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

郑德智

2023 年 12 月 10 日

摘要

该论文提出了一个在仅给定单个 RGB 图像的情况下改变物体相机视点的框架。为了在这种不受约束的情境中进行新颖视角的合成，该论文充分利用了大规模扩散模型对自然图像的几何先验知识。他们的条件扩散模型使用合成数据集学习相对相机视点的控制，从而允许在指定的相机变换下生成同一物体的新图像。尽管它是在合成数据集上训练的，但他们的模型保持了强大的零样本泛化能力，适用于超出分布数据集和野外图像，包括印象派绘画。他们的视点条件扩散方法还可用于从单个图像进行 3D 重建的任务。

关键词：扩散模型；单图像物体重建

1 引言

仅通过单一摄像机视角，人类通常能够想象物体的三维形状和外观。这种能力对于日常任务非常重要，比如物体操纵和在复杂环境中导航，同时也是视觉创造力的关键，比如绘画。尽管这种能力可以部分地通过对称性等几何先验来解释，但人类似乎能够推广到更具挑战性的对象，轻松突破物理和几何约束。事实上，人类可以预测那些在物理世界中不存在（甚至不能存在）的对象的三维形状。为了实现这种广泛泛化的程度，人类依赖于通过一生的视觉探索积累的先验知识。相比之下，由于依赖于昂贵的三维注释（例如 CAD 模型）或特定类别的先验条件，大多数现有的三维图像重建方法在封闭世界环境中运作。最近，一些方法通过在大规模、多样化的数据集上进行预训练，如 CO3D，在开放世界三维重建方向上取得了重大进展。然而，这些方法通常仍然需要与训练相关的几何信息，例如立体视图或摄像机姿势。因此，与最近的互联网规模的文本-图像集合 [17] 相比，这些方法使用的数据规模和多样性仍然微不足道，而这些文本-图像集合是大规模扩散模型 [12,15,16] 成功的关键。已经证明，互联网规模的预训练赋予这些模型丰富的语义先验，但它们捕捉几何信息的程度仍然很大程度上未被探索。在本文中，他们展示了他们能够学习控制机制，以操作大规模扩散模型（例如 Stable Diffusion [15]）的摄像机视点，以执行零样本新视角合成和三维形状重建。在给定单个 RGB 图像的情况下，这两个任务都受到严重的约束不足。然而，由于现代生成模型可用的训练数据规模之大（超过 50 亿张图像），扩散模型是自然图像分布的最先进表示，其支持涵盖了大量对象的许多视点。尽管它们是在没有任何摄像机对应关系的 2D 单眼图像上训练的，可以微调模型以学习在生成过程中进行相对摄像机旋转和平移的控制。这些控制允许模型对任意图像进行编码，然后解码为选择的不同摄像机视点。本文的主要贡献在于展示大规

模扩散模型已经学到了关于视觉世界的丰富的三维先验知识，尽管它们只是在二维图像上进行训练。还展示了在新视角合成和零样本三维物体重建方面的最新成果，而且这些成果都是基于单个 RGB 图像的。

2 相关工作

2.1 3D 生成模型

结合大规模图文数据集 [17] 的生成图像架构的最新进展使得合成多样场景和对象的高保真度成为可能 [12,16]。特别是，扩散模型在使用去噪目标的情况下已经被证明在学习可扩展的图像生成器方面非常有效。然而，将它们扩展到 3D 领域将需要大量昂贵的注释 3D 数据。相反，最近的方法依赖于将预训练的大规模 2D 扩散模型转移到 3D，而无需使用任何地面真实的 3D 数据。神经辐射场或 NeRFs [11] 因其能够以高保真度编码场景的能力而成为强大的表示形式。通常，NeRF 用于单场景重建，其中提供了涵盖整个场景的许多姿势图像。然后，任务是从未观察到的角度预测新的视图。DreamField [7] 表明 NeRF 是一个更多才多艺的工具，也可以用作 3D 生成系统的主要组件。各种后续工作 [8,13,18] 用 CLIP 替代了来自 2D 扩散模型的蒸馏损失，这些模型被重新定位为从文本输入生成高保真度的 3D 对象和场景。他们的工作探索了一种非传统的新视角合成方法，将其建模为一种以视点为条件的图像到图像的翻译任务，使用扩散模型。所学模型还可以与 3D 蒸馏结合，从单个图像重构出 3D 形状。之前的工作 [20] 采用了类似的流程，但没有展示零样本泛化的能力。同时进行的方法 [3,9,24] 提出了类似的技术，使用语言引导的先验和文本反演 [5] 执行图像到 3D 生成。相比之下，他们的方法通过合成数据集学习视点的控制，并展示了对野外图像的零样本泛化。

2.2 单视图物体重建

从单一视图重建三维物体是一个极具挑战性的问题，需要强大的先验知识。其中一种方法是通过依赖表示为网格 [21,25]、体素 [6,23] 或点云 [4,10] 的 3D 基元的集合，并使用图像编码器进行条件建模。这些模型受使用的 3D 数据集合的多样性的限制，由于这种类型的条件的全局性质，它们显示出较差的泛化能力。此外，它们需要额外的姿势估计步骤以确保估计的形状与输入的对齐。另一方面，局部条件模型 [19] 旨在直接使用场景重建的局部图像特征，并展现出更强的跨域泛化能力，尽管通常局限于附近的视角重建。最近，MCC [22] 从 RGB-D 视图学习了通用的三维重建表示，并在以物体为中心的视频的大规模数据集上进行训练。在他们的工作中，他们展示了可以直接从预训练的 Stable Diffusion 模型中提取丰富的几何信息，减轻了对额外深度信息的需求。

3 本文方法

3.1 本文方法概述

给定一个单一的 RGB 图像 $x \in \mathbb{R}^{H\times W\times 3}$，表示一个物体，目标是合成该物体的图像，使其具有不同的相机视点。设 $R \in \mathbb{R}^{3\times 3}$ 和 $T \in \mathbb{R}^3$ 分别为所需视点的相对相机旋转和平移。目

标是学习一个模型 f，该模型能够在给定相机变换的情况下合成新的图像：

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

，其中 $\hat{x}R, T$ 表示合成的图像。希望估计得到的 $\hat{x}R, T$ 在感知上与真实但未观察到的新视图 xR, T 相似。从单 ocular RGB 图像进行新视图合成是一个严重的欠约束问题。他们的方法将利用大规模扩散模型，如稳定扩散，以执行此任务，因为它们在从文本描述生成多样图像方面展现出非凡的零样本能力。由于它们的训练数据规模 [17]，预训练的扩散模型是当今自然图像分布的最先进表示。然而，他们必须克服两个挑战来创建 f。首先，尽管大规模生成模型模型是在不同视角下对各种对象进行大量训练的，但这些表示并未明确编码视角之间的对应关系。其次，生成模型继承了在互联网上反映的视角偏见。

网络结构如图 1所示：

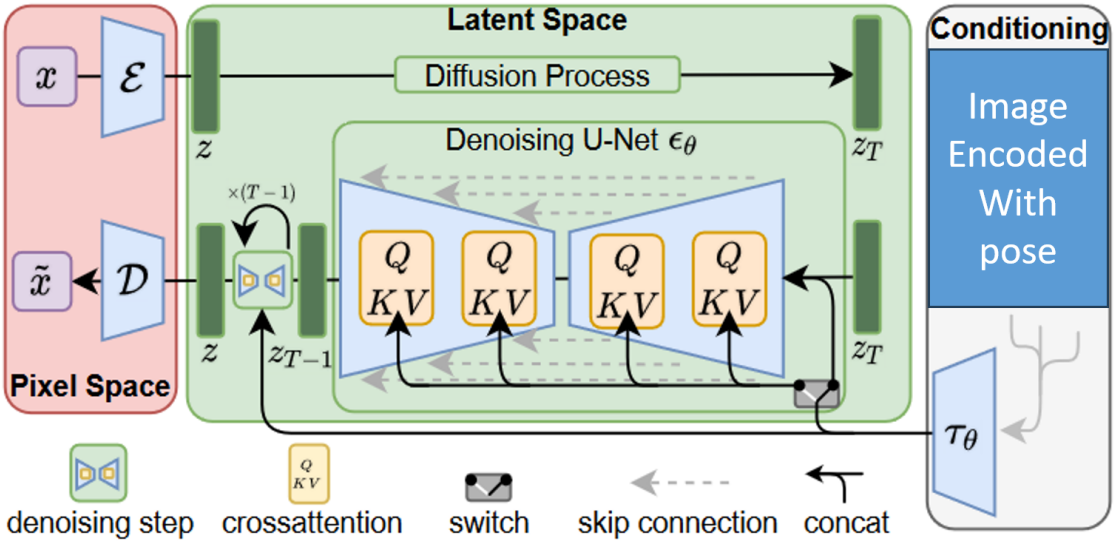

图 1. 方法示意图

3.2 学习控制相机视角

由于扩散模型已经在互联网规模的数据上进行了训练，它们对自然图像分布的支持可能涵盖了大多数对象的大多数视角，但这些视角在预训练模型中无法受控。一旦能够教给模型一种机制来控制拍摄照片的相机外参，那么就能够解锁执行新视图合成的能力。为此，考虑一个配对图像及其相对相机外参的数据集 $(x, x_{R,T}, R, T)$，模型的方法如图 1所示，对预训练的扩散模型进行微调，以学习对相机参数的控制，同时不破坏其余表示。根据 [15]，使用具有编码器 E、去噪 U-Net ϵ_θ 和解码器 D 的潜在扩散架构。在扩散时间步 $t \sim [1, 1000]$，设 $c(x, R, T)$ 为输入视图和相对相机外参的嵌入。然后，求解以下目标以微调模型：

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

这里，z_t 是在扩散时间步 t 的样本，ϵ 是来自均值为 0，方差为 1 的正态分布的噪声。这个目标函数表达了模型在噪声条件下对于输入视图和相机外参的适应性，通过调整模型参数 θ 来实现微调。在模型 θ 训练完成后，推理模型 f 可以通过从一个高斯噪声图像开始进行迭代去

噪 [15]，条件是 $c(x, R, T)$，从而生成图像。本文的主要结果是，通过这种方式微调预训练的扩散模型使其能够学习一种通用的机制，用于控制相机视角，该机制在微调数据集中所见对象之外进行外推。换句话说，这种微调允许控制被“添加”，并且扩散模型仍然保留生成逼真图像的能力，只不过现在可以控制视点。这种组合性确立了模型的零样本能力，最终模型可以为在微调集中缺乏 3D 资源且从未出现在微调集中的对象类别合成新视图。

3.3 以视角为条件的扩散模型

从单一图像进行的 3D 重建需要同时进行低层次感知（深度、阴影、纹理等）和高层次理解（类型、功能、结构等）。因此，采用了一种混合调节机制。在一个流中，将输入图像的 CLIP [14] 嵌入与（R，T）连接以形成“姿势 CLIP”嵌入 $c(x, R, T)$。应用跨注意力机制来调节去噪 U-Net，从而提供输入图像的高级语义信息。在另一个流中，输入图像与正在去噪的图像进行通道连接，帮助模型保留正在合成对象的身份和细节。为了能够应用无分类器引导，在推理期间将输入图像和姿势 CLIP 嵌入设置为零向量，并在缩放条件信息。

3.4 3D 重建

在许多应用中，合成对象的新视图并不足够。希望进行全面的三维重建，捕捉对象的外观和几何特征。采用最近开源的框架 Score Jacobian Chaining（SJC）[18]，通过文本到图像扩散模型的先验进行三维表示的优化。然而，由于扩散模型的概率性质，梯度更新具有很高的随机性。SJC 中采用的一个关键技术，受 DreamFusion [13] 启发，是将无分类器的引导值设置为显著高于通常值。这种方法减少了每个样本的多样性，但提高了重建的保真度。然后，用高斯噪声 $\epsilon \sim$ N（0,1）扰动生成的图像，并通过应用 U-Net $\epsilon\theta$（以输入图像 x、姿态 CLIP 嵌入 c（x，R，T）和时间步 t 为条件）对其进行去噪，以近似得到朝向非噪声输入 x 的分数：

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

其中 LSJC 是由 [18] 引入的 PAAS 分数。此外，通过均方误差损失优化输入视图。为了进一步规范 NeRF 表示，还对每个采样视点应用深度平滑损失，并对近视图一致性损失进行规范，以规范相邻视图之间的外观变化。

3.5 数据集

他们使用最近发布的 Objaverse [2] 数据集进行微调，这是一个大规模的开源数据集，包含由 100,000 多名艺术家创建的 800,000 多个 3D 模型。虽然它没有像 ShapeNet [1] 那样的明确类别标签，但 Objaverse 包含大量高质量的 3D 模型，具有丰富的几何结构，其中许多模型具有细致的细节和材质属性。对于数据集中的每个对象，随机采样 12 个相机外参矩阵 Me，指向对象中心，并使用光线追踪引擎渲染 12 个视图。在训练时，可以为每个对象随机采样两个视图，形成一个图像对 (x, xR,T)。可以轻松从两个外参矩阵中导出定义两个透视图之间映射的相应相对视点变换 (R, T)。

4 复现细节

4.1 与已有开源代码对比

我没有参考已有的代码，凭借着自己的能力在 diffusers 上实现了 Zero123 模型。

4.2 实验环境搭建

在 Ubuntu20.04 上，CUDA11.8，python3.8 上实现。环境搭建只需要安装相应的 python 依赖

4.3 创新点

官方使用 openai 的框架实现，而我使用 huggingface 的 diffuers 框架实现。diffusers 使用简单，并且集成了多种优化加速模块。可以加快推理速度，优化显存占用。并且在原模型权重的基础上，并且在 ObjaverseXL 数据集上，继续训练，取得了比论文中原模型更好的结果。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。表格 1 展示了在 ObjaverseXL [2] 上继续训练的结果，评测指标用 PSNR,SSIM，LPIPS,FID

Zero123	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)
Paper	18.378	0.877	0.088	0.027
复现结果	**18.545**	**0.880**	**0.084**	**0.024**

表 1. 在 ObjaverseXL 数据集上，继续训练的结果

等，这些指标在原文中均有使用，我遵循原文的设置进行测试。可以明显地看出，在所有指标上，继续训练的模型都比原文的模型要好。从表 1 中的各项量化指标中，可以看到训练是十分成功的。

	latency	Speed-up
original	4.34s	x1
fp16	1.65s	x2.63
channels last	1.51s	x2.88
traced UNet	1.47s	x2.96
memory efficient attention	1.20s	x3.61

表 2. 在 diffuesrs 框架下，推理速度与原框架速度对比

另外，由于模型使用 diffusers 框架实现的，diffusers 框架下内置了许多加速模块，可以轻松的使用。表 2 是我在 diffusers 框架下用不同加速模块与原官方实现的对比结果。可以看

见几乎都达到了一倍以上的提升，另外，显存占用从原本的 11G 下降到了 5G。在显存和推理速度上对比原文都有非常大的提升。

图 2. 可视化结果

6 总结与展望

本部分对整个文档的内容进行归纳并分析目前实现过程中的不足以及未来可进一步进行研究的方向。

参考文献

[1] 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.

[2] 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.

[3] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023.

[4] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

[5] 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. *arXiv preprint arXiv:2208.01618*, 2022.

[6] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 484–499. Springer, 2016.

[7] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 857–866, 2022.

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

[9] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023.

[10] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

[11] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[12] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[13] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[14] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

[16] 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, 2022. *URL https://arxiv. org/abs/2205.11487*, 4.

[17] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[18] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023.

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.

[20] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

[21] Markus Worchel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, and Peter Eisert. Multi-view mesh reconstruction with neural deferred shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6187–6197, 2022.

[22] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9065–9075, 2023.

[23] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in neural information processing systems*, 30, 2017.

[24] 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 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4479–4489, 2023.

[25] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019.