

Zero-1-to-3: 从单张图像到三维物体的零样本生成

Zero-1-to-3: Zero-shot One Image to 3D Object

Ruoshi Liu¹ Rundi Wu¹ Basile Van Hoorick¹ Pavel Tokmakov² Sergey Zakharov² Carl Vondrick¹

¹ Columbia University ² Toyota Research Institute

zero123.cs.columbia.edu

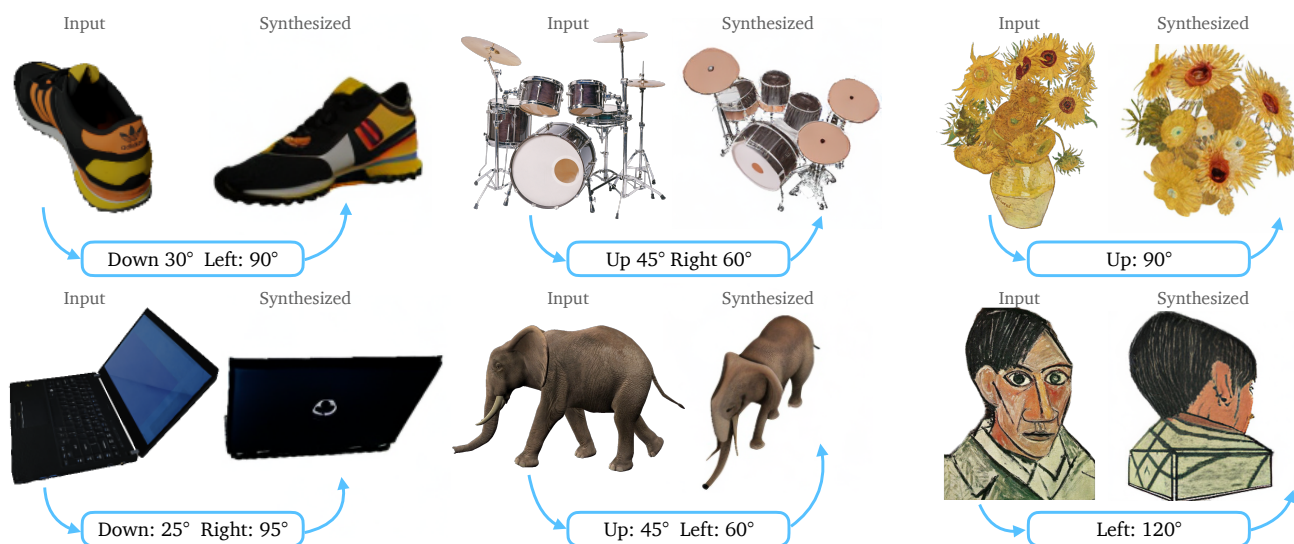


图 1: 给定物体的单张 RGB 图像, 我们提出 *Zero-1-to-3*——一种从指定相机视角合成图像的方法。即便相对变换很大, 我们的方法也能合成出富含细节、与输入视图保持一致的视角。它对几何复杂、风格艺术化的物体也取得了很强的零样本表现。

摘要

我们提出 *Zero-1-to-3*, 一个仅凭单张 RGB 图像就能改变物体相机视角的框架。要在这样一个欠约束的设定下完成新视角合成, 我们借助大规模扩散模型从自然图像中学到的几何先验。我们的条件扩散模型在合成数据集上学习对相对相机视角的控制, 从而能在指定的相机变换下为同一物体生成新图像。尽管只在合成数据上训练, 模型对分布外数据集乃至真实世界图像——印象派画作也包括在内——仍保持了很强的零样本泛化能力。这种以视角为条件的扩散方法还可进一步用于单图像三维重建。定性与定量实验表明, 借

助互联网规模的预训练, 我们的方法大幅超越了当前最先进的单视图三维重建与新视角合成模型。

1. 引言

仅凭单一相机视角, 人类往往就能想象出物体的三维形状与外观。这种能力对日常任务很重要, 比如物体操控 [17] 和在复杂环境中导航 [7], 对绘画 [32] 这样的视觉创造也至关重要。它部分可归因于对称性等几何先验, 但人类似乎能轻松推广到更具挑战性的物体, 哪怕它们打破了物理和几何约束。事实上, 我们能预测出那些在物理世界并不存在、甚至不可能存在的物体的三维形状 (见图 1 第三列)。要达到如此程度的泛化,

人类依靠的是一生视觉探索所积累的先验知识。

相比之下，现有的三维图像重建方法大多工作在封闭世界设定下，因为它们依赖昂贵的三维标注（如 CAD 模型）或特定类别的先验 [37, 21, 36, 67, 68, 66, 25, 24]。最近，一些方法在开放世界三维重建方向上取得重大进展，做法是在 CO3D [43, 30, 36, 15] 等大规模、多样化的数据集上预训练。但这些方法在训练时通常仍需要几何相关信息，例如立体视图或相机位姿。因此，与近期支撑起大型扩散模型成功 [45, 44, 33] 的互联网规模图文数据集 [47] 相比，它们所用数据的规模和多样性仍显得微不足道。已有研究表明，互联网规模的预训练赋予了这些模型丰富的语义先验，但它们在多大程度上捕捉到几何信息，至今仍鲜有探索。

本文表明，我们能够学到一种控制机制，去操纵 Stable Diffusion [44] 这类大规模扩散模型中的相机视角，从而完成零样本的新视角合成与三维形状重建。给定单张 RGB 图像，这两项任务都是严重欠约束的。不过，得益于现代生成模型可用训练数据的规模（超过 50 亿张图像），扩散模型是当今自然图像分布的最先进表示，其覆盖范围囊括了大量物体的众多视角。这些模型虽然训练于二维单目图像、没有任何相机对应关系，我们仍可对其微调，让它在生成过程中学会控制相对的相机旋转与平移。借助这些控制，我们可以把任意图像编码，再将其解码到我们指定的另一相机视角。图 1 给出了若干结果示例。

本文的主要贡献在于证明：大型扩散模型虽然只在二维图像上训练，却已学到关于视觉世界的丰富三维先验。我们还展示了新视角合成的最先进结果，以及从单张 RGB 图像对物体做零样本三维重建的最先进结果。下文先在第 2 节简要回顾相关工作；第 3 节介绍我们如何通过微调大型扩散模型来学习对相机外参的控制；最后在第 4 节给出若干定量与定性实验，评估从单张图像出发的零样本视角合成，以及几何与外观的三维重建。我们将公开全部代码、模型以及一个在线演示。

2. 相关工作

三维生成模型。 生成式图像架构的最新进展，加上大规模图文数据集 [47]，已能合成高保真且多样的场景与物体 [33, 40, 45]。其中，扩散模型被证明能以去噪目标高效地学习可扩展的图像生成器 [6, 48]。然而，要把



图 2: 文本到图像模型中的视角偏差。我们展示 Dall-E-2 与 Stable Diffusion v2 在提示词 “a chair” 下的采样结果。多数样本呈现的都是正面朝向、处于标准姿态的椅子。

它们扩展到三维领域，需要大量昂贵的带标注三维数据。为此，近期方法转而把预训练的大规模二维扩散模型迁移到三维，而不使用任何三维真值数据。神经辐射场 (NeRF) [31] 凭借高保真编码场景的能力，成为一种强大的表示。NeRF 通常用于单场景重建：给定覆盖整个场景的大量带位姿图像，任务是从未观测角度预测新视角。DreamFields [22] 表明 NeRF 是一种更通用的工具，还能充当三维生成系统的核心组件。一系列后续工作 [38, 26, 53] 用来自二维扩散模型的蒸馏损失替换 CLIP，把该模型改造为从文本输入生成高保真的三维物体与场景。

我们的工作探索了一条非常规的新视角合成路线，把它建模为一个以视角为条件、用扩散模型完成的图到图翻译任务。学到的模型还可与三维蒸馏结合，从单张图像重建三维形状。已有工作 [56] 采用了类似流程，但未展现零样本泛化能力。若干并行工作 [9, 29, 61] 提出了相近技术，借助语言引导的先验和文本反演 [14] 来做图到三维的生成。相比之下，我们的方法通过合成数据集学习对视角的控制，并展现出对真实世界图像的零样本泛化。

单视图物体重建。 从单一视图重建三维物体是一个高度困难的问题，需要很强的先验。一类工作依赖以网格 [58, 62]、体素 [16, 60] 或点云 [12, 30] 表示的三维基元集合来构建先验，并用图像编码器提供条件。这类模型受限于所用三维数据集的多样性，又因这种条件方式是全局性的，泛化能力较差。此外，它们还需一个额外的位姿估计步骤，以确保估计形状与输入对齐。另

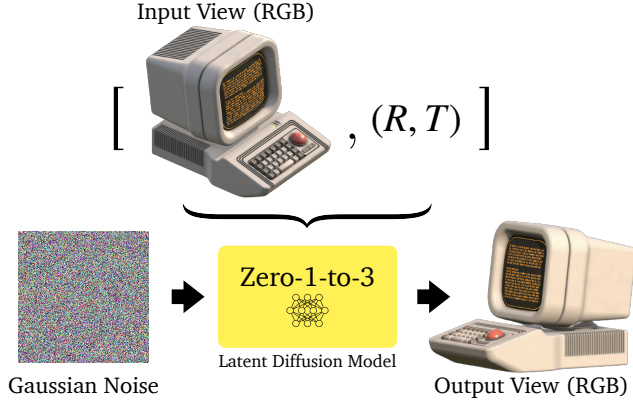


图 3: **Zero-1-to-3** 是一个以视角为条件的图像翻译模型，采用条件潜在扩散架构。输入视图和一个相对视角变换都作为条件信息使用。

一方面，局部条件模型 [46, 63, 54, 51, 52] 试图直接利用局部图像特征来重建场景，跨域泛化能力更强，但通常只能重建邻近视角。最近，MCC [59] 从 RGB-D 视图学习一种通用的三维重建表示，并在大规模的以物体为中心的视觉数据集上训练。

在本工作中，我们表明可以直接从预训练的 Stable Diffusion 模型中提取丰富的几何信息，从而免去对额外深度信息的需求。

3. 方法

给定物体的单张 RGB 图像 $x \in \mathbb{R}^{H \times W \times 3}$ ，我们的目标是从另一相机视角合成该物体的图像。设 $R \in \mathbb{R}^{3 \times 3}$ 和 $T \in \mathbb{R}^3$ 分别为目标视角的相对相机旋转与平移。我们希望学到一个模型 f ，在此相机变换下合成新图像：

$$\hat{x}_{R,T} = f(x, R, T) \quad (1)$$

其中 $\hat{x}_{R,T}$ 记为合成出的图像。我们希望估计得到的 $\hat{x}_{R,T}$ 在感知上接近真实但未被观测到的新视角 $x_{R,T}$ 。

从单目 RGB 图像做新视角合成是严重欠约束的。我们的方法将借助 Stable Diffusion 这类大型扩散模型来完成，因为它们能从文本描述生成多样图像时展现出非凡的零样本能力。得益于训练数据的规模 [47]，预训练扩散模型是当今自然图像分布的最先进表示。

然而，要构造 f ，我们必须克服两个挑战。其一，大规模生成模型虽然在大量不同视角的物体上训练，其表示却没有显式编码视角之间的对应关系。其二，生成

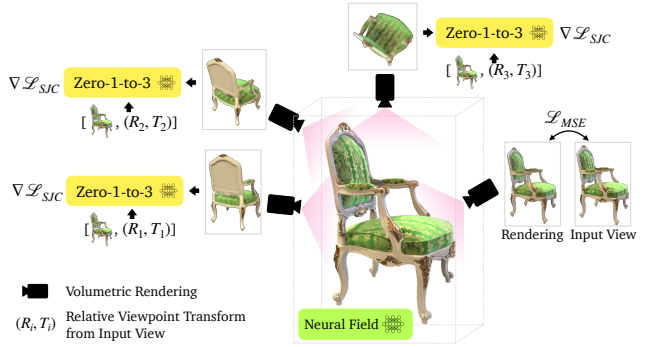


图 4: 用 **Zero-1-to-3** 做三维重建。Zero-1-to-3 可用于优化一个神经场，以完成单图像三维重建任务。训练时，我们随机采样视角，并用 Zero-1-to-3 来监督三维重建。

模型继承了互联网上反映出的视角偏差。如图 2 所示，Stable Diffusion 倾向于生成正面朝向、处于标准姿态的椅子。这两个问题极大地妨碍了从大规模扩散模型中提取三维知识。

3.1. 学习控制相机视角

由于扩散模型在互联网规模的数据上训练，其对自然图像分布的覆盖很可能已包含大多数物体的大多数视角，但在预训练模型中这些视角无法被控制。一旦我们能教会模型一种机制，去控制拍摄照片所用的相机外参，就解锁了执行新视角合成的能力。

为此，给定一组配对图像及其相对相机外参 $\{(x, x_{(R,T)}, R, T)\}$ 构成的数据集，我们的方法（见图 3）对预训练扩散模型做微调，从而在不破坏其余表示的前提下学会对相机参数的控制。沿用 [44]，我们采用潜在扩散架构，包含编码器 $\mathcal{E}$ 、去噪 U-Net ϵ_θ 和解码器 $\mathcal{D}$ 。在扩散时间步 $t \sim [1, 1000]$ ，设 $c(x, R, T)$ 为输入视图与相对相机外参的嵌入。随后我们求解以下目标来微调模型：

$$\min_{\theta} \mathbb{E}_{z \sim \mathcal{E}(x), t, \epsilon \sim \mathcal{N}(0,1)} \|\epsilon - \epsilon_\theta(z, t, c(x, R, T))\|_2^2. \quad (2)$$

模型 ϵ_θ 训练完成后，推理模型 f 便能以 $c(x, R, T)$ 为条件，从一张高斯噪声图像 [44] 出发迭代去噪，生成图像。

本文的主要结果是：以这种方式微调预训练扩散模型，能让它学到一种控制相机视角的通用机制，并可外推到微调数据集之外的物体。换言之，这种微调把控制

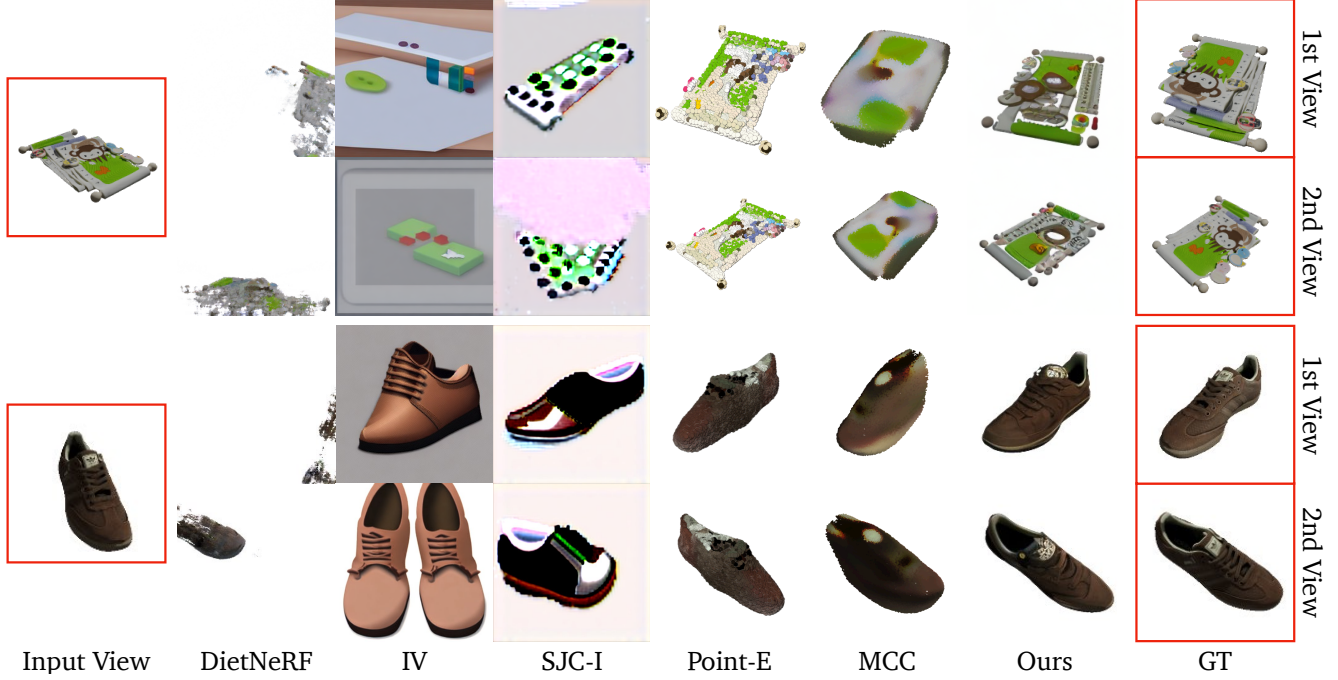


图 5: 在 Google Scanned Objects [10] 上的新视角合成。左侧的输入视图用于合成两个随机采样的新视角，右侧为对应的真值视图。与各基线相比，我们合成的新视角包含丰富的纹理与几何细节，且与真值高度一致，而基线方法则明显丢失了高频细节。

能力“外挂”上去，扩散模型仍保留生成逼真图像的能力，只是如今多了对视角的控制。这种可组合性赋予模型零样本能力：最终模型能为那些缺乏三维资产、从未在微调集中出现过的物体类别合成新视角。

3.2. 以视角为条件的扩散

从单张图像做三维重建，既需要低层感知（深度、明暗、纹理等），也需要高层理解（类别、功能、结构等）。为此，我们采用一种混合条件机制。在一条通路上，把输入图像的 CLIP [39] 嵌入与 (R, T) 拼接，构成“带位姿的 CLIP”嵌入 $c(x, R, T)$ 。我们用交叉注意力对去噪 U-Net 施加条件，为其提供输入图像的高层语义信息。在另一条通路上，把输入图像与正在去噪的图像做通道拼接，帮助模型保持所合成物体的身份与细节。为了能施加无分类器引导 [19]，我们沿用 [3] 提出的类似机制：随机地把输入图像和带位姿的 CLIP 嵌入置为零向量，并在推理时对条件信息做缩放。

3.3. 三维重建

在许多应用中，仅合成物体的新视角还不够；人们希望得到同时刻画外观与几何的完整三维重建。我们采用近期开源的 Score Jacobian Chaining (SJC) [53] 框架，用文本到图像扩散模型的先验来优化三维表示。然而，由于扩散模型的概率本质，梯度更新高度随机。SJC 受 DreamFusion [38] 启发，采用了一项关键技术：把无分类器引导的取值设得远高于常规。这种做法降低了单个样本的多样性，却提升了重建的保真度。

如图 4 所示，与 SJC 类似，我们随机采样视角并执行体渲染。随后用高斯噪声 $\epsilon \sim \mathcal{N}(0, 1)$ 扰动所得图像，并以输入图像 x 、带位姿的 CLIP 嵌入 $c(x, R, T)$ 和时间步 t 为条件，施加 U-Net ϵ_θ 去噪，从而逼近朝向无噪输入 x_π 的分数：

$$\nabla \mathcal{L}_{SJC} = \nabla_{I_\pi} \log p_{\sqrt{2}\epsilon}(x_\pi) \quad (3)$$

其中 $\nabla \mathcal{L}_{SJC}$ 是 [53] 提出的 PAAS 分数。

此外，我们用 MSE 损失优化输入视图。为进一步正则化 NeRF 表示，我们还对每个采样视角施加深度平

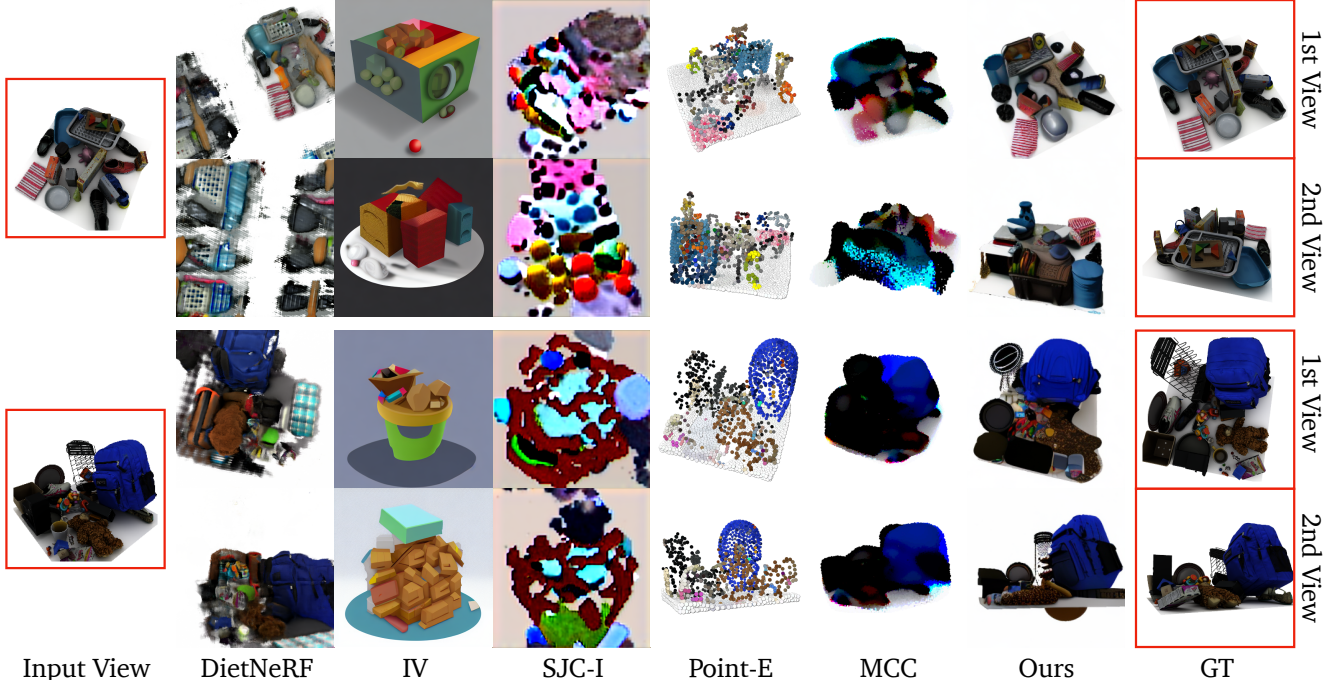


图 6: 在 RTMV [50] 上的新视角合成。左侧的输入视图用于合成 2 个随机采样的新视角，右侧为对应的真值视图。即使相机视角变化很大，我们合成的视图仍保持高保真，而多数其他方法的质量急剧下降。

滑损失，并用邻近视角一致性损失来约束相邻视角之间的外观变化。

3.4. 数据集

我们使用近期发布的 *Objaverse* [8] 数据集做微调，它是一个大规模开源数据集，包含由 10 万多名创作者制作的 80 多万个三维模型。虽然它不像 ShapeNet [4] 那样带有显式的类别标签，Objaverse 却囊括了大量高质量、几何丰富的三维模型，其中许多带有精细的细节和材质属性。对数据集中的每个物体，我们随机采样 12 个指向物体中心的相机外参矩阵 $\mathcal{M}_1$ ，并用光线追踪引擎渲染 12 个视图。训练时，可为每个物体采样两个视图，构成图像对 $(x, x_{R,T})$ 。定义两个视角之间映射关系的相对视角变换 (R, T) ，可以很容易地从这两个外参矩阵推导出来。

4. 实验

我们评估模型在零样本新视角合成与三维重建上的表现。经 Objaverse 作者确认，本文所用的数据集和图像都在 Objaverse 数据集之外，因此可视为零样本结

果。我们在不同复杂度的合成物体与场景上，将模型与最先进方法做定量比较。我们还用多样的真实世界图像给出定性结果，从我们自己拍摄的日常物品到画作皆有。

4.1. 任务

我们描述两项密切相关的任务，二者都以单视图 RGB 图像为输入，并以零样本方式应用。

新视角合成。新视角合成是计算机视觉中一个由来已久的三维问题，要求模型隐式地学习物体的深度、纹理与形状。仅有单一视图这一极其有限的输入信息，迫使新视角合成方法去利用先验知识。近来流行的方法依赖优化隐式神经网络，并对随机采样的视角施加 CLIP 一致性目标 [23]。而我们以视角为条件的图像生成方法与之正交，因为我们颠倒了三维重建与新视角合成的先后顺序，同时仍保留输入图像所描绘物体的身份。这样一来，绕物体旋转时因自遮挡带来的偶然不确定性，可由一个概率生成模型来建模，而大型扩散模型学到的语义与几何先验也能被有效利用。

三维重建。我们也可以改用 SJC [53] 或 DreamFusion [38] 这类随机三维重建框架，来构造一个最可能



图 7: 真实世界图像上的新视角合成。第 1、3、4 行是用 iPhone 拍摄图像的结果, 第 2 行是从互联网下载图像的结果。我们的方法对不同表面材质和几何的物体都很鲁棒。我们随机采样了 5 个不同视角, 未经挑选直接展示结果。更多未经筛选的结果见补充材料。

的三维表示。我们将其参数化为体素辐射场 [5, 49, 13], 随后对密度场执行 Marching Cubes 提取网格。把我们可以视角为条件的扩散模型用于三维重建, 提供了一条可行的路径, 能把该扩散模型学到的丰富二维外观先验导向三维几何。

4.2. 基线

为与我们方法的适用范围保持一致, 我们只与那些在零样本设定下工作、且以单视图 RGB 图像为输入的方法比较。

在新视角合成方面, 我们与若干最先进的单图像算法比较。具体而言, 我们对标 DietNeRF [23], 它用跨视角的 CLIP 图到图一致性损失来正则化 NeRF。此外, 我们比较了 Image Variations (IV) [1], 它是一个经微调、以图像而非文本提示为条件的 Stable Diffusion

模型, 可看作以 Stable Diffusion 为基础的语义最近邻搜索引擎。最后, 我们改造了 SJC [53]——一个基于扩散的文本到三维模型, 把其中原本以文本为条件的扩散模型替换为以图像为条件的扩散模型, 我们称之为 SJC-I。

在三维重建方面, 我们用两个最先进的单视图算法作为基线: (1) 多视图压缩编码 (MCC) [59], 一种基于神经场的方法, 把 RGB-D 观测补全为三维表示; (2) Point-E [34], 一个在带颜色点云上的扩散模型。MCC 在 CO3Dv2 [43] 上训练, 而 Point-E 值得注意的是训练于 OpenAI 一个大得多的内部三维数据集。我们同样与 SJC-I 比较。

由于 MCC 需要深度输入, 我们直接使用现成的 MiDaS [42, 41] 做深度估计。我们假定一组在整个测试集上看起来都合理的标准尺度与偏移值, 把得到的相



图 8: **新视角合成的多样性**。给定一个输入视图, 我们固定另一个视角, 随机生成多个条件样本。这些不同的结果, 反映出输入视图中缺失的几何与外观信息上的一系列多样性。

	DietNeRF [23]	Image Variation [1]	SJC-I [53]	本文		DietNeRF [23]	Image Variation [1]	SJC-I [53]	本文
PSNR $\uparrow$	<u>8.933</u>	5.914	6.573	18.378	PSNR $\uparrow$	7.130	6.561	<u>7.953</u>	10.405
SSIM $\uparrow$	<u>0.645</u>	0.540	0.552	0.877	SSIM $\uparrow$	0.406	0.442	<u>0.456</u>	0.606
LPIPS $\downarrow$	<u>0.412</u>	0.545	0.484	0.088	LPIPS $\downarrow$	<u>0.507</u>	0.564	0.545	0.323
FID $\downarrow$	<u>12.919</u>	22.533	19.783	0.027	FID $\downarrow$	<u>5.143</u>	10.218	10.202	0.319

表 1: 在 **Google Scanned Objects** 上的新视角合成结果。所有指标都表明, 我们的方法能以显著优势超越各基线。

表 2: 在 **RTMV** 上的新视角合成结果。RTMV 中的场景相对 Objaverse 训练数据属于分布外, 但我们的模型仍以显著优势超越各基线。

对视差图转换为绝对的伪度量深度图。

4.3. 基准与指标

我们在 Google Scanned Objects (GSO) [10] 和 RTMV [50] 上评估两项任务: 前者是一个高质量扫描家居物品的数据集, 后者由复杂场景构成, 每个场景包含 20 个随机物体。所有实验中, 三维重建都用各自对应的真值三维模型来评估。

在新视角合成方面, 我们用四个覆盖图像相似度不同方面的指标, 对方法与基线做全面的数值评估: PSNR、SSIM [55]、LPIPS [64] 和 FID [18]。在三维重建方面, 我们测量倒角距离 (Chamfer Distance) 和体素 IoU。

4.4. 新视角合成结果

数值结果见表 1 和表 2。图 5 表明, 在 GSO 上与所有基线相比, 我们的方法能生成高度逼真、与真值

高度一致的图像。图 6 显示, 即便 RTMV 场景相对 Objaverse 数据集属于分布外, 同样的趋势也依然成立。在各基线中, 我们观察到 Point-E 的结果往往远好于其他基线, 保持了令人印象深刻的零样本泛化能力。然而, 其生成点云规模偏小, 极大限制了 Point-E 在新视角合成上的适用性。

在图 7 中, 我们进一步展示了模型对几何与纹理都富有挑战的物体的泛化表现, 以及它在保持物体类别、身份和低层细节的同时合成高保真视角的能力。

样本间的多样性。从单张图像做新视角合成是一项严重约束的任务, 就捕捉背后的不确定性而言, 这使扩散模型相比 NeRF 成为格外合适的架构选择。由于输入图像是二维的, 它们始终只描绘了物体的局部视图, 许多部分未被观测。图 8 举例说明了从新视角采样出的、既合理又高质量的图像所具有的多样性。

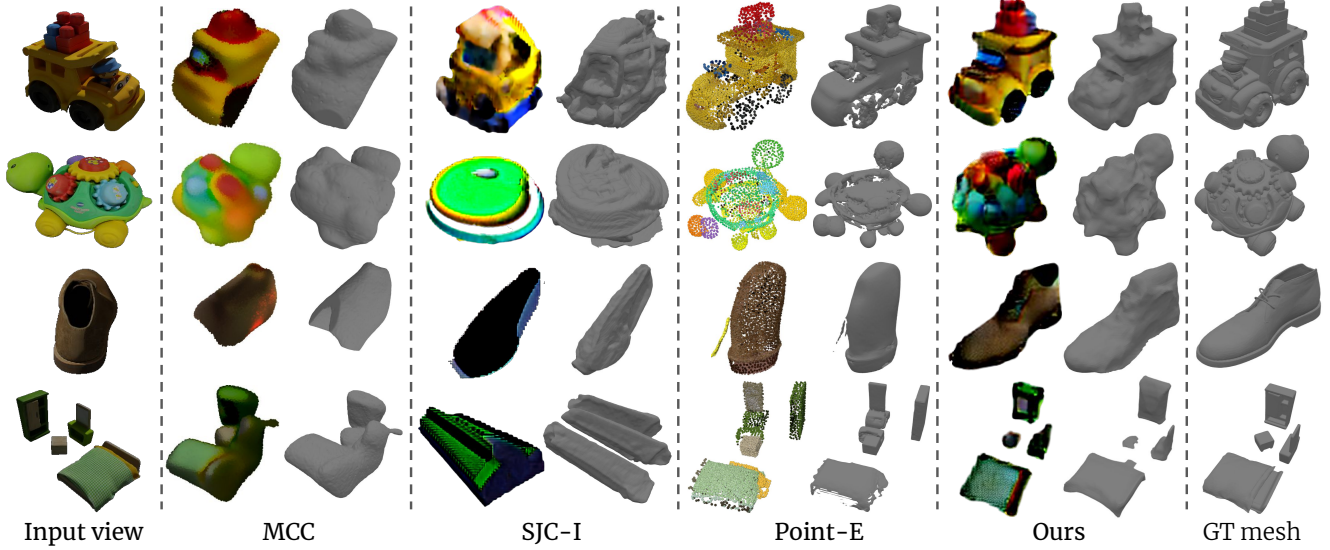


图 9: 三维重建的定性示例。输入视图图像显示在左侧。对每种方法，我们展示从另一角度渲染的视图及重建出的三维网格。真值网格显示在右侧。

	MCC [59]	SJC-I [53]	Point-E [34]	本文
CD ↓	0.1230	0.2245	<u>0.0804</u>	0.0717
IoU ↑	0.2343	0.1332	<u>0.2944</u>	0.5052

表 3: 在 GSO 上的单视图三维重建结果。注意，我们的体素 IoU 大幅优于所比较的方法。

	MCC [59]	SJC-I [53]	Point-E [34]	本文
CD ↓	0.1578	<u>0.1554</u>	0.1565	0.1352
IoU ↑	<u>0.1550</u>	0.1380	0.0784	0.2196

表 4: 在 RTMV 上的单视图三维重建结果。由于 RTMV 由包含许多物体的杂乱场景构成，所研究的方法似乎都表现欠佳。不过，尽管我们的方法并未针对三维重建任务显式训练，它仍是其中最好的。

4.5. 三维重建结果

数值结果见表 3 和表 4。图 9 定性地表明，我们的方法重建出与真值一致的高保真三维网格。MCC 往往能较好地估计从输入视图可见的表面，但常常无法正确推断物体背面的几何。

SJC-I 也常常无法重建出有意义的几何。另一方面，Point-E 具有令人印象深刻的零样本泛化能力，能对物体几何给出合理的估计。然而，它生成的是仅含 4,096 个点、分布不均的稀疏点云，（按其提供的网格转换方法）有时会在重建表面上留下孔洞。因此，它取得了不错的 CD 分数，却在体素 IoU 上表现欠佳。我们的方法利用了以视角为条件的扩散模型学到的多视图先验，并将其与 NeRF 式表示的优势结合。如表 3 和表 4 所示，这两方面因素在 CD 和体素 IoU 上都带来了相较以往工作的提升。

4.6. 从文本到图像再到三维

除真实世界图像外，我们还在 Dall-E-2 [40] 等文本到图像模型生成的图像上测试了方法。如图 10 所示，我们的模型能为这些图像生成新视角，同时保留物体的身份。我们相信这在许多文本到三维的生成应用中会非常有用。

5. 讨论

在本工作中，我们提出了一种全新的方法 *Zero-1-to-3*，用于零样本、单图像的新视角合成与三维重建。我们的方法借助 Stable Diffusion 模型——它在互联网规模的数据上预训练，捕捉到丰富的语义与几何先验。为提取这些信息，我们在合成数据上对模型做微调，学会对相机视角的控制。所得方法凭借利用 Stable Diffusion 所学强物体形状先验的能力，在多个基准上取得了最

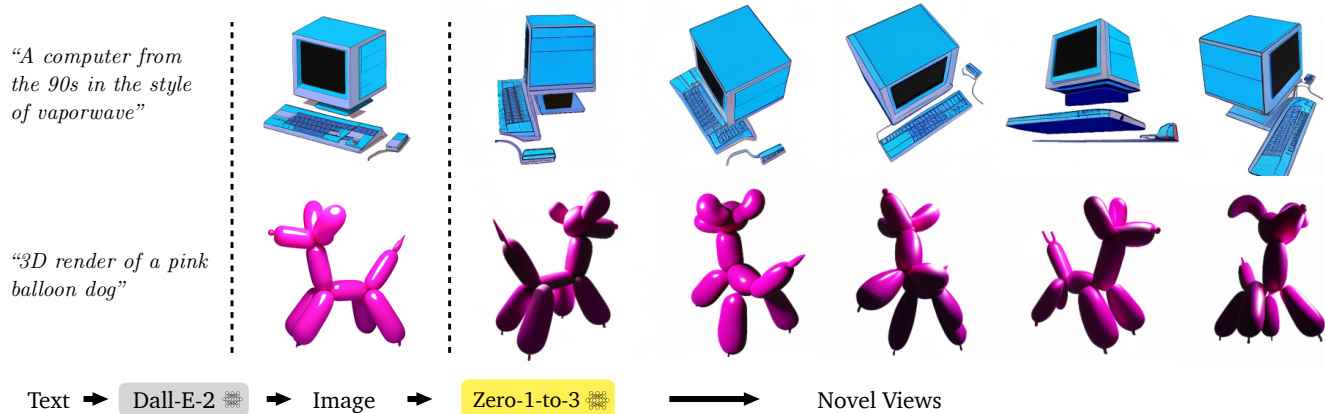


图 10: 从 Dall-E-2 生成图像的新视角合成。多个物体的构图（第 1 行）和光照细节（第 2 行），在我们合成的新视角中都得以保留。

先进的结果。

5.1. 未来工作

从物体到场景。我们的方法在单个物体、纯色背景的数据集上训练。虽然我们在 RTMV 数据集上展示了对含多个物体场景的高度泛化，但与 GSO 的分布内样本相比，质量仍有下降。因此，对复杂背景场景的泛化仍是我们方法面临的一个重要挑战。

从场景到视频。若能从单一视角推理动态场景的几何，将开辟全新的研究方向，例如理解遮挡 [52, 27] 和动态物体操控。近来已有若干基于扩散的视频生成方法 [20, 11]，把它们扩展到三维将是打开这些机遇的关键。

将图形学流程与 Stable Diffusion 结合。本文展示了一个从 Stable Diffusion 中提取物体三维知识的框架。像 Stable Diffusion 这样强大的自然图像生成模型，还隐含着关于光照、明暗、纹理等的其他知识。未来工作可以探索类似机制，去完成传统图形学任务，例如场景重光照。

致谢：我们感谢 Changxi Zheng 和 Samir Gadre 提供的宝贵反馈，也感谢 SJC [53]、NeRDi [9]、SparseFusion [65] 和 Objaverse [8] 的作者们所作的有益讨论。本研究部分得到 Toyota Research Institute、DARPA MCS 计划（联邦协议编号 N660011924032）以及 NSF NRI 奖项 #1925157 的支持。

参考文献

- [1] Stable diffusion image variations - a hugging face space by lambdalabs. 6, 7, 13
- [2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam. 13
- [3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023. 4
- [4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5
- [5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensorRF: Tensorial radiance fields. In *ECCV*, 2022. 6
- [6] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. In *ICLR*, 2021. 2
- [7] Ken Cheng. Reflections on geometry and navigation. *Connection Science*, 17(1-2):5–21, 2005. 1
- [8] 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. *arXiv preprint arXiv:2212.08051*, 2022. 5, 9, 13

- [9] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchun Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. NeRDi: Single-view NeRF synthesis with language-guided diffusion as general image priors. *arXiv preprint arXiv:2212.03267*, 2022. 2, 9
- [10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *ICRA*, 2022. 4, 7
- [11] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 9
- [12] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 2
- [13] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qin-hong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 6
- [14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2
- [15] 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 *NeurIPS*, 2022. 2
- [16] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2
- [17] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009. 1
- [18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7
- [19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2022. 4
- [20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 9
- [21] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Planes vs. chairs: Category-guided 3D shape learning without any 3D cues. In *ECCV*, 2022. 2
- [22] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022. 2
- [23] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a Diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021. 5, 6, 7, 14
- [24] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2
- [25] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 2
- [26] 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. *arXiv preprint arXiv:2211.10440*, 2022. 2
- [27] Ruoshi Liu, Sachit Menon, Chengzhi Mao, Dennis Park, Simon Stent, and Carl Vondrick. Shadows shed light on 3D objects. *arXiv preprint arXiv:2206.08990*, 2022. 9
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 13
- [29] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. RealFusion: 360° reconstruction of any object from a single image. *arXiv preprint arXiv:2302.10663*, 2023. 2
- [30] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, 2019. 2
- [31] 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. 2
- [32] Gillian M Morriss-Kay. The evolution of human artistic creativity. *Journal of anatomy*, 216(2):158–176, 2010. 1
- [33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photo-realistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2
- [34] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A system for generating 3D point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 6, 8
- [35] OPHoperHPO. Ophoperhpo/ image-background-remove-tool: automated high-quality background removal framework for an image using neural networks. 13, 14
- [36] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2
- [37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2
- [38] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023. 2, 4, 5
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4
- [40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2, 8
- [41] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 6, 14
- [42] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020. 6, 14
- [43] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordon, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*, 2021. 2, 6
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3
- [45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2
- [46] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 3
- [47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2, 3
- [48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2
- [49] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 6
- [50] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, and Stan Birchfield. RTMV: A ray-traced multi-view synthetic dataset for novel view synthesis. *ECCVW*, 2022. 5, 7

- [51] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 3
- [52] Basile Van Hoorick, Purva Tendulkar, Dídac Surís, Dennis Park, Simon Stent, and Carl Vondrick. Revealing occlusions with 4D neural fields. In *CVPR*, 2022. 3, 9
- [53] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2D diffusion models for 3D generation. In *CVPR*, 2023. 2, 4, 5, 6, 7, 8, 9
- [54] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning multi-view image-based rendering. In *CVPR*, 2021. 3
- [55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7
- [56] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 2
- [57] Wikipedia. Spherical coordinate system — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Spherical%20coordinate%20system&oldid=1142703172>, 2023. [Online; accessed 14-March-2023]. 13
- [58] Markus Worchel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, and Peter Eisert. Multi-view mesh reconstruction with neural deferred shading. In *CVPR*, 2022. 2
- [59] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3D reconstruction. *arXiv:2301.08247*, 2023. 3, 6, 8, 14
- [60] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. MarrNet: 3D shape reconstruction via 2.5D sketches. In *NeurIPS*, 2017. 2
- [61] DeJia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. NeuralLift-360: Lifting an in-the-wild 2D photo to a 3D object with 360° views. In *CVPR*, 2023. 2
- [62] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. In *NeurIPS*, 2019. 2
- [63] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 3
- [64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7
- [65] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction, 2022. 9
- [66] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-D safari: Learning to estimate zebra pose, shape, and texture from images” in the wild”. In *ICCV*, 2019. 2
- [67] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *CVPR*, 2018. 2
- [68] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, 2017. 2

附录

A. 坐标系与相机模型

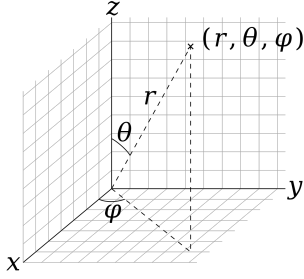


图 11: 球坐标系 [57]。

我们用球坐标系表示相机位置及其相对变换。如图 11 所示，假定物体中心为坐标系原点，我们可用 θ 、 ϕ 、 r 分别表示极角、方位角和半径（到中心的距离）。构建数据集时，我们把所有资产归一化到 XYZ 单位立方体 $[-0.5, 0.5]^3$ 内。然后采样相机视角，使 $\theta \in [0, \pi]$ 、 $\phi \in [0, 2\pi]$ 均匀覆盖单位球面， r 在区间 $[1.5, 2.2]$ 内均匀采样。训练时，当采样到来自两个不同视角的图像，设其相机位置为 (θ_1, ϕ_1, r_1) 和 (θ_2, ϕ_2, r_2) ，我们把它们的相对相机变换记为 $(\theta_2 - \theta_1, \phi_2 - \phi_1, r_2 - r_1)$ 。由于相机始终指向坐标系中心，外参矩阵由相机在球坐标系中的位置唯一确定。我们假设相机的水平视场角为 49.1° ，并采用针孔相机模型。

由于方位角的不连续性，我们用 $\phi \mapsto [\sin(\phi), \cos(\phi)]$ 对其编码。随后，在训练和推理时，都把表示相对相机视角变化的四个值 $[\theta, \sin(\phi), \cos(\phi), r]$ 连同一张输入图像一起送入模型，以生成新视角。

B. 数据集构建

我们用 Blender [2] 渲染微调数据集的训练图像。具体渲染代码沿用自 Objaverse [8] 作者公开发布的仓库¹。对 Objaverse 中的每个物体，我们随机采样 12 个视角，使用 Blender 的 Cycles 引擎、每条光线 128 个采样并配合一步去噪来渲染每张图像。所有图像以 512×512 分辨率渲染，透明背景用白色填充。我们还施加随机化的面光源照明。总计，我们为微调渲染了约 1000 万张图像的数据集。

¹<https://github.com/allenai/objaverse-rendering>

C. 微调 Stable Diffusion

我们用渲染得到的数据集微调一个预训练的 Stable Diffusion 模型来执行新视角合成。由于原始 Stable Diffusion 网络并非以多模态文本嵌入为条件，要让它能接收来自图像的条件信息，需要对原始 Stable Diffusion 架构做改动并微调。这一改动已由 [1] 完成，我们使用他们发布的检查点。为进一步让模型能同时接收来自图像的条件信息与相对相机位姿，我们把图像的 CLIP 嵌入（维度 768）与位姿向量（维度 4）拼接，并初始化另一个全连接层（ $772 \mapsto 768$ ），以确保与扩散模型架构兼容。该层的学习率放大为其他层的 $10\times$ 。网络架构的其余部分与原始 Stable Diffusion 保持一致。

C.1. 训练细节

训练使用 AdamW [28]，学习率为 10^{-4} 。起初，我们尝试在保持原始分辨率（图像尺寸 512×512 ，潜在维度 64×64 ）的情况下用批大小 192 训练。然而，我们发现这导致收敛更慢、批次间方差更大。由于原始 Stable Diffusion 的训练流程使用批大小 3072，我们随后把图像尺寸减小到 256×256 （相应地潜在维度降为 32×32 ），以便把批大小增大到 1536。批大小的提升带来了更好的训练稳定性和显著改善的收敛速度。我们在一台 $8 \times A100-80GB$ 的机器上微调模型 7 天。

C.2. 推理细节

在一块 RTX A6000 GPU 上，Zero-1-to-3 生成一个新视角仅需 2 秒。注意，以往工作通常要训练一个 NeRF 才能渲染新视角，耗时长得多。相比之下，我们的方法颠倒了三维重建与新视角合成的先后顺序，使新视角合成过程既快速、又在不确定性下保有多多样性。由于本文处理的是从单张图像到三维物体的问题，推理时若使用真实世界图像，我们会先对每张图像施加一个现成的背景去除工具 [35]，再将其作为 Zero-1-to-3 的输入。

D. 三维重建

与原始的 Score Jacobian Chaining (SJC) 实现不同，我们移除了“空缺损失” (emptiness loss) 和“中心损

失” (center loss)。为正则化 VoxelRF 表示, 我们可微分地渲染一张深度图, 并对其施加平滑损失。这基于一个先验: 物体的几何通常比其纹理包含更少的高频信息。该损失在消除物体表示中的孔洞上尤其有用。我们还施加邻近视角一致性损失, 来约束从某一视角渲染的图像与从邻近随机采样视角渲染的另一图像之间的差异。我们发现这对提升物体纹理的跨视角一致性很有帮助。全部实现细节见作为附录一部分提交的代码。在一块 RTX A6000 GPU 上, 对一张图像完成一次完整三维重建约需 30 分钟。

网络提取。我们按如下方式从 VoxelRF 表示中提取三维网格。首先在分辨率 200^3 上查询密度网格。然后用尺寸为 $(7, 7, 7)$ 的均值滤波器平滑密度网格, 再施加尺寸为 $(5, 5, 5)$ 的腐蚀算子。最后, 对所得密度网格运行 Marching Cubes。设 $\bar{d}$ 为密度网格的平均值。对 GSO 数据集, 我们使用 $8\bar{d}$ 的密度阈值。对 RTMV 数据集, 我们使用 $4\bar{d}$ 的密度阈值。

评估。先把真值三维形状和预测三维形状归一化到单位立方体内。为计算倒角距离 (CD), 我们随机采样 2000 个点。对 Point-E 和 MCC, 直接从它们预测的点云中采样。对我们的方法和 SJC-I, 则从重建出的三维网格上采样点。我们在分辨率 64^3 上计算体素 IoU。对我们的方法、Point-E 和 SJC-I, 用 Marching Cubes 把重建出的三维表面网格体素化。对 MCC, 则直接按占据情况把预测的稠密点云体素化。

E. 基线

为与我们方法的适用范围保持一致, 我们只与满足以下条件的方法比较: (1) 在零样本设定下工作; (2) 以单视图 RGB 图像为输入; (3) 有官方参考实现可在线获取, 且能在合理时间内改造适配。下文各节描述这些基线的实现细节。

E.1. DietNeRF

我们使用 GitHub² 上的官方实现。截至撰稿时, 该代码支持用联合的 MSE 与一致性损失从头做少视角 NeRF 优化, 但未提供与微调 PixelNeRF 相关的功能。为公平起见, 我们采用与 [23] 中在 NeRF 合成数据集

²<https://github.com/ajayjain/DietNeRF>

上所做实验相同的超参数。为评估新视角合成, 我们从测试集中指定的相机位姿渲染所得 NeRF。

E.2. Point-E

我们使用 GitHub³ 上的官方实现和预训练模型。我们保留所有超参数, 并按其演示示例从单张输入图像做三维重建。预测结果已经归一化, 因此无需做任何缩放来对齐真值。表面网格提取采用其默认方法, 网格尺寸为 128。

E.3. MCC

我们使用 GitHub⁴ 上的官方实现。由于该方法要求以带颜色的点云而非 RGB 图像为输入, 我们先施加一个在线现成的前景分割方法 [35], 以及一个最先进的深度估计方法 [42, 41] 做预处理。为公平起见, 我们把所有超参数保持与 [59] 中所述零样本、真实世界实验相同。为评估三维重建, 我们对预测结果归一化、按相机外参旋转, 再与三维真值比较。

³<https://github.com/openai/point-e>

⁴<https://github.com/facebookresearch/MCC>