ZeroNVS：从单张图像实现零样本 360 度视角合成

ZeroNVS: Zero-Shot 360-Degree View Synthesis from a Single Image

Kyle Sargent[1], Zizhang Li[1], Tanmay Shah[2], Charles Herrmann[2], Hong-Xing Yu[1], Yunzhi Zhang[1], Eric Ryan Chan[1], Dmitry Lagun[2], Li Fei-Fei[1], Deqing Sun[2], Jiajun Wu[1]

[1]Stanford University, [2]Google Research

摘要

We introduce a 3D-aware diffusion model, ZeroNVS, for single-image novel view synthesis for in-the-wild scenes. While existing methods are designed for single objects with masked backgrounds, we propose new techniques to address challenges introduced by in-the-wild multi-object scenes with complex backgrounds. Specifically, we train a generative prior on a mixture of data sources that capture object-centric, indoor, and outdoor scenes. To address issues from data mixture such as depth-scale ambiguity, we propose a novel camera conditioning parameterization and normalization scheme. Further, we observe that Score Distillation Sampling (SDS) tends to truncate the distribution of complex backgrounds during distillation of 360-degree scenes, and propose "SDS anchoring" to improve the diversity of synthesized novel views. Our model sets a new state-of-the-art result in LPIPS on the DTU dataset in the zero-shot setting, even outperforming methods specifically trained on DTU. We further adapt the challenging Mip-NeRF 360 dataset as a new benchmark for single-image novel view synthesis, and demonstrate strong performance in this setting. Code and models are available at this url.

我们提出一种具备 *3D* 感知能力的扩散模型 *ZeroNVS*，用于真实场景（*in-the-wild*）下的单图像新视角合成。现有方法大多针对背景经掩码处理的单个物体而设计，真实场景中的多物体与复杂背景则带来了新的难题，我们为此提出若干新技术加以应对。具体而言，我们在一个混合数据源上训练生成先验，这些数据涵盖以物体为中心、室内以及室外的各类场景。数据混合会引入深度-尺度歧义等问题，我们为此设计了一种新的相机条件参数化与归一化方案。此外，我们观察到在蒸馏 *360* 度场景时，分数蒸馏采样（*Score Distillation Sampling*，*SDS*）往往会截断复杂背景的分布，于是提出"*SDS* 锚定"来提升合成新视角的多样性。在零样本设定下，我们的模型在 *DTU* 数据集上取得了 *LPIPS* 指标的最新最优结果，甚至超过了专门在 *DTU* 上训练的方法。我们进一步将颇具挑战的 *Mip-NeRF 360* 数据集改造为单图像新视角合成的新基准，并在该设定下展现出优异的性能。代码与模型见此网址。

1. 引言（Introduction）

Models for single-image, 360-degree novel view synthesis (NVS) should produce *realistic* and *diverse* results: the synthesized images should look natural and 3D-consistent to humans, and they should also capture the many possible explanations of unobservable regions. This challenging problem has typically been studied in the context of single objects without backgrounds, where the requirements on both realism and diversity are simplified. Recent progress relies on large 3D datasets like Objaverse-XL [9] which have enabled training conditional diffusion [19] models to perform photorealistic and 3D-consistent NVS via Score Dis-

本文为 AI 辅助翻译版本，仅供学习交流；引用请以英文原文为准：arXiv:2310.17994。

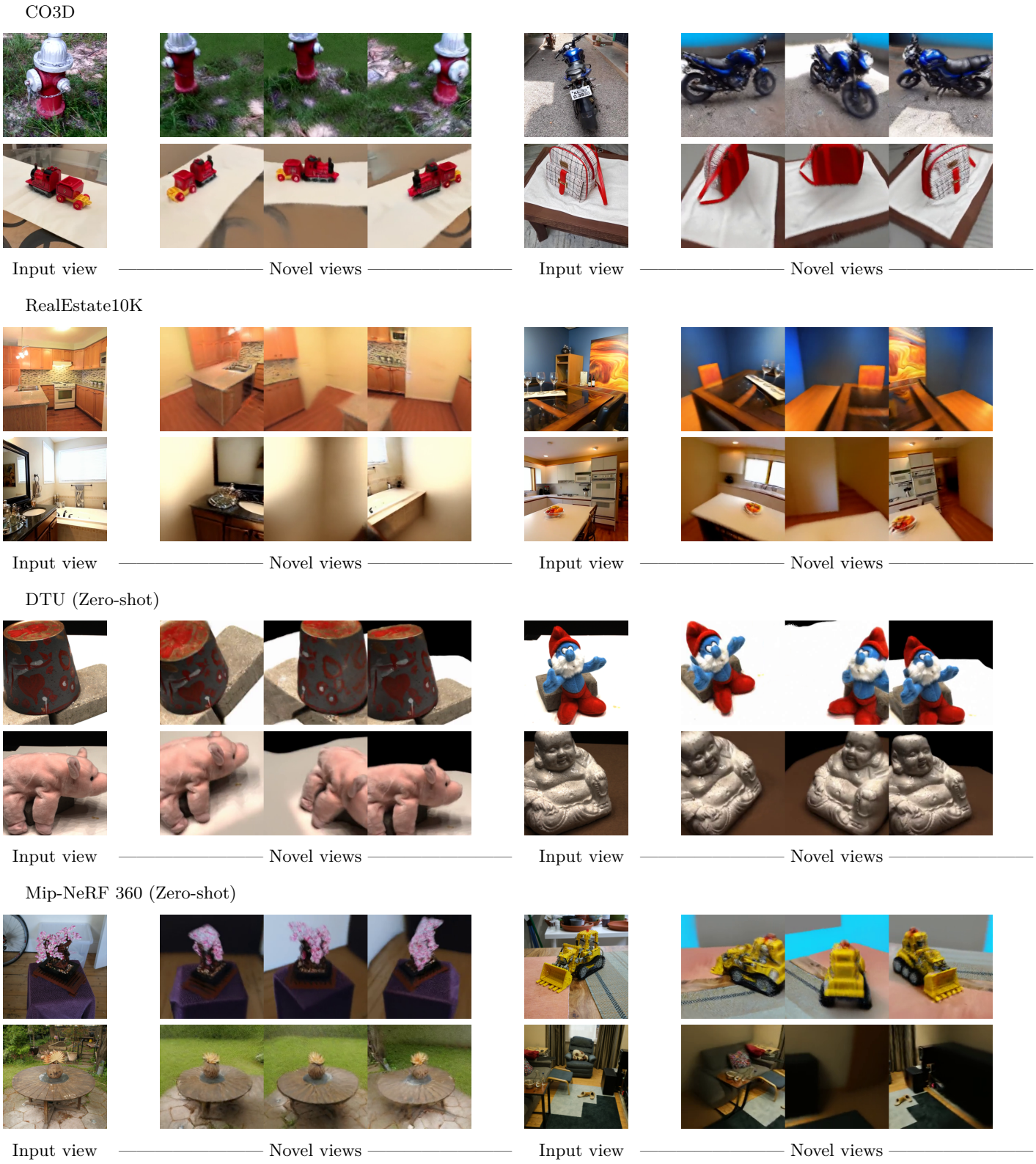

图 1. Results for view synthesis from a single image. All NeRFs are predicted by the *same* model. 单图像视角合成结果展示。所有 NeRF 由同一模型预测。

tillation Sampling [SDS; 27]. Meanwhile, since image diversity mostly lies in the background, not the object, the ignorance of background significantly lowers the expectation of synthesizing diverse images–in fact, most object-centric methods do not consider diversity metrics [19, 22, 28].

单图像 360 度新视角合成（NVS）模型应生成逼真且多样的结果。合成图像应在视觉上显得自然且具有 3D 一致性，同时应捕捉对不可观测区域的多种合理解释。这一问题通常在不含背景的单个物体背景下研究，相应地放宽了真实感和多样性的要求。最近的进展得益于 Objaverse-XL [9] 等大规模三维数据集，使研究者能够训练条件扩散模型 [19]，通过分数蒸馏采样（SDS）[SDS; 27] 生成逼真且 3D 一致的新视角合成。然而，由于图像多样性主要取决于背景而非物体本身，忽视背景大幅降低了生成多样图像的期望。事实上，大多数以物体为中心的方法都未考虑多样性指标 [19, 22, 28]。

Neither assumption holds for the more challenging problem of zero-shot, 360-degree novel view synthesis on real-world scenes. There is no single, large-scale dataset of scenes with ground-truth geometry, texture, and camera parameters, analogous to Objaverse-XL for objects. The background, which cannot be ignored anymore, also needs to be well modeled for synthesizing diverse results.

对于更具挑战性的零样本 360 度真实场景新视角合成问题，上述两个假设均不成立。不存在单个大规模的场景数据集，具有真值几何、纹理和相机参数，无法与物体领域的 Objaverse-XL 相对应。背景不能再被忽视，也需要精心建模以生成多样化的结果。

We address both issues with our new model, ZeroNVS. Inspired by previous object-centric methods [19, 22, 28], ZeroNVS also trains a 2D conditional diffusion model followed by 3D distillation. But unlike them, ZeroNVS works well on scenes due to two technical innovations: a new camera parametrization and normalization scheme for conditioning, which allows training the diffusion model on diverse scene datasets, and an "SDS anchoring" mechanism, improving the background diversity over standard SDS.

我们通过新模型 ZeroNVS 来解决这两个问题。受先前以物体为中心的方法 [19, 22, 28] 启发，ZeroNVS 也训练一个二维条件扩散模型，然后进行 3D 蒸馏。但与它们不同的是，ZeroNVS 在场景上表现出色，得益于两项技术创新：用于条件化的新相机参数化和归一化方案，允许在多样场景数据集上训练扩散模型；以及"SDS 锚定" 机制，相比标准 SDS 改进背景多样性。

To overcome the key challenge of limited training data, we propose training the diffusion model on a massive mixed dataset comprised of all scenes from CO3D [31], RealEstate10K [45], and ACID [17], so that the model may potentially handle complex in-the-wild scenes. The mixed data of such scale and diversity are captured with a variety of camera settings and have several different types of 3D ground truth, e.g., computed with COLMAP [33] or ORB-SLAM [24]. We show that while the camera conditioning representations from prior methods [19] are too ambiguous or inexpressive to model in-the-wild scenes, our new camera parametrization and normalization scheme allows exploiting such diverse data sources and leads to superior NVS on real-world scenes.

为克服有限训练数据的关键挑战，我们提出在包含来自 CO3D [31]、RealEstate10K [45] 和 ACID [17] 的全部场景的大规模混合数据集上训练扩散模型，使模型能处理复杂的真实场景。这类大规模且多样的混合数据采用多种相机设置获取，具有不同类型的三维真值，如用 COLMAP [33] 或 ORB-SLAM [24] 计算的。我们发现，先前方法 [19] 的相机条件表示对于真实场景而言过于模糊或缺乏表达力，而我们的新相机参数化和归一化方案能够有效利用这些多样数据源，在真实世界场景上实现更优的新视角合成。

Building a 2D conditional diffusion model that works effectively for in-the-wild scenes enables us to then study the limitations of SDS in the scene setting. In particular, we observe limited diversity from SDS in the generated scene backgrounds when synthesizing long-range (e.g., 180-degree) novel views. We therefore propose "SDS anchoring" to ameliorate the issue. In SDS anchoring, we propose to first sample several "anchor" novel views using the standard Denoising Diffusion Implicit Model (DDIM) sampling [37]. This yields a collection of pseudo-ground-truth novel views with diverse contents, since DDIM is not prone to mode collapse like SDS. Then, rather than using these views as

RGB supervision, we sample from them randomly as conditions for SDS, which enforces diversity while still ensuring 3D-consistent view synthesis.

构建对真实场景有效的二维条件扩散模型，使我们能进一步研究 SDS 在场景设置中的局限。特别地，在合成长距离（如 180 度）新视角时，我们观察到 SDS 对生成的场景背景多样性不足。因此我们提出"SDS 锚定" 来缓解该问题。在 SDS 锚定中，我们首先使用标准去噪扩散隐式模型（DDIM）采样 [37] 采样多个" 锚点" 新视角。这产生一组具有多样内容的伪真值新视角，因为 DDIM 不像 SDS 那样易陷入模式坍缩。随后，我们从这些视角中随机采样作为 SDS 的条件，而非用作 RGB 监督，从而在保证 3D 一致视角合成的同时强制多样性。

ZeroNVS achieves strong zero-shot generalization to unseen data. We set a new state-of-the-art LPIPS score on the challenging DTU benchmark, even outperforming methods that were directly fine-tuned on this dataset. Since the popular benchmark DTU consists of scenes captured by a forward-facing camera rig and cannot evaluate more challenging pose changes, we propose to use the Mip-NeRF 360 dataset [2] as a single-image novel view synthesis benchmark. ZeroNVS achieves the best LPIPS performance on this benchmark. Finally, we show the potential of SDS anchoring for addressing diversity issues in background generation via a user study.

ZeroNVS 在未见数据上达到强大的零样本泛化性能。我们在具有挑战性的 DTU 基准上创设新的最先进 LPIPS 得分，甚至超过在该数据集上直接微调的方法。由于广泛使用的 DTU 基准由前向相机装置捕获的场景组成，无法评估更复杂的位姿变化，我们建议将 Mip-NeRF 360 数据集 [2] 作为单图像新视角合成基准。ZeroNVS 在此基准上达到最佳 LPIPS 性能。最后，我们通过用户研究展示了 SDS 锚定在解决背景生成多样性问题中的潜力。

To summarize, we make the following contributions:

总结起来，我们做出如下贡献：

- We propose ZeroNVS, which enables full-scene NVS from real images. ZeroNVS first demonstrates that SDS distillation can be used to lift scenes that are not object-centric and may have complex backgrounds to 3D.

  我们提出 ZeroNVS，让完整场景的新视角合成成为可能。ZeroNVS 首次展示 SDS 蒸馏可用于将非以物体为中心、包含复杂背景的场景提升至三维。
- We show that the formulations on handling cameras and scene scale in prior work are either inexpressive or ambiguous for in-the-wild scenes. We propose a new camera conditioning parameterization and a scene normalization scheme. These enable us to train a single model on a large collection of diverse training data consisting of CO3D, RealEstate10K and ACID, allowing strong zero-shot generalization for NVS on in-the-wild images.

  我们指出先前工作中处理相机和场景尺度的公式对于真实场景而言缺乏表达力或存在歧义。我们提出新的相机条件参数化和场景归一化方案。这使我们能在包含 CO3D、RealEstate10K 和 ACID 的大量多样训练数据上训练单一模型，在真实图像上实现强大的零样本泛化。
- We study the limitations of SDS distillation as applied to scenes. Similar to prior work, we identify a diversity issue, which manifests in this case as novel view predictions with monotone backgrounds. We propose SDS anchoring to ameliorate the issue.

  我们研究了 SDS 蒸馏应用于场景时的局限。类似先前工作，我们识别出多样性问题，表现为新视角预测背景单调。我们提出 SDS 锚定来缓解该问题。
- We show state-of-the-art LPIPS results on DTU *zero-shot*, surpassing prior methods finetuned on this dataset. Furthermore, we introduce the Mip-NeRF 360 dataset as a scene-level single-image novel view synthesis benchmark and analyze the performances of our and other methods. Finally, we show that our proposed SDS anchoring is preferred via a user study.

  我们在 DTU 上取得最先进的 LPIPS 结果（零样本设置），超越在该数据集上微调的其他方法。此外，我们引入 Mip-NeRF 360 数据集作为场景级单图像新视角合成基准，并分析我们和其他方法的性能。最后，我们通过用户研究证明用户倾向于 SDS 锚定。

2. 相关工作（Related Work）

3D generation. DreamFusion [27] proposed Score Distillation Sampling (SDS) as a way of leveraging a diffusion model to extract a NeRF given a user-provided text prompt. After DreamFusion, follow-up works such as Magic3D [16], ATT3D [21], ProlificDreamer [39], and Fantasia3D [7] improved the quality, diversity, resolution, or run-time. Other types of 3D generative models include GAN-based 3D generative models, which are primarily restricted to single object categories [4, 5, 13, 25, 26, 34] or to synthetic data [12]. Recently, 3DGP [35] adapted the GAN approach to train 3D generative models on ImageNet. VQ3D [32] and IVID [41] leveraged vector quantization and diffusion, respectively, to learn generative models on ImageNet. One critical critical challenge for scene-based 3D-aware methods 360-degree viewpoint change. Both VQ3D and 3DGP demonstrate only limited camera motion, while IVID generally focuses on small camera motion but can achieve 360-degree views for a small subset of scenes.

3D 生成。 DreamFusion [27] 提出了分数蒸馏采样（Score Distillation Sampling, SDS），借助扩散模型从用户给定的文本提示中提取出一个 NeRF。在 DreamFusion 之后，Magic3D [16]、ATT3D [21]、ProlificDreamer [39]、Fantasia3D [7] 等后续工作分别在质量、多样性、分辨率或运行速度上有所改进。另一类 3D 生成模型是基于 GAN 的 3D 生成模型，它们大多局限于单一物体类别 [4, 5, 13, 25, 26, 34] 或合成数据 [12]。近期，3DGP [35] 将 GAN 方法用于在 ImageNet 上训练 3D 生成模型。VQ3D [32] 与 IVID [41] 则分别借助向量量化与扩散，在 ImageNet 上学习生成模型。对于基于场景的 3D 感知方法而言，一个关键挑战是 360 度视角变化。VQ3D 与 3DGP 都只能实现有限的相机运动；IVID 一般也聚焦于小幅相机运动，仅在少数场景上能达到 360 度视角。

Single-image novel view synthesis. PixelNeRF [44] and DietNeRF [15] learn to infer NeRFs from sparse views via training an image-based 3D feature extractor or semantic consistency losses, respectively. However, these approaches do not produce renderings resembling crisp natural images from a single image. Several recent diffusion-based approaches achieve high quality results by separating the problem into two stages. First, a (potentially 3D-aware) diffusion model is trained, and second, the diffusion model is used to distill 3D-consistent scene representations given an input image via techniques like score distillation sampling [27], score Jacobian chaining [38], textual inversion or semantic guidance leveraging the diffusion model [10, 22], or explicit 3D reconstruction from multiple sampled views of the diffusion model [18, 20]. Another diffusion-based work, GeNVS [6], achieves 360 camera motion but only for specific object categories such as fire hydrants. By contrast, ZeroNVS generates 360-degree camera motion by default for a variety of scene categories. This is enabled by innovations such as novel camera conditioning representations and SDS anchoring, which enable us to train on massive real scene datasets and then to perform scene-level NVS with up to 360-degree viewpoint change on diverse scene types.

单图像新视角合成。 PixelNeRF [44] 与 DietNeRF [15] 分别通过训练基于图像的 3D 特征提取器、以及语义一致性损失，学习从稀疏视角推断 NeRF。但这些方法无法从单张图像生成媲美清晰自然图像的渲染结果。近期若干基于扩散的方法把问题拆成两个阶段，从而取得高质量的结果：第一阶段训练一个（可能具备 3D 感知能力的）扩散模型；第二阶段则利用该扩散模型，从输入图像蒸馏出 3D 一致的场景表示，所用技术包括分数蒸馏采样 [27]、分数雅可比链式法 [38]、借助扩散模型的文本反演或语义引导 [10, 22]，以及从扩散模型采样的多个视角做显式 3D 重建 [18, 20]。另一项基于扩散的工作 GeNVS [6] 实现了 360 度相机运动，但仅限于消防栓等特定物体类别。相比之下，ZeroNVS 默认就能为各类场景生成 360 度的相机运动。这得益于新的相机条件表示、SDS 锚定等创新，使我们能够在海量真实场景数据集上训练，进而在多样的场景类型上实现视角变化最高达 360 度的场景级 NVS。

3. 方法（Approach）

We consider the problem of scene-level novel view synthesis from a single real image. Similar to prior work [19, 28], we first train a diffusion model $\mathbf{p}_\theta$ to perform novel view syn-

thesis, and then leverage it to perform 3D SDS distillation. Unlike prior work, we focus on scenes rather than objects.

我们考虑从单张真实图像做场景级新视角合成的问题。类似于既有工作 [19, 28]，我们首先训练扩散模型 $\mathbf{p}_\theta$ 以执行新视角合成，然后利用它执行 3D SDS 蒸馏。与既有工作不同的是，我们专注于场景而非单个物体。

Scenes present several unique challenges. First, prior works use representations for cameras and scale which are either ambiguous or insufficiently expressive for scenes. Second, the inference procedure of prior works is based on SDS, which has a known mode collapse issue and which manifests in scenes through greatly reduced background diversity in predicted views. We will attempt to address these challenges through improved representations and inference procedures for scenes compared with prior work [19, 28].

场景呈现出多项独有的挑战。首先，既有工作采用的相机和尺度表示对于场景而言要么具有歧义性，要么表达力不足。其次，既有工作的推理过程基于 SDS，这存在已知的模式坍缩问题，在场景中表现为预测视角的背景多样性大幅降低。我们将通过与既有工作 [19, 28] 相比改进的表示和推理过程来试图解决这些挑战。

We shall begin by introducing some general notation. Let a scene S consist of a set of images $X = \{X_i\}_{i=1}^n$, depth maps $D = \{D_i\}_{i=1}^n$, extrinsics $E = \{E_i\}_{i=1}^n$, and a shared field-of-view f. We note that an extrinsics matrix E_i can be identified with its rotation and translation components, defined by $E_i = (E_i^R, E_i^T)$. We preprocess our data to consist of square images and assume intrinsics are shared within a given scene, and that there is no skew, distortion, or off-center principal point.

我们首先介绍一些通用记号。设场景 S 由一组图像 $X = \{X_i\}_{i=1}^n$、深度图 $D = \{D_i\}_{i=1}^n$、外参 $E = \{E_i\}_{i=1}^n$ 和共享的视场角 f 组成。注意外参矩阵 E_i 可以用其旋转和平移分量表示，定义为 $E_i = (E_i^R, E_i^T)$。我们对数据预处理，使其由正方形图像组成，并假设内参在给定的场景内共享，且不存在倾斜、畸变或非中心主点。

We will focus on the design of the conditional information which is passed to the view synthesis diffusion model $\mathbf{p}_\theta$ in addition to the input image. This conditional information can be represented via a function, $\mathbf{M}(D, f, E, i, j)$, which computes a conditioning embedding given the depth maps and extrinsics for the scene, the field of view, and the indices i, j of the input and target view respectively. We learn a generative model over novel views $\mathbf{p}_\theta$ given by

我们重点关注传递给视角合成扩散模型 $\mathbf{p}_\theta$ 的条件信息的设计，该条件信息是对输入图像的补充。这个条件信息可以通过函数 $\mathbf{M}(D, f, E, i, j)$ 表示，该函数根据场景的深度图、外参、视场角以及输入和目标视角的索引 i, j，计算条件嵌入。我们学习新视角上的生成模型 $\mathbf{p}_\theta$，给定为

$$X_j \sim \mathbf{p}_\theta(X_j | X_i, \mathbf{M}(D, f, E, i, j)) \ .$$

The output of $\mathbf{M}$ and the input image X_i are the only information used by the model for NVS. Both Zero-1-to-3 (Section 3.1) and our model, as well as several intermediate models that we will study (Sections 3.2 and 3.3), can be regarded as different choices for $\mathbf{M}$. As we illustrate in Figures 2, 3, 5 and 6, and verify in experiments, different choices for $\mathbf{M}$ can have drastic impacts on performance.

$\mathbf{M}$ 的输出和输入图像 X_i 是模型用于 NVS 的唯一信息。Zero-1-to-3（第 3.1 节）和我们的模型，以及我们将研究的几个中间模型（第 3.2 和 3.3 节），都可以看作对 $\mathbf{M}$ 的不同选择。如我们在图 2、3、5 和 6 中所示，并在实验中验证的那样，对 $\mathbf{M}$ 的不同选择可能对性能产生巨大影响。

3.1. 为视角合成表示物体（Representing Objects for View Synthesis）

Zero-1-to-3 [19] represents poses with 3 degrees of freedom, given by an elevation θ, azimuth ϕ, and radius z. Let $\mathbf{P} : \mathrm{SE}(3) \to \mathbb{R}^3$ be the projection to (θ, ϕ, z), then

Zero-1-to-3 [19] 用 3 个自由度表示位姿，由仰角 θ、方位角 ϕ 和半径 z 给出。设 $\mathbf{P} : \mathrm{SE}(3) \to \mathbb{R}^3$ 为投影到 (θ, ϕ, z) 的函数，则

$$\mathbf{M}_{\mathrm{Zero-1-to-3}}(D, f, E, i, j) = \mathbf{P}(E_i) - \mathbf{P}(E_j)$$

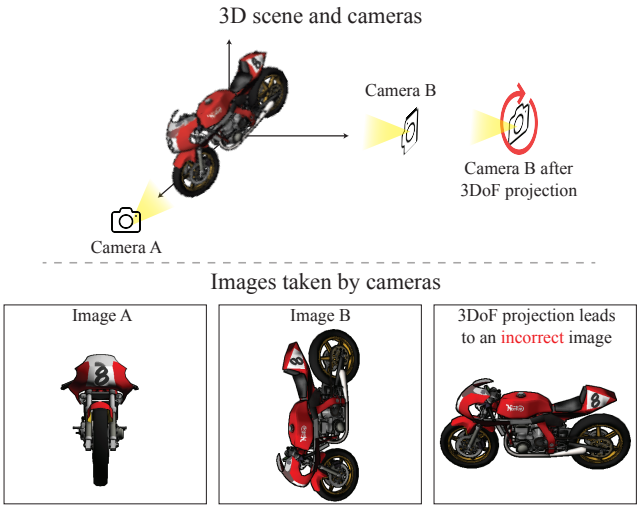

图 2. 3 自由度相机位姿捕捉仰角、方位角和半径，但无法表示相机的滚转（如图所示）或任意位置和方向的相机。

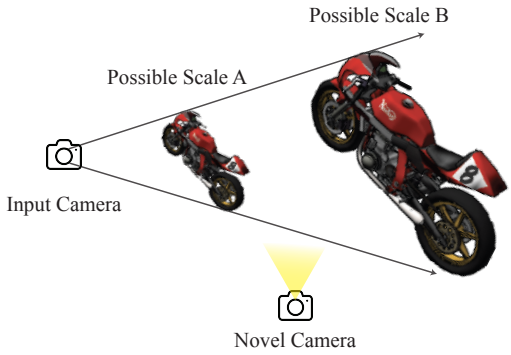

图 3. 对于单目相机，靠近相机的小物体（左）和远处的大物体（右）看起来相同，尽管代表不同的场景。输入视角中的尺度歧义导致多个可信的新视角。

is the camera conditioning representation used by Zero-1-to-3. For object mesh datasets such as Objaverse [8] and Objaverse-XL [9], this representation is appropriate because the data is known to consist of single objects without backgrounds, aligned and centered at the origin and imaged from training cameras generated with three degrees of freedom. However, such a parameterization limits the model's ability to generalize to non-object-centric images, and to real-world data.

是 Zero-1-to-3 使用的相机条件表示。对于 Objaverse [8] 和 Objaverse-XL [9] 等物体网格数据集，这种表示是合适的，因为数据已知由无背景的单个物体组成，对齐并以原点为中心，从用三个自由度生成的训练相机拍摄。然而，这样的参数化限制了模型对非物体中心图像和真实世界数据的泛化能力。

In real-world data, poses can have six degrees of freedom, incorporating both rotation (pitch, roll, yaw) and 3D translation. An illustration of a failure of the 3DoF camera representation is shown in Figure 2.

在真实世界数据中，位姿可以有六个自由度，包括旋转（俯仰、滚转、偏航）和 3D 平移。3 自由度相机表示失败的示意图如图 2 所示。

3.2. 为视角合成表示场景（Representing Scenes for View Synthesis）

For scenes, we should use a camera representation with six degrees of freedom that can capture all possible positions and orientations. One straightforward choice is the relative pose parameterization [40]. We propose to also include the field of view as an additional degree of freedom. We term this combined representation "6DoF+1". This gives us

对于场景，我们应该使用具有六个自由度的相机表示，能够捕捉所有可能的位置和方向。一个直接的选择是相对位姿参数化 [40]。我们建议将视场角作为额外的自由度。我们将这种组合表示称为"6DoF+1"。这给出

$$\mathbf{M}_{\text{6DoF+1}}(D, f, E, i, j) = [E_i^{-1} E_j, f].$$

Importantly, $\mathbf{M}_{\text{6DoF+1}}$ is invariant to any rigid transformation $\tilde{E}$ of the scene, so that we have

重要的是，$\mathbf{M}_{\text{6DoF+1}}$ 对场景的任何刚体变换 $\tilde{E}$ 都是不变的，因此我们有

$$\mathbf{M}_{\text{6DoF+1}}(D, f, \tilde{E} \cdot E, i, j) = [E_i^{-1} E_j, f] \ .$$

This is useful given the arbitrary nature of the poses for our datasets which are determined by COLMAP or ORB-SLAM. The poses discovered via these algorithms are not related to any meaningful alignment of the scene, such

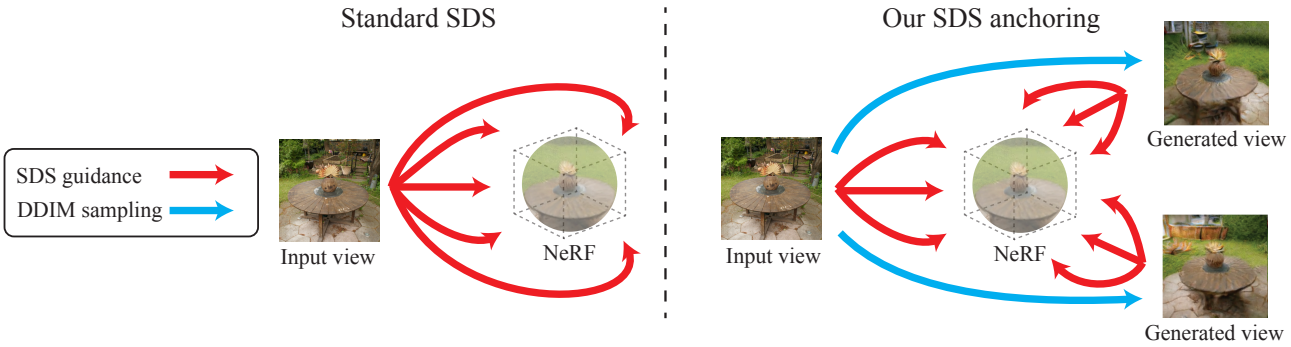

图 4. 基于 SDS 的 NeRF 蒸馏（左）对所有 360 度新视角使用相同的引导图像。我们的"SDS 锚定"（右）首先通过 DDIM [36] 采样新视角，然后使用最近的图像（输入或采样的新视角）作引导。

as a rigid transformation and scale transformation which align the scene to some canonical frame and unit of scale. Although we have seen that $\mathbf{M}_{6\text{DoF}+1}$ is invariant to rigid transformations of the scene, it is not invariant to scale. The scene scales determined by COLMAP and ORB-SLAM are also arbitrary and may vary significantly. One solution is to directly normalize the camera locations. Let $\mathbf{R}(E, \lambda) : \text{SE}(3) \times \mathbb{R} \to \text{SE}(3)$ be a function that scales the translation component of E by λ. Then we define

这很有用，因为我们数据集中由 COLMAP 或 ORB-SLAM 确定的位姿具有任意性。这些算法发现的位姿与任何有意义的场景对齐无关，如刚体变换和尺度变换等用于将场景对齐到某个规范坐标系和单位尺度的变换。虽然我们已经看到 $\mathbf{M}_{6\text{DoF}+1}$ 对场景的刚体变换是不变的，但它对尺度并不是不变的。由 COLMAP 和 ORB-SLAM 确定的场景尺度也是任意的，可能差异很大。一个解决方案是直接归一化相机位置。设 $\mathbf{R}(E, \lambda) : \text{SE}(3) \times \mathbb{R} \to \text{SE}(3)$ 是一个按 λ 缩放 E 的平移分量的函数。然后我们定义

$$s = \frac{1}{n}\sum_{i=1}^{n} \|E_i^T - \frac{1}{n}\sum_{j=1}^{n} E_j^T\|_2 \ ,$$

$$\mathbf{M}_{6\text{DoF}+1,\ \text{norm.}}(D, f, E, i, j) = \left[\mathbf{R}\left(E_i^{-1}E_j, \frac{1}{s}\right), f\right] \ ,$$

where s is the average norm of the camera locations when the mean of the camera locations is chosen as the origin. In $\mathbf{M}_{6\text{DoF}+1,\ \text{norm.}}$, the camera locations are normalized via rescaling by $\frac{1}{s}$, in contrast to $\mathbf{M}_{6\text{DoF}+1}$ where the scales are arbitrary. This choice of $\mathbf{M}$ assures that scenes from our mixture of datasets will have similar scales.

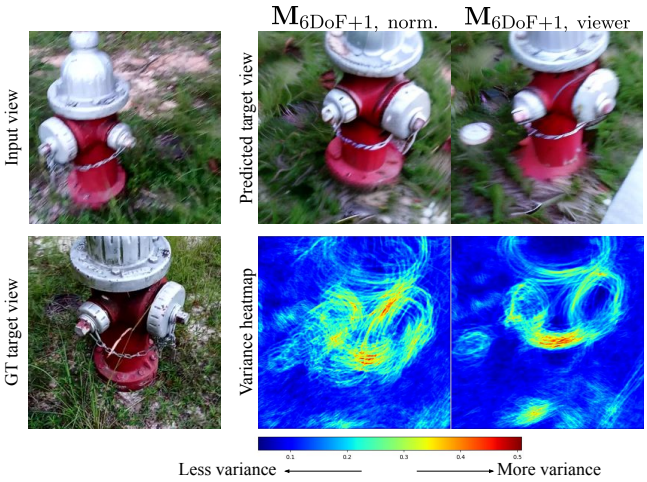

图 5. ZeroNVS 多个样本的 Sobel 边缘样本和方差热图。$\mathbf{M}_{6\text{DoF}+1,\ \text{viewer}}$ 减少了尺度歧义的随机性。

其中 s 是当相机位置的均值作为原点时相机位置的平均范数。在 $\mathbf{M}_{6\text{DoF}+1,\ \text{norm.}}$ 中，相机位置通过按 $\frac{1}{s}$ 重新缩放归一化，与 $\mathbf{M}_{6\text{DoF}+1}$ 中尺度任意的情况相反。这种 $\mathbf{M}$ 的选择确保了我们混合数据集中的场景具有相似的尺度。

3.3. 用新的归一化方案解决尺度歧义 （Addressing Scale Ambiguity with a New Normalization Scheme）

The representation $\mathbf{M}_{6\text{DoF}+1,\ \text{norm.}}$ achieves reasonable performance on real scenes by addressing issues in prior representations with limited degrees of freedom and handling of scale. However, a normalization scheme that better addresses scale ambiguity may lead to improved performance. Scene scale is ambiguous given a monocular input image [30, 43]. This complicates NVS, as we illustrate in Figure 3.

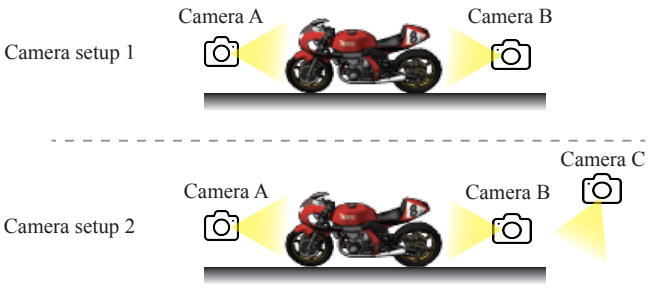

图 6. 上：两个相机面向物体的场景。下：添加一个面向地面的新相机的相同场景。在 $\mathbf{M}_{\text{6DoF+1, agg.}}$ 下添加相机 C 会大幅改变场景的尺度。$\mathbf{M}_{\text{6DoF+1, viewer}}$ 避免了这种情况。

表示 $\mathbf{M}_{\text{6DoF+1, norm.}}$ 通过解决先前表示中自由度有限和尺度处理的问题，在真实场景上实现了合理的性能。然而，更好地解决尺度歧义的归一化方案可能会导致性能提升。给定单目输入图像，场景尺度是有歧义的 [30, 43]。这使 NVS 复杂化，如我们在图 3 中所示。

We therefore choose to introduce information about the scale of the visible content to our conditioning embedding function $\mathbf{M}$. Rather than normalize by camera locations, Stereo Magnification [45] takes the 5-th quantile of each depth map of the scene, and then takes the 10-th quantile of this aggregated set of numbers, and declares this as the scene scale. Let $\mathbf{Q}_k$ be a function which takes the k-th quantile of a set of numbers, then we define

因此，我们选择向条件嵌入函数 $\mathbf{M}$ 引入可见内容尺度的信息。Stereo Magnification [45] 不是按相机位置归一化，而是取场景每个深度图的第 5 分位数，然后取这个聚合数集的第 10 分位数，并将其声明为场景尺度。设 $\mathbf{Q}_k$ 是一个取数集的 k 分位数的函数，则我们定义

$$q = \mathbf{Q}_{10}(\{\mathbf{Q}_5(D_i)\}_{i=1}^{n}) \ ,$$
$$\mathbf{M}_{\text{6DoF+1, agg.}}(D, f, E, i, j) = \left[\mathbf{R}\left(E_i^{-1}E_j, \frac{1}{q}\right), f\right] \ ,$$

where in $\mathbf{M}_{\text{6DoF+1, agg.}}$, q is the scale applied to the translation component of the scene's cameras before computing the relative pose. In this way $\mathbf{M}_{\text{6DoF+1, agg.}}$ is different from $\mathbf{M}_{\text{6DoF+1, norm.}}$ because the camera conditioning representation contains information about the scale of the visible content from the depth maps D_i. Although conditioning with $\mathbf{M}_{\text{6DoF+1, agg.}}$ improves performance, there are two issues. The first arises from aggregating the quantiles over all the images. In Figure 6, adding an additional Camera C to the scene changes the value of $\mathbf{M}_{\text{6DoF+1, agg.}}$ despite nothing else having changed about the scene. This makes the view synthesis task from either Camera A or Camera B more ambiguous. To ensure this is impossible, we can simply eliminate the aggregation step over the quantiles of all depth maps in the scene. The second issue arises from different depth statistics within the mixture of datasets we use for training. ORB-SLAM generally produces sparser depth maps than COLMAP, and therefore the value of $\mathbf{Q}_k$ may have different meanings for each. We therefore use an off-the-shelf depth estimator [29] to fill holes in the depth maps. We denote the depth D_i infilled in this way as $\bar{D}_i$. We then apply $\mathbf{Q}_k$ to dense depth maps $\bar{D}_i$ instead. We emphasize that the depth estimator is *not* used during inference or distillation. Its purpose is only for the model to learn a consistent definition of scale during training. These two fixes lead to our proposed normalization, which is fully viewer-centric. We define it as

其中在 $\mathbf{M}_{\text{6DoF+1, agg.}}$ 中，q 是应用于场景相机平移分量以计算相对位姿的尺度。这样，$\mathbf{M}_{\text{6DoF+1, agg.}}$ 与 $\mathbf{M}_{\text{6DoF+1, norm.}}$ 不同，因为相机条件表示包含了来自深度图 D_i 的可见内容尺度的信息。虽然以 $\mathbf{M}_{\text{6DoF+1, agg.}}$ 为条件能改进性能，但存在两个问题。第一个问题源于在所有图像上聚合分位数。在图 6 中，向场景添加额外的相机 C 改变了 $\mathbf{M}_{\text{6DoF+1, agg.}}$ 的值，尽管场景的其他部分没有改变。这会让从相机 A 或相机 B 做视角合成的任务更加模糊。为了确保这是不可能的，我们可以简单地消除在场景中所有深度图分位数上的聚合步骤。第二个问题源于我们用于训练的混合数据集中的不同深度统计。ORB-SLAM 通常比 COLMAP 生成更稀疏的深度图，因此 $\mathbf{Q}_k$ 的值对每个数据集可能有不同的含义。因此，我们使用现成的深度估计器 [29] 填充深度图中的孔洞。我们将以这种方式填充的深度 D_i 表示为 $\bar{D}_i$。我们然后将 $\mathbf{Q}_k$ 应用于密集深度图 $\bar{D}_i$。我们强调深度估计器在推理或蒸馏期间不被使用。它的目的仅是让模型在训练期间学习尺度的一致定义。这两个修复导致了我们提议的归一化，这是完全以观察者为中心的。我们将其定义为

DTU 上的 NVS	LPIPS ↓	PSNR ↑	SSIM ↑
DS-NeRF†	0.649	12.17	0.410
PixelNeRF	0.535	15.55	0.537
SinNeRF	0.525	**16.52**	**0.560**
DietNeRF	0.487	14.24	0.481
NeRDi	0.421	14.47	0.465
ZeroNVS （本文）	**0.380**	13.55	0.469

表 1. **与最先进方法的对比**。尽管是唯一未在 DTU 上微调的方法，我们在 DTU 上的 LPIPS 实现了最先进性能。† = Xu 等 [42] 中报告的性能。

$$q_i = \mathbf{Q}_{20}(\bar{D}_i) \ ,$$

$$\mathbf{M}_{\text{6DoF+1, viewer}}(D, f, E, i, j) = \left[\mathbf{R}\left(E_i^{-1}E_j, \frac{1}{q_i}\right), f\right] \ ,$$

where in $\mathbf{M}_{\text{6DoF+1, viewer}}$, the scale q_i applied to the cameras is dependent only on the depth map in the input view $\bar{D}_i$, different from $\mathbf{M}_{\text{6DoF+1, agg.}}$ where the scale q computed by aggregating over all D_i. At inference the value of q_i can be chosen heuristically without compromising performance. Correcting for the scale ambiguities in this way improves metrics, which we show in Section 4.

其中在 $\mathbf{M}_{\text{6DoF+1, viewer}}$ 中，应用于相机的尺度 q_i 仅取决于输入视角 $\bar{D}_i$ 中的深度图，与 $\mathbf{M}_{\text{6DoF+1, agg.}}$ 中通过在所有 D_i 上聚合计算的尺度 q 不同。在推理时，q_i 的值可以启发式地选择而不影响性能。以这种方式纠正尺度歧义改进了度量，我们在第 4 节中展示。

3.4. 通过 SDS 锚定改进多样性（Improving Diversity with SDS Anchoring）

Diffusion models trained with the improved camera conditioning representation $\mathbf{M}_{\text{6DoF+1, viewer}}$ achieve superior view synthesis results via 3D SDS distillation. However, for large viewpoint changes, novel view synthesis is also a generation problem, and it may be desirable to generate diverse and plausible contents rather than contents that are only optimal on average for metrics such as PSNR, SSIM, and LPIPS. However, Poole 等 [27] noted that even when the underlying generative model produces diverse images, SDS distillation of that model tends to seek a single mode.

NVS	LPIPS ↓	PSNR ↑	SSIM ↑
Mip-NeRF 360 数据集			
Zero-1-to-3	0.667	11.7	0.196
PixelNeRF	0.718	**16.5**	**0.556**
ZeroNVS （本文）	**0.625**	13.2	0.240
DTU 数据集			
Zero-1-to-3	0.472	10.70	0.383
PixelNeRF	0.738	10.46	0.397
ZeroNVS （本文）	**0.380**	**13.55**	**0.469**

表 2. **零样本对比**。与在我们混合数据集上重新训练的基线的对比。即使在 Zero-1-to-3 用相同场景数据训练的情况下，ZeroNVS 也超越了 Zero-1-to-3。广泛的视频对比在补充材料中。

For novel view synthesis of scenes via SDS, we observe a unique manifestation of this diversity issue: lack of diversity is especially apparent in inferred backgrounds. Often, SDS distillation predicts a gray or monotone background for regions not observed by the input camera.

用改进的相机条件表示 $\mathbf{M}_{\text{6DoF+1, viewer}}$ 训练的扩散模型通过 3D SDS 蒸馏实现优异的视角合成结果。然而，对于大的视角变化，新视角合成也是一个生成问题，可能希望生成多样且合理的内容，而不是仅在 PSNR、SSIM 和 LPIPS 等指标上平均最优的内容。然而，Poole 等 [27] 指出即使底层生成模型产生多样化的图像，该模型的 SDS 蒸馏也倾向于寻求单一模式。对于通过 SDS 的场景新视角合成，我们观察到这个多样性问题的独特表现：缺乏多样性在推理的背景中特别明显。通常，SDS 蒸馏对于输入相机未观察到的区域预测灰色或单调背景。

To remedy this, we propose "SDS anchoring" (Figure 4).

为了解决这个问题，我们提出"SDS 锚定"（图 4）。

With SDS anchoring, we first directly sample k novel views $\hat{\boldsymbol{X}}_k = \{\hat{X}_j\}_{j=1}^k$ with $\hat{X}_j \sim p(X_j | X_i, \mathbf{M}(D, f, E, i, j))$ from poses evenly spaced in azimuth for maximum scene coverage. We sample the novel views via DDIM [36], which does not have the mode collapse issues of SDS. Each novel view is generated conditional on the input view. Then, when optimizing the SDS objective, we condition the diffu-

输入视角　　真值新视角　　ZeroNVS （本文）　　PixelNeRF

PSNR=10.8, SSIM=0.22　　PSNR=12.2, SSIM=0.30

图 7. PSNR 和 SSIM 对于视角合成评估的限制。不对齐可能导致语义上更合理的预测获得更差的 PSNR 和 SSIM 值。

DTU 上的 NVS	LPIPS ↓	PSNR ↑	SSIM ↑
全部数据集	**0.421**	**12.2**	**0.444**
-ACID	0.446	11.5	0.405
-CO3D	0.456	10.7	0.407
-RealEstate10K	0.435	12.0	0.429

表 3. [**Ablation study on training data.** Training on all datasets improves performance.] **训练数据消融实验**。在所有数据集上训练提升了性能。

sion model not on the input view, but on the nearest view.

通过 SDS 锚定，我们首先直接从方位角均匀间隔的位姿采样 k 个新视角 $\hat{\boldsymbol{X}}_k = \{\hat{X}_j\}_{j=1}^{k}$，其中 $\hat{X}_j \sim p(X_j|X_i, \mathbf{M}(D, f, E, i, j))$，以实现最大场景覆盖。我们通过 DDIM [36] 采样新视角，它没有 SDS 的模式坍缩问题。每个新视角都是以输入视角为条件生成的。然后，当优化 SDS 目标时，我们不以输入视角、而以最近的视角作为扩散模型的条件。

As shown quantitatively in a user study in Section 4 and qualitatively in Figure 9, SDS anchoring produces more diverse background contents. We provide more details about the setup of SDS anchoring in the supplementary.

如第 4 节中用户研究中定量所示，以及图 9 中定性所示，SDS 锚定产生了更多样化的背景内容。我们在补充材料中提供了有关 SDS 锚定设置的更多详细信息。

4. 实验（Experiments）

4.1. 设置（Setup）

Datasets. Our models are trained on a mixture dataset consisting of CO3D [31], ACID [17], and RealEstate10K [45]. Each example is sampled uniformly at random from the three datasets. We train at 256×256 resolution, center-cropping and adjusting the intrinsics for each image and scene as necessary. We train using our representation $\mathbf{M}_{\text{6DoF+1, viewer}}$ unless otherwise specified. We provide more training details in the supplementary.

数据集。我们的模型在由 CO3D、ACID 和 RealEstate10K 组成的混合数据集上训练。各样例从三个数据集中均匀随机采样。我们以 256×256 分辨率训练，对每个图像和场景进行中心裁剪并按需调整内参。除非另有说明，否则使用我们的表示 $\mathbf{M}_{\text{6DoF+1, viewer}}$ 训练。

我们在补充材料中提供了更多训练细节。

We evaluate our trained diffusion models on held-out subsets of CO3D, ACID, and RealEstate10K respectively, for 2D novel view synthesis. Our main evaluations are for zero-shot 3D consistent novel view synthesis, where we compare against other techniques on the DTU benchmark [1] and on the Mip-NeRF 360 dataset [2]. We evaluate at 256×256 resolution except for DTU, for which we use 400×300 resolution to be comparable to prior art.

我们在 CO3D、ACID 和 RealEstate10K 的留出

图 8. [Qualitative comparison between baseline methods and our method.] 基线方法与本方法的定性对比。

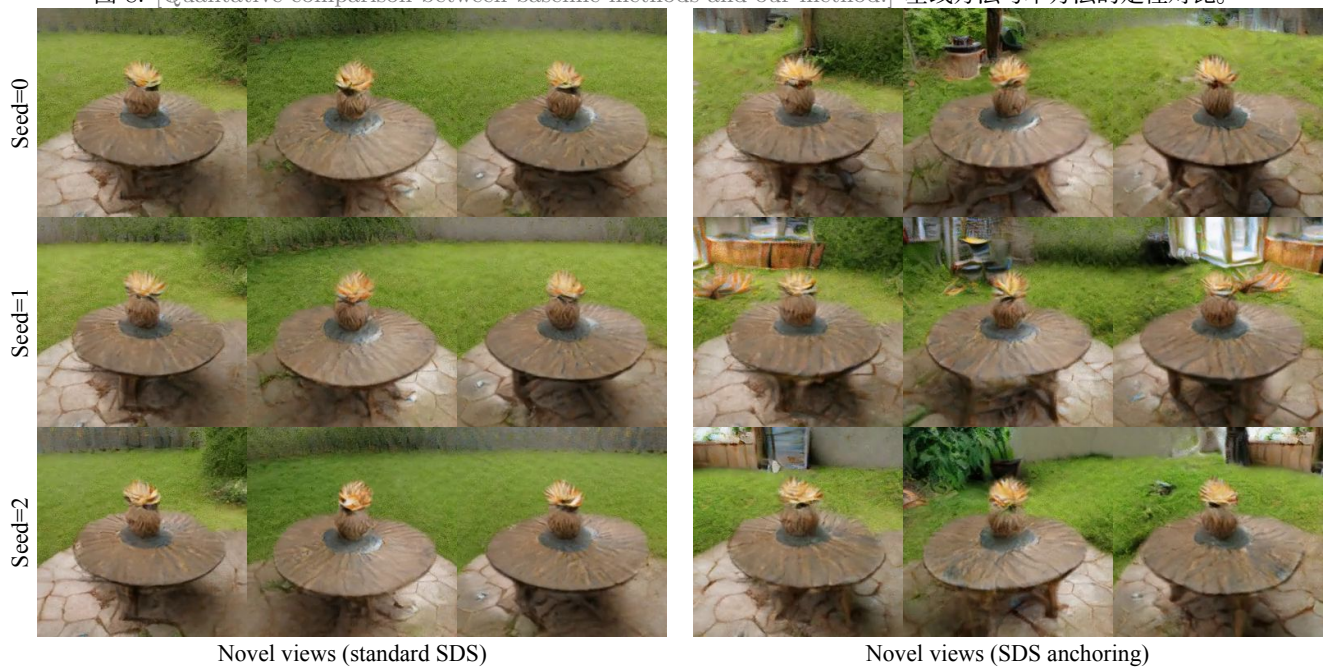

图 9. [Whereas standard SDS (left) tends to predict monotonous backgrounds, our SDS anchoring (right) generates more diverse background contents. Additionally, SDS anchoring generates noticeably different results depending on the seed.] 标准 SDS（左图）易生成单调背景，而我们的 SDS 锚定（右图）可生成更多样化的背景内容。此外，SDS 锚定会根据随机种子生成明显不同的结果。

子集上分别评估训练得到的扩散模型用于二维新视角合成。主要评估针对零样本三维一致新视角合成，我们在 DTU 基准和 Mip-NeRF 360 数据集上与其他方法比较。除 DTU 外，我们均以 256×256 分辨率评估；对于 DTU，为与先前工作比较，使用 400×300 分辨率。

Implementation details. Our diffusion model training code is written in PyTorch and based on the public code for Zero-1-to-3 [19]. We initialize from the pretrained Zero-1-to-3-XL, swapping out the conditioning module to accommodate our novel parameterizations. Our distillation code is implemented in Threestudio [14]. We use a custom NeRF

	2D 新视角合成									3D NeRF 蒸馏		
	CO3D			RealEstate10K			ACID			DTU		
条件表示	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
$\mathbf{M}_{\text{Zero-1-to-3}}$	12.0	.366	.590	11.7	.338	.534	15.5	.371	.431	10.3	.384	.477
$\mathbf{M}_{\text{6DoF+1}}$	12.2	.370	.575	12.5	.380	.483	15.2	.363	.445	9.5	.347	.472
$\mathbf{M}_{\text{6DoF+1, norm.}}$	12.9	.392	.542	12.9	.408	.450	16.5	.398	.398	11.5	.422	.421
$\mathbf{M}_{\text{6DoF+1, agg.}}$	13.2	.402	.527	**13.5**	**.441**	.417	16.9	.411	.378	**12.2**	.436	**.420**
$\mathbf{M}_{\text{6DoF+1, viewer}}$	**13.4**	**.407**	**.515**	**13.5**	.440	**.414**	**17.1**	**.415**	**.368**	**12.2**	**.444**	.421

表 4. [**Ablation study on the conditioning representation M.** Our $\mathbf{M}_{\text{6DoF+1, viewer}}$ matches or outperforms other representations.] **条件表示 M 的消融实验**。我们的 $\mathbf{M}_{\text{6DoF+1, viewer}}$ 与其他表示方法相当或更优。

network combining features of Mip-NeRF 360 with Instant-NGP [23]. The noise schedule is annealed following Wang 等 [39]. For details please see the supplementary.

实现细节。我们的扩散模型训练代码用 PyTorch 编写，基于 Zero-1-to-3 的公开代码。我们从预训练的 Zero-1-to-3-XL 初始化，并替换条件模块以适应我们的新参数化。蒸馏代码在 Threestudio 中实现。我们使用结合 Mip-NeRF 360 和 Instant-NGP 特性的自定义 NeRF 网络。噪声调度按 Wang 等 [39] 的方法退火。详见补充材料。

4.2. 主要结果（Main Results）

We evaluate all methods using the standard set of novel view synthesis metrics: PSNR, SSIM, and LPIPS. We weigh LPIPS more heavily in the comparison due to the well-known issues with PSNR and SSIM as discussed in [6, 10]. We confirm that PSNR and SSIM do not correlate well with performance in our setting, as illustrated in Figure 7.

我们使用标准的新视角合成指标（PSNR、SSIM 和 LPIPS）评估所有方法。在比较中我们更看重 LPIPS，因为 PSNR 和 SSIM 的已知局限性在 [6, 10] 中有讨论。我们证实在我们的设置中 PSNR 和 SSIM 与性能相关性不强，如图 7 所示。

The results are shown in Table 1. We first compare against baseline methods DS-NeRF [11], PixelNeRF [44], SinNeRF [42], DietNeRF [15], and NeRDi [10] on DTU. All these methods are trained on DTU, but we achieve a state-of-the-art LPIPS despite being fully zero-shot. We show visual comparisons in Figure 8.

结果见表 1。首先，我们在 DTU 上与 DS-NeRF、PixelNeRF、SinNeRF、DietNeRF 和 NeRDi 等基线方法比较。虽然这些方法均在 DTU 上训练，但我们在完全零样本的设置下仍取得最先进的 LPIPS 性能。定性对比见图 8。

DTU scenes are limited to relatively simple forward-facing scenes. Therefore, we introduce a more challenging benchmark dataset, the Mip-NeRF 360 dataset, to benchmark the task of 360-degree view synthesis from a single image. We use this benchmark as a zero-shot benchmark, and train three baseline models on our mixture dataset to compare zero-shot performance. Our method is the best on LPIPS for this dataset. On DTU, we exceed Zero-1-to-3 and the zero-shot PixelNeRF model on all metrics, not just LPIPS. Performance is shown in Table 2. All numbers for our method and Zero-1-to-3 are for NeRFs predicted from SDS distillation unless otherwise noted.

DTU 场景仅限于相对简单的前向场景。因此，我们引入更具挑战性的基准数据集 Mip-NeRF 360，用于基准测试单张图像的 360 度视角合成任务。我们将其作为零样本基准，并在我们的混合数据集上训练三个基线模型来比较零样本性能。我们的方法在该数据集上的 LPIPS 性能最优。在 DTU 上，我们在所有指标上都超过了 Zero-1-to-3 和零样本 PixelNeRF，不仅仅是 LPIPS。性能结果见表 2。除非另有说明，否则我们

的方法和 Zero-1-to-3 的数字均来自 SDS 蒸馏预测的 NeRF。

Limited diversity is a known issue with SDS-based methods, but the long run time of SDS-based methods makes typical generation-based metrics such as FID cost-prohibitive. Therefore, we quantify the improved diversity from SDS anchoring via a user study of 21 users on the Mip-NeRF 360 dataset. Users were asked to compare scenes predicted with and without SDS anchoring along three dimensions: Realism, Creativity, and Overall Preference. The preferences for SDS anchoring were: Realism (78%), Creativity (82%), and Overall Preference (80%). The supplementary provides more details about the setup of the study. Figure 9 includes qualitative examples that show the advantages of SDS anchoring, and the supplementary webpage contains the videos which were shown in the study.

多样性有限是 SDS 类方法的已知问题，但 SDS 方法运行时间长，FID 等基于生成的指标计算代价过高。因此，我们通过 Mip-NeRF 360 数据集上的 21 位用户参与的用户研究来量化 SDS 锚定改进的多样性。用户被要求沿着真实感、创意性和总体偏好三个维度比较有无 SDS 锚定的预测场景。对 SDS 锚定的偏好分别为：真实感（78

We conduct multiple ablations to verify our contributions. We verify the benefits of each of our multiple multi-view scene datasets in Table 3. Removing any of the three datasets on which ZeroNVS is trained reduces performance. In Table 4, we analyze the diffusion model's performance on held-out subsets of our datasets, with the various parameterizations discussed in Section 3. We see that as the conditioning parameterization is further refined, the performance continues to increase. Due to computational constraints, we train the ablation diffusion models for fewer steps than our main model, hence the slightly worse performance relative to Table 1. We provide more details in the supplementary.

我们做了多项消融实验来验证我们的贡献。表 3 验证了多视图场景数据集各部分的益处。移除任何一个训练 ZeroNVS 所用的三个数据集都会降低性能。在表 4 中，我们分析了扩散模型在数据集留出子集上的性能，使用了第 3 节讨论的各种参数化方法。随着条件参数化的进一步改进，性能不断提升。由于计算资源限制，消融扩散模型的训练步数少于主模型，因此相对于表 1 性能略低。详见补充材料。

5. 结论（Conclusion）

We have introduced ZeroNVS, a system for 3D-consistent novel view synthesis from a single image for generic scenes. We showed its state-of-the-art performance on existing NVS benchmarks and proposed the Mip-NeRF 360 dataset as a more challenging benchmark for single-image NVS.

我们提出了 ZeroNVS，一个从单张图像实现通用场景三维一致新视角合成的系统。我们展示了其在现有 NVS 基准上的最先进性能，并提议 Mip-NeRF 360 数据集作为单图像 NVS 的更具挑战性的基准。

致谢（Acknowledgments）。 The work is in part supported by NSF CCRI #2120095, RI #2211258, ONR MURI N00014-22-1-2740, and Google.

本工作部分受 NSF CCRI #2120095、RI #2211258、ONR MURI N00014-22-1-2740 以及 Google 的资助。

参考文献

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. 11

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 4, 11, 20

[3] Thomas Breuel. Webdataset library, 2020. 18, 19

[4] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *CVPR*, 2021. 5

[5] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh

Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2021. 5

[6] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *ICCV*, 2023. 5, 13

[7] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation. In *ICCV*, 2023. 5

[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. 7

[9] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. *arXiv preprint arXiv:2307.05663*, 2023. 1, 3, 7

[10] 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 *CVPR*, 2022. 5, 13

[11] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022. 13

[12] 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. 5

[13] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A Style-based 3D-aware Generator for High-resolution Image Synthesis. In *ICLR*, 2022. 5

[14] 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, and Song-Hai Zhang. threestudio: A unified framework for 3D content generation, 2023. 12

[15] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021. 5, 13

[16] 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 *CVPR*, 2023. 5

[17] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021. 3, 11

[18] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 5

[19] 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. In *CVPR*, 2023. 1, 3, 5, 6, 12, 18

[20] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 5

[21] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. ATT3D: Amortized text-to-3D object synthesis. In *ICCV*, 2023. 5

[22] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. RealFusion: 360° reconstruction of any object from a single image. In *CVPR*, 2023. 3, 5

[23] 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, 2022. 13

[24] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 3

[25] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *ICCV*, 2019. 5

[26] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 5

[27] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2022. 3, 5, 10

[28] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3, 5, 6, 20

[29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 9, 18

[30] 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 (TPAMI)*, 44 (3), 2022. 8, 9, 18

[31] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*, 2021. 3, 11

[32] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. VQ3D: Learning a 3D-aware generative model on ImageNet. In *ICCV*, 2023. 5

[33] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3

[34] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. EpiGRAF: Rethinking training of 3D GANs. In *NeurIPS*, 2022. 5

[35] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3D generation on ImageNet. In *ICLR*, 2023. 5

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv::2010.02502*, 2020. 8, 10, 11

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 4

[38] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score Jacobian chaining: Lifting pretrained 2D diffusion models for 3D generation. *arXiv preprint arXiv:2212.00774*, 2022. 5

[39] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 5, 13

[40] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 7

[41] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3D-aware image generation using 2D diffusion models. In *ICCV*, 2023. 5

[42] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. SinNeRF: Training neural radiance fields on complex scenes from a single image. In *ECCV*, 2022. 10, 13

[43] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3D scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 8, 9

[44] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 5, 13

[45] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 3, 9, 11

A. 细节：扩散模型训练（Details: Diffusion model training）

A.1. 模型（Model）

We train diffusion models for various camera conditioning parameterizations: $\mathbf{M}_{\text{Zero-1-to-3}}$, $\mathbf{M}_{\text{6DoF+1}}$, $\mathbf{M}_{\text{6DoF+1, norm.}}$, $\mathbf{M}_{\text{6DoF+1, agg.}}$, and $\mathbf{M}_{\text{6DoF+1, viewer}}$. Our runtime is identical to Zero-1-to-3 [19] as the camera conditioning novelties we introduce add negligible overhead and can be done mainly in the dataloader. We train our main model for $60,000$ steps with batch size 1536. We find that performance tends to saturate after about $20,000$ steps for all models, though it does not decrease. For inference of the 2D diffusion model, we use 500 DDIM steps and guidance scale 3.0.

我们为各种相机条件参数化训练扩散模型：$\mathbf{M}_{\text{Zero-1-to-3}}$、$\mathbf{M}_{\text{6DoF+1}}$、$\mathbf{M}_{\text{6DoF+1, norm.}}$、$\mathbf{M}_{\text{6DoF+1, agg.}}$ 和 $\mathbf{M}_{\text{6DoF+1, viewer}}$。我们的运行时间与 Zero-1-to-3 [19] 相同，因为我们引入的相机条件新方法的开销可以忽略，主要可在数据加载器中完成。我们以批量大小 1536 为主模型训练 60,000 步。我们发现所有模型在约 20,000 步后性能趋向饱和，但不会下降。对于 2D 扩散模型推理，我们使用 500 DDIM 步和 3.0 的引导尺度。

关于 $\mathbf{M}_{\text{6DoF+1}}$ 的细节（Details for）：

To embed the field of view f in radians, we use a 3-dimensional vector consisting of $[f, \sin(f), \cos(f)]$. When concatenated with the $4 \times 4 = 16$-dimensional relative pose matrix, this gives a 19-dimensional conditioning vector.

为了以弧度嵌入视场角 f，我们使用由 $[f, \sin(f), \cos(f)]$ 组成的 3 维向量。当与 $4 \times 4 = 16$ 维相对位姿矩阵连接时，得到 19 维条件向量。

关于 $\mathbf{M}_{\text{6DoF+1, viewer}}$ 的细节（Details for）：

We use the DPT-SwinV2-256 depth model [29] to infill depth maps from ORB-SLAM and COLMAP on the ACID, RealEstate10K, and CO3D datasets. We infill the invalid depth map regions only after aligning the disparity from the monodepth estimator to the ground-truth sparse depth map via the optimal scale and shift following Ranftl 等 [30]. We downsample the depth map $4\times$ so that the quantile function is evaluated quickly.

我们使用 DPT-SwinV2-256 深度模型 [29] 在 ACID、RealEstate10K 和 CO3D 数据集上从 ORB-SLAM 和 COLMAP 填补深度图。仅在通过最优尺度与平移将单目深度估计器的视差对齐到真值稀疏深度图后，我们才填补无效深度图区域，遵循 Ranftl 等 [30]。我们对深度图下采样 $4\times$，以便快速评估分位数函数。

At inference time, the value of $\mathbf{Q}_{20}(\bar{D})$ may not be known since input depth map D is unknown. Therefore there is a question of how to compute the conditioning embedding at inference time. Values of $\mathbf{Q}_{20}(\bar{D})$ between $.7-1$. work for most images and it can be chosen heuristically. For instance, for DTU we uniformly assume a value of .7, which seems to work well. Note that any value of $\mathbf{Q}_{20}(\bar{D})$ is presumably possible; it is only when this value is incompatible with the desired SDS camera radius that distillation may fail, since the cameras may intersect the visible content.

在推理时，由于输入深度图 D 未知，$\mathbf{Q}_{20}(\bar{D})$ 的值可能不知道。因此存在如何在推理时计算条件嵌入的问题。$\mathbf{Q}_{20}(\bar{D})$ 在 $.7-1$ 之间的值对大多数图像都有效，可以启发式地选择。例如，对于 DTU，我们统一假设值为 .7，这似乎效果很好。注意，$\mathbf{Q}_{20}(\bar{D})$ 的任何值都可能存在；只有当该值与所需的 SDS 相机半径不兼容时，蒸馏才可能失败，因为相机可能与可见内容相交。

A.2. 数据加载器（Dataloader）

One significant engineering component of our work is our design of a streaming dataloader for multiview data, built on top of WebDataset [3]. Each dataset is sharded and each shard consists of a sequential tar archive of scenes. The shards can be streamed in parallel via multiprocessing. As a shard is streamed, we yield random pairs of views from scenes according to a "rate" parameter that determines how densely to sample each scene. This parameter allows a trade-off between fully random sampling (lower rate) and biased sampling (higher rate) which can be tuned according to the available network bandwidth. Individual streams from each dataset are then combined and sampled

randomly to yield the mixture dataset. We will release the code together with our main code release.

我们工作中的一个重要工程组件是我们为多视图数据设计的流式数据加载器，构建于 WebDataset [3] 之上。每个数据集都被分片，每个分片包含一个顺序的场景 tar 归档。这些分片可以通过多进程并行流式传输。在流式传输分片时，我们根据控制每个场景采样密度的" 速率" 参数从场景中生成随机视角对。该参数允许在完全随机采样（低速率）和有偏采样（高速率）之间权衡，可根据可用网络带宽调整。来自每个数据集的各个流随后被合并并随机采样以生成混合数据集。代码将随主代码一并发布。

B. 细节：NeRF 预测与蒸馏（Details: NeRF prediction and distillation）

B.1. SDS 锚定（SDS Anchoring）

We propose SDS anchoring in order to increase the diversity of synthesized scenes. We sample 2 anchors at 120 and 240 degrees of azimuth relative to the input camera.

我们提出 SDS 锚定以增加合成场景的多样性。我们在相对于输入相机的方位角 120 和 240 度处采样 2 个锚点。

One potential issue with SDS anchoring is that if the samples are 3D-inconsistent, the resulting generations may look unusual. Furthermore, traditional SDS already performs quite well except if the criterion is diverse backgrounds. Therefore, to implement anchoring, we randomly choose with probability .5 either the input camera and view or the nearest sampled anchor camera and view as guidance. If the guidance is an anchor, we "gate" the gradients flowing back from SDS according to the depth of the NeRF render, so that only depths above a certain threshold (1.0 in our experiments) receive guidance from the anchors. This seems to mostly mitigate artifacts from 3D-inconsistency of foreground content, while still allowing for rich backgrounds. We show video results for SDS anchoring on the webpage.

SDS 锚定的一个潜在问题是，如果样本存在 3D 不一致，生成的结果可能看起来不寻常。此外，传统 SDS 表现相当好，除非标准是多样化背景。因此，为了实现锚定，我们以概率 .5 随机选择输入相机与视角或最近采样的锚点相机与视角作为引导。如果引导是锚点，我们根据 NeRF 渲染的深度对从 SDS 流回的梯度门控，只有深度超过某个阈值（我们的实验中为 1.0）才能从锚点获得引导。这似乎能在很大程度上减轻前景内容 3D 不一致导致的伪影，同时仍然允许丰富的背景。我们在网页上展示了 SDS 锚定的视频结果。

B.2. 超参数（Hyperparameters）

NeRF distillation via involves numerous hyperparameters such as for controlling lighting, shading, camera sampling, number of training steps, training at progressively increasing resolutions, loss weights, density blob initializations, optimizers, guidance weight, and more. We will share a few insights about choosing hyperparameters for scenes here, and release the full configs as part of our code release.

NeRF 蒸馏涉及众多超参数，如控制光照、着色、相机采样、训练步数、渐进式增分分辨率训练、损失权重、密度团初始化、优化器、引导权重等。我们将在这里分享一些关于为场景选择超参数的见解，并作为代码发布的一部分发布完整配置。

噪声调度（Noise scheduling）： We found that ending training with very low maximum noise levels such as .025 seemed to benefit results, particularly perceptual metrics like LPIPS. We additionaly found a significant benefit on 360-degree scenes such as in the Mip-NeRF 360 dataset to scheduling the noise "anisotropically;" that is, reducing the noise level more slowly on the opposite end from the input view. This seems to give the optimization more time to solve the challenging 180-degree views at higher noise levels before refining the predictions at low noise levels.

我们发现以非常低的最大噪声水平（如 .025）结束训练似乎对结果有益，特别是对于感知指标如 LPIPS。我们还发现对于 360 度场景（如 Mip-NeRF 360 数据集）各向异性地调度噪声有显著益处；即在距离输入视角的相反一端更缓慢地降低噪声水平。这似乎给优化更多时间在更高噪声水平处解决具有挑战性的 180 度视角，然后在低噪声水平处细化预测。

其他（Miscellaneous）： Progressive azimuth and elevation sampling following [28] was also found to be very important for training stability. Training resolution progresses stagewise, first with batch size 6 at 128x128 and then with batch size 1 at 256×256.

渐进方位角和仰角采样遵循 [28] 也被发现对训练稳定性非常重要。训练分辨率分阶段进行，首先批量大小 6、128×128 分辨率，然后批量大小 1、256×256 分辨率。

C. 实验设置（Experimental setups）

For our main results on DTU and Mip-NeRF 360, we train our model and Zero-1-to-3 for $60,000$ steps. Performance for our method seems to saturate earlier than for Zero-1-to-3, which trained for about $100,000$ steps; this may be due to the larger dataset size. Objaverse, with $800,000$ scenes, is much larger than the combination of RealEstate10K, ACID, and CO3D, which are only about $95,000$ scenes in total.

对于我们在 DTU 和 Mip-NeRF 360 上的主要结果，我们将模型和 Zero-1-to-3 训练 $60,000$ 步。我们方法的性能似乎比 Zero-1-to-3 更早饱和，后者训练了约 $100,000$ 步；这可能是由于较大的数据集大小。包含 $800,000$ 个场景的 Objaverse 远大于 RealEstate10K、ACID 和 CO3D 的组合，后者总共仅约 $95,000$ 个场景。

For the retrained PixelNeRF baseline, we retrained it on our mixture dataset of CO3D, ACID, and RealEstate10K for about $560,000$ steps.

对于重新训练的 PixelNeRF 基线，我们在 CO3D、ACID 和 RealEstate10K 的混合数据集上重新训练了约 $560,000$ 步。

C.1. 主要结果（Main results）

For all single-image NeRF distillation results, we assume the camera elevation, field of view, and content scale are given. These parameters are identical for all DTU scenes but vary across the Mip-NeRF 360 dataset. For DTU, we use the standard input views and test split from from prior work. We select Mip-NeRF 360 input view indices manually based on two criteria. First, the views are well-approximated by a 3DoF pose representation in the sense of geodesic distance between rotations. This is to ensure fair comparison with Zero-1-to-3, and for compatibility with Threestudio's SDS sampling scheme, which also uses 3 degrees of freedom. Second, as much of the scene content as possible must be visible in the view. The exact values of the input view indices are given in Table 5.

对于所有单图像 NeRF 蒸馏结果，我们假设相机仰角、视场角和内容尺度已给定。这些参数对所有 DTU 场景相同，但在 Mip-NeRF 360 数据集中变化。对于 DTU，我们使用来自先前工作的标准输入视角和测试分割。我们基于两个标准手动选择 Mip-NeRF 360 输入视角索引。首先，视角由 3DoF 姿态表示很好地近似，在旋转间的测地距离意义上。这是为了确保与 Zero-1-to-3 的公平比较，以及与 Threestudio 的 SDS 采样方案兼容，后者也使用 3 个自由度。其次，尽可能多的场景内容必须在视角中可见。输入视角索引的确切值在表 5 中给出。

The field of view is obtained via COLMAP. The camera elevation is set automatically via computing the angle between the forward axis of the camera and the world's XY-plane, after the cameras have been standardized via PCA following Barron 等 [2].

视场角通过 COLMAP 获得。相机仰角通过计算相机前向轴与世界 XY 平面之间的角度自动设置，相机经过 PCA 标准化后遵循 Barron 等 [2]。

One challenge is that for both the Mip-NeRF 360 and DTU datasets, the scene scales are not known by the zero-shot methods, namely Zero-1-to-3, our method, and our retrained PixelNeRF. Therefore, for the zero-shot methods, we manually grid search for the optimal world scale in intervals of .1 to find the appropriate world scale for each scene in order to align the predictions to the generated scenes. Between five to nine samples within $[.3, .4, .5, .6, .7, .8, .9, 1., 1.1, 1.2, 1.3, 1.4, 1.5]$ generally suffices to find the appropriate scale. Even correcting for the scale misalignment issue in this way, the zero-shot methods generally do worse on pixel-aligned metrics like SSIM and PSNR compared with methods that have been

fine-tuned on DTU.

一个挑战是对于 Mip-NeRF 360 和 DTU 数据集，零样本方法（即 Zero-1-to-3、我们的方法和重新训练的 PixelNeRF）不知道场景尺度。因此，对于零样本方法，我们手动以 .1 的间隔网格搜索最优世界尺度，以为每个场景找到适当的世界尺度，以便将预测对齐到生成的场景。在 $[.3, .4, .5, .6, .7, .8, .9, 1., 1.1, 1.2, 1.3, 1.4, 1.5]$ 范围内通常 5 到 9 个样本就足以找到适当的尺度。即使以这种方式更正尺度失配问题，与在 DTU 上微调过的方法相比，零样本方法通常在像素对齐指标（如 SSIM 和 PSNR）上表现更差。

C.2. 用户研究（User study）

We conduct a user study on the seven Mip-NeRF 360 scenes, comparing our method with and without SDS anchoring. We received 21 respondents. For each scene, respondents were shown 360-degree novel view videos of the scene inferred both with and without SDS anchoring. The videos were shown in a random order and respondents were unaware which video corresponded to the use of SDS anchoring. Respondents were asked:

我们对七个 Mip-NeRF 360 场景开展用户研究，比较我们的方法在使用和不使用 SDS 锚定的情况下的表现。我们收到了 21 位受访者。对于每个场景，受访者被展示了用 SDS 锚定和不用 SDS 锚定两种方式推理出的场景的 360 度新视角视频。视频以随机顺序显示，受访者不知道哪个视频对应于使用 SDS 锚定。受访者被问到：

1. Which scene seems more realistic?

   哪个场景看起来更逼真?
2. Which scene seems more creative?

   哪个场景看起来更有创意?
3. Which scene do you prefer?

   你更喜欢哪个场景?

Respondents generally preferred the scenes produced by SDS anchoring, especially with respect to “Which scene seems more creative?”

受访者普遍更喜欢 SDS 锚定生成的场景，特别是对于” 哪个场景看起来更有创意？ ” 这个问题。

场景名	输入视角索引	内容尺度
bicycle	98	.9
bonsai	204	.9
counter	95	.9
garden	63	.9
kitchen	65	.9
room	151	2.
stump	34	.9

表 5. Mip-NeRF 360 数据集的设置 [Setup for the Mip-NeRF 360 dataset]

.

C.3. 消融实验（Ablation studies）

We perform ablation studies on dataset selection and camera representations. For 2D novel view synthesis metrics, we compute metrics on a held-out subset of scenes from the respective datasets, randomly sampling pairs of input and target novel views from each scene. For 3D SDS distillation and novel view synthesis, our settings are identical to the NeRF distillation settings for our main results except that we use shorter-trained diffusion models. We train them for 25,000 steps as opposed to 60,000 steps for computational constraint reasons.

我们对数据集选择和相机表示做消融实验。对于 2D 新视角合成指标，我们在来自各自数据集的保留子集上计算指标，从每个场景中随机采样输入和目标新视角对。对于 3D SDS 蒸馏和新视角合成，我们的设置与我们主要结果的 NeRF 蒸馏设置相同，除了我们使用训练较少的扩散模型。由于计算限制，我们训练 25,000 步而不是 60,000 步。