 41.9 |
| local | w/ | rendering | w/o | | | 44.1 |
| local | w/ | rendering | w/ | | | 45.1 |

3D扩散模块消融研究的结果。该研究强调了多视图图像对于模块有效性的重要性。当模块在没有多视图条件下运行时，仅依赖单一输入视图的全局CLIP特征（行a和f），性能显著下降。相反，One-2-3-45++方法通过构建具有已知投影矩阵的3D特征体积，利用了多视图局部特征。仅将多个视图的全局CLIP特征简单连接也会损害性能（行b和f），突显了多视图局部条件的重要性。然而，输入视图的全局CLIP特征提供了全局形状语义；它们的移除会导致性能下降（行c和e）。尽管One-2-3-45++使用预测的多视图图像进行3D重建，但在3D扩散模块的训练过程中加入这些预测图像可能导致性能下降（行d和e），因为预测的多视图图像与实际的3D真值网格之间可能存在不匹配。为了有效训练模块，我们使用真值渲染。认识到预测的多视图图像可能存在缺陷，我们在训练过程中对投影矩阵引入随机扰动，以增强处理预测多视图图像时的鲁棒性（行e和f）。多视图生成对比我们还评估了我们的多视图生成模块与现有方法的对比，包括Zero123[35]及其扩展变体[10]，以及两个同期工作：Sync-

表5. 不同多视角生成方法的比较。

在完整的GSO [13]数据集上进行评估。

| | Target Elevations | PSNR | LPIPS | Mask IoU |
|--------------------|-------------------|-------|-------|----------|
| Zero123 [35] | | 20.32 | 0.110 | 0.856 |
| Zero123 XL [10] | 30 and - 20 | 20.11 | 0.113 | 0.869 |
| Ours | | 22.12 | 0.110 | 0.878 |
| SyncDreamer [37] | 30 | 21.67 | 0.095 | 0.894 |
| Wonder3D [40] | | 18.67 | 0.130 | 0.635 |

Dreamer [37] 和 Wonder3D [40]。我们的比较使用了 GSO [13] 数据集，其中每个对象我们渲染一张输入图像，并要求这些方法生成多视角图像。对于 Zero123 和 Zero123 XL，我们使用了与我们方法相同的目标姿态。然而，对于 Wonder3D 和 SyncDreamer，我们采用了这些方法预设的目标姿态，因为它们在推理过程中不支持改变相机位置。如表 5 所示，我们的方法在 PSNR、LPIPS 和前景掩码 IoU 方面超过了当前的方法。值得注意的是，Wonder3D [40] 在训练阶段使用了正交投影，这在处理推理过程中的透视图像时降低了其鲁棒性。SyncDreamer [37] 仅在 30° 的仰角生成视图，这比我们的设置更为简单。此外，由于这些指标不评估视图之间的三维一致性，请参阅补充材料以获取更多定性比较和讨论。

5. 结论

在本文中，我们介绍了One-2-3-45++，这是一种创新的方法，能够将任意物体的单张图像转换为带纹理的3D网格。与现有的文本到3D模型相比，这种方法提供了更精确的控制，并且能够快速生成高质量的网格——通常在60秒以内。此外，生成的网格对原始输入图像表现出高度的保真度。展望未来，通过结合2D扩散模型的额外引导条件以及RGB图像，有望进一步提升几何结构的鲁棒性和细节。

致谢

我们感谢Google Cloud和Lambda Labs在提供计算资源方面的宝贵支持。

参考文献

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, 和 Leonidas Guibas. 学习三维点云的表示和生成模型。在国际机器学习会议上，第40-49页。PMLR, 2018. 2
- [2] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, 和 Gordon Wetzstein. Genvs: 使用3D感知扩散模型的生成新颖视角合成，2023. 3
- [3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, 和 Hao Su. Mvsnerf: 从多视图立体视觉快速泛化的辐射场重建。在IEEE/CVF国际计算机视觉会议上，第14124-14133页，2021. 3
- [4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, 和 Hao Su. Tensorf: 张量辐射场。在欧洲计算机视觉会议上，第333-350页。Springer, 2022. 5
- [5] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, 和 Matthias Nießner. Text2tex: 通过扩散模型的文本驱动纹理合成。arXiv预印本arXiv:2303.11396, 2023. 2
- [6] Rui Chen, Yongwei Chen, Ningxin Jiao, 和 Kui Jia. Fantasia3d: 解耦几何和外观以实现高质量文本到3D内容创建。arXiv预印本arXiv:2303.13873, 2023. 2
- [7] Zilong Chen, Feng Wang, 和 Huaping Liu. 使用高斯喷射的文本到3D。arXiv预印本arXiv:2309.16585, 2023. 2
- [8] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, 和 Liang-Yan Gui. Sdfusion: 多模态3D形状补全、重建和生成。在IEEE/CVF计算机视觉和模式识别会议上，第4456-4465页，2023. 2
- [9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, 和 Silvio Savarese. 3d-r2n2: 单视图和多视图3D物体重建的统一方法。在计算机视觉-ECCV 2016: 第14届欧洲会议，荷兰阿姆斯特丹，2016年10月11-14日，第VIII部分14，第628-644页。Springer, 2016. 2
- [10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian LaForte, Vikram Voleti, Samir Yitzhak Gadre, 等. Objaverse-xl: 1000万个以上3D对象的宇宙。arXiv预印本arXiv:2307.05663, 2023. 6, 7, 8
- [11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, 和 Ali Farhadi. Objaverse: 一个带有注释的3D对象宇宙。在IEEE/CVF计算机视觉和模式识别会议上，第13142-13153页，2023. 4, 5
- [12] 邓聪悦, 江驰宇, Charles R Qi, 闫欣辰, 周寅, Leonidas Guibas, Dragomir Anguelov 等. Nerdi: 利用语言引导的扩散作为通用图像先验进行单视图NeRF合成。在IEEE/CVF计算机视觉与模式识别会议上，页码20637–20647, 2023. 2
- [13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, 和 Vincent Vanhoucke. Google扫描对象: 一个高质量的3D扫描家用物品数据集。在2022年国际机器人与自动化会议(ICRA)上，页码2553–2560. IEEE, 2022. 6, 7, 8
- [14] Ziya Erkoc, Fangchang Ma, Qi Shan, Matthias Nießner, 和 Angela Dai. Hyper扩散: 通过权重空间扩散生成隐式神经场。arXiv预印本arXiv:2303.17015, 2023. 2
- [15] Haoqiang Fan, Hao Su, 和 Leonidas J Guibas. 用于从单张图像重建3D对象的点集生成网络。在IEEE计算机视觉与模式识别会议上，页码605–613, 2017. 2
- [16] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daqing Li, Or Litany, Zan Gojcic, 和 Sanja Fidler. Get3D: 从图像中学习高质量3D纹理形状的生成模型。在神经信息处理系统进展中, 35:31841–31854, 2022. 2
- [17] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, 和 Mathieu Aubry. 一种学习3D表面生成的纸糊方法。在IEEE计算机视觉与模式识别会议上，页码216–224, 2018. 2
- [18] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian LaForte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, 和 Song-Hai Zhang. three-studio: 一个统一的3D内容生成框架。https://github.com/threestudio-project/threestudio, 2023. 6, 7
- [19] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, 和 Barlas O'guz. 3DGen: 用于纹理网格生成的三平面潜在扩散。arXiv预印本arXiv:2303.05371, 2023. 2
- [20] Jonathan Ho, Ajay Jain, 和 Pieter Abbeel. 去噪扩散概率模型。在神经信息处理系统进展中, 33:6840–6851, 2020. 5
- [21] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, 和 Ziwei Liu. AvatarClip: 零样文本驱动的3D头像生成与动画。arXiv预印本arXiv:2205.08535, 2022. 2
- [22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, 和 Weizhu Chen. LoRA: 大型语言模型的低秩适应。arXiv预印本arXiv:2106.09685, 2021. 4
- [23] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, 和 Ben Poole. 零样文本引导的对象生成与梦幻场。在IEEE/CVF计算机视觉与模式识别会议上，页码867–876, 2022. 2
- [24] Mohammad Mahdi Johari, Yann Lepoittevin, 和 François Fleuret. GeoNeRF: 利用几何先验泛化NeRF。在IEEE/CVF计算机视觉与模式识别会议上，

2022年计算机视觉与模式识别会议，第18365-18375页。

3

- [25] Heewoo Jun和Alex Nichol。Shap-e：生成条件3D隐函数。arXiv预印本arXiv:2305.02463，2023年。2, 6, 7
- [26] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, 和 Jitendra Malik。从图像集合中学习类别特定的网格重建。在欧洲计算机视觉会议（ECCV）的论文集上，第371-386页，2018年。2
- [27] Animesh Karnewar, Andrea Vedaldi, David Novotny, 和 Niloy J Mitra。Holodfusion：使用2D图像训练3D扩散模型。在IEEE/CVF计算机视觉与模式识别会议的论文集上，第18423-18433页，2023年。3
- [28] Jon'a's Kulhanek, Erik Derner, Torsten Sattler, 和 Robert Babuška。Viewformer：利用变压器从少量图像中进行无NeRF的神经渲染。在欧洲计算机视觉会议的论文集上，第198-216页。Springer, 2022年。3
- [29] Han-Hung Lee和Angel X Chang。理解纯Clip指导对于体素网格NeRF模型的应用。arXiv预印本arXiv:2209.15172，2022年。2
- [30] Muheng Li, Yueqi Duan, Jie Zhou, 和 Jiwen Lu。Diffusion-sdf：通过体素化扩散实现文本到形状的转换。在IEEE/CVF计算机视觉与模式识别会议的论文集上，第12642-12651页，2023年。2
- [31] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, 和 Tsung-Yi Lin。Magic3d：高分辨率文本到3D内容创建。在IEEE/CVF计算机视觉与模式识别会议的论文集上，第300-309页，2023年。2
- [32] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, 和 Xiu Li。Consistent123：使用案例感知扩散先验将一张图像转换为高度一致的3D资产。arXiv预印本arXiv:2309.17261，2023年。3
- [33] Minghua Liu, Minhyuk Sung, Radomir Mech, 和 Hao Su。Deepmetahandles：使用双调和坐标学习3D网格的变形元手柄。在IEEE/CVF计算机视觉与模式识别会议的论文集上，第12-21页，2021年。2
- [34] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, 等。One-2-3-45：在45秒内将任何单张图像转换为3D网格，无需每形状优化。arXiv预印本arXiv:2306.16928，2023年。2, 6, 7
- [35] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, 和 Carl Vondrick。Zero-1-to-3：零样本单张图像到3D对象的转换。在IEEE/CVF国际计算机视觉会议的论文集上，第9298-9309页，2023年。2, 5, 8
- [36] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, 和 Wenping Wang。用于遮挡感知基于图像渲染的神经光线。在IEEE/CVF计算机视觉与模式识别会议的论文集上，第7824-7833页，2022年。3

[37] 刘源, 林成, 曾子娇, 龙晓晓, 刘凌杰, Taku Komura, 和 王文平。Syncdreamer: 从单视图图像生成多视图一致图像。arXiv预印本 arXiv:2309.03453, 2023. 3, 5, 6, 8

[38] 刘振, 冯瑶, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, 和 刘伟洋。Meshfusion: 基于分数的生成3D网格建模。arXiv预印本 arXiv:2303.08133, 2023. 2

[39] 龙晓晓, 林成, 王鹏, Taku Komura, 和 王文平。Sparseneus: 从稀疏视图快速泛化神经表面重建。在欧洲计算机视觉会议上，页码210–227. Springer, 2022. 2, 3, 7

[40] 龙晓晓, 郭元辰, 林成, 奚志扬, 刘凌杰, 马悦欣, 张松海, Marc Habermann, Christian Theobalt, 和 王文平。Wonder3d: 使用跨域扩散从单张图像生成3D, 2023. 3, 8

[41] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, 和 Andrea Vedaldi。Realfusion: 从单张图像进行360度物体重建。在IEEE/CVF计算机视觉与模式识别会议上，页码8446–8455, 2023. 2

[42] Luke Melas-Kyriazi, Christian Rupprecht, 和 Andrea Vedaldi。Pc2: 用于单张图像3D重建的投影条件点云扩散。在IEEE/CVF计算机视觉与模式识别会议上，页码12923–12932, 2023. 2

[43] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, 和 Andreas Geiger。占用网络: 在函数空间中学习3D重建。在IEEE/CVF计算机视觉与模式识别会议上，页码4460–4470, 2019. 2

[44] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, 和 Daniel Cohen-Or。潜在-nerf用于形状引导的3D形状和纹理生成。在IEEE/CVF计算机视觉与模式识别会议上，页码12663–12673, 2023. 2

[45] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, 和 Rana Hanocka。Text2mesh: 文本驱动的神经风格化网格。在IEEE/CVF计算机视觉与模式识别会议上，页码13492–13502, 2022.

[46] Nasir Mohammad Khalid, 谢天豪, Eugene Belilovsky, 和 Tiberiu Popa。Clip-mesh: 使用预训练的图像-文本模型从文本生成纹理网格。在SIGGRAPH Asia 2022会议论文中，页码1–8, 2022. 2

[47] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, 和 Peter Battaglia。Polygen: 3D网格的自回归生成模型。在国际机器学习会议上，页码7220–7229. PMLR, 2020. 2

[48] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, 和 Mark Chen。Point-e: 从复杂提示生成3D点云的系统。arXiv预印本 arXiv:2212.08751, 2022. 2

[49] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, 和 Steven Lovegrove。Deepsdf: 学习连续有符号距离函数以进行形狀表示。

在IEEE/CVF计算机视觉与模式识别会议论文集，第165-174页，2019年。2

- [50] Ben Poole, Ajay Jain, Jonathan T Barron, 和 Ben Mildenhall. Dreamfusion: 使用2D扩散进行文本到3D生成。arXiv预印本arXiv:2209.14988, 2022年。2, 5, 7
- [51] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov等。Magic123: 使用2D和3D扩散先验从一张图像生成高质量3D物体。arXiv预印本arXiv:2306.17843, 2023年。2
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark等。从自然语言监督中学习可迁移的视觉模型。在国际机器学习会议论文集，第8748-8763页。PMLR, 2021年。2
- [53] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron等。Dreambooth3D: 主题驱动的文本到3D生成。arXiv预印本arXiv:2303.13508, 2023年。2
- [54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, 和 Ilya Sutskever. 零样文本到图像生成。在国际机器学习会议论文集，第8821-8831页。PMLR, 2021年。2
- [55] Yufan Ren, Tong Zhang, Marc Pollefeys, Sabine Süsstrunk, 和 Fangjinhua Wang. Volrecon: 使用有符号光线距离函数的体积渲染进行通用多视图重建。在IEEE/CVF计算机视觉与模式识别会议论文集，第16685-16695页，2023年。3
- [56] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, 和 Daniel Cohen-Or. Texture: 文本引导的3D形状纹理化。arXiv预印本arXiv:2302.01721, 2023年。2
- [57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, 和 Björn Ommer. 高分辨率图像合成与潜在扩散模型，2021年。2
- [58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, 和 Björn Ommer. 高分辨率图像合成与潜在扩散模型。在IEEE/CVF计算机视觉与模式识别会议论文集，第10684-10695页，2022年。2
- [59] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans等。具有深度语言理解的逼真文本到图像扩散模型。神经信息处理系统进展, 35:36479-36494, 2022年。2
- [60] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, 和 Kamal Rahimi Malekshan. Clip-forge: 零样文本到形状生成。在IEEE/CVF计算机视觉与模式识别会议论文集，第18603-18613页，2022年。2
- [61] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee,

和Seungryong Kim。让2D扩散模型了解3D一致性以实现稳健的文本到3D生成。arXiv预印本arXiv:2303.07937, 2023. 2

- [62] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, 和 Xiao Yang. MVDream: 用于3D生成的多视图扩散。arXiv预印本arXiv:2308.16512, 2023. 3, 7
- [63] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, 和 Gang Zeng. DreamGaussian: 用于高效3D内容创建的生成高斯溅射。arXiv预印本arXiv:2309.16653, 2023. 2, 5, 6
- [64] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, 和 Dong Chen. Make-it-3D: 利用扩散先验从单张图像进行高保真3D创建。arXiv预印本arXiv:2303.14184, 2023. 2
- [65] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Reznikov, Joshua B Tenenbaum, Frédo Durand, William T Freeman, 和 Vincent Sitzmann。通过前向模型进行扩散：在没有直接监督的情况下解决随机逆问题。arXiv预印本arXiv:2306.11719, 2023. 3
- [66] Alex Trevithick 和 Bo Yang. GRF: 学习用于3D表示和渲染的通用辐射场。在IEEE/CVF国际计算机视觉会议上，第15182-15192页，2021年。
- [67] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, 和 Greg Shakhnarovich. 分数雅可比链：将预训练的2D扩散模型提升为3D生成。在IEEE/CVF计算机视觉和模式识别会议上，第12619-12629页，2023. 2
- [68] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, 和 Yu-Gang Jiang. Pixel2Mesh: 从单张RGB图像生成3D网格模型。在欧洲计算机视觉会议(ECCV)上，第52-67页，2018. 2
- [69] Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, 等。注意力是NERF所需的一切吗？arXiv预印本arXiv:2207.13298, 2022. 3
- [70] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, 和 Thomas Funkhouser。IBRNet: 学习基于多视图图像的渲染。在IEEE/CVF计算机视觉和模式识别会议上，第4690-4699页，2021. 3
- [71] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, 和 Jun Zhu. ProlificDreamer: 通过变分分数蒸馏实现高保真和多样化的文本到3D生成。arXiv预印本arXiv:2305.16213, 2023. 2, 7
- [72] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, 和 Lei Zhang. Consistent123: 改进从单张图像到3D对象合成的连续性。arXiv预印本arXiv:2310.08092, 2023. 3
- [73] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, 和 Georgia Gkioxari。用于3D重建的多视图压缩编码。在IEEE/CVF计算机视觉和模式识别会议上，第9065-9075页，2023. 2
- [74] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, 和 Josh Tenenbaum。MarrNet: 3D形状重建。

通过2.5D草图进行重建。神经信息处理系统进展，30卷，2017年。2

[75] 谢浩哲，姚鸿勋，孙晓帅，周上琛，张圣平。Pix2vox：从单视图和多视图图像进行上下文感知的3D重建。在IEEE/CVF国际计算机视觉会议上，2019年，第2690–2698页。2

[76] 徐德嘉，姜一凡，王培豪，范志文，王毅，王章阳。Neurallift-360：将野外2D照片提升为具有360度视图的3D物体。在IEEE/CVF计算机视觉与模式识别会议上，2023年，第4479–4489页。2

[77] 徐嘉乐，王新涛，程伟豪，曹彦佩，单颖，邝小虎，高盛华。Dream3d：利用3D形状先验和文本到图像扩散模型进行零样本文本到3D合成。在IEEE/CVF计算机视觉与模式识别会议上，2023年，第20908–20918页。2

[78] 徐强康，王伟悦，杜杜·切兰，梅赫罗米尔·梅赫，乌尔里希·纽曼。DISN：用于高质量单视图3D重建的深度隐式表面网络。神经信息处理系统进展，32卷，2019年。2

[79] 杨浩，洪兰青，李傲雪，胡天阳，李振国，李金和，王立伟。Contranerf：通过对比学习实现合成到真实新颖视图合成的可泛化神经辐射场。在IEEE/CVF计算机视觉与模式识别会议上，2023年，第16508–16517页。3

[80] 叶江龙，王鹏，李克杰，施义春，王恒。Consistent-1-to-3：通过几何感知扩散模型实现一致的图像到3D视图合成。arXiv预印本arXiv:2310.03020，2023年。3

[81] 余朝辉，周奇，李敬良，张哲，王志斌，王凡。Points-to-3d：弥合稀疏点与可控文本到3D生成之间的差距。arXiv预印本arXiv:2307.13908，2023年。2

[82] 王宇，钱雪林，霍敬阳，黄铁军，赵波，付彦伟。推动大规模3D形状生成的极限。arXiv预印本arXiv:2306.11510，2023年。2

[83] 曾晓辉，阿沙·瓦赫达特，弗朗西斯·威廉姆斯，赞·戈伊奇，奥尔·利塔尼，桑贾·菲德勒，卡斯滕·克雷布斯。Lion：用于3D形状生成的潜在点扩散模型。arXiv预印本arXiv:2210.06978，2022年。2

[84] 张彪，唐家鹏，马蒂亚斯·尼斯特纳，彼得·旺卡。3dshape2vecset：用于神经场和生成扩散模型的3D形状表示。arXiv预印本arXiv:2301.11445，2023年。2

[85] 张吕敏。仅参考控制。<https://github.com/Mikubill/sd-webui-controlnet/discussions/1236>，2023年。4

[86] 赵子博，刘文，陈鑫，曾贤芳，王锐，程培，付斌，陈涛，余刚，高盛华。米开朗基罗：基于形状-图像-文本对齐潜在表示的条件3D形状生成。arXiv预印本arXiv:2306.17115，2023年。2

[87] 郑新阳，潘浩，王鹏舒，佟鑫，刘洋，沈向洋。局部注意力SDF

可控三维形状生成的扩散方法。arXiv预印本
arXiv:2305.04461，2023. 2, 4, 5

[88] Zhizhuo Zhou和Shubham Tulsiani。Sparsefusion：提取视图条件扩散以进行三维重建。在IEEE/CVF计算机视觉与模式识别会议论文集，第12588–12597页，2023年。3