Wonder3D: Single Image to 3D using Cross-Domain Diffusion

Wonder3D: 使用跨域扩散实现单张图像到三维模型的转换

Xiaoxiao Long [1,3*] , Yuan-Chen Guo [2*‡] , Cheng Lin [1†] , Yuan Liu [1] , Zhiyang Dou [1]

龙笑笑 [1,3*] ， 郭元辰 [2*‡] ， 林程 [1†] ， 刘源 [1] ， 窦智洋 [1]

Lingjie Liu [4] , Yuexin Ma [5] , Song-Hai Zhang [2] , Marc Habermann [6] , Christian Theobalt [6] , Wenping Wang [7†]

刘凌杰 [4] ， 马悦欣 [5] ， 张松海 [2] ， 马克·哈伯曼 [6] ， 克里斯蒂安·特奥巴尔 [6] ， 王文平 [7†]

[1] The University of Hong Kong [2] Tsinghua University [3] VAST

[1] 香港大学 [2] 清华大学 [3] 海量数据先进技术研究中心 (VAST)

[4] University of Pennsylvania [5] Shanghai Tech University [6] MPI Informatik [7] Texas A&M University

[4] 宾夕法尼亚大学 [5] 上海科技大学 [6] 马克斯·普朗克信息研究所 (MPI Informatik) [7] 德克萨斯农工大学

*Core contributions † Corresponding authors ‡ Intern at VAST

* 核心贡献 † 通讯作者 ‡ 海量数据先进技术研究中心 (VAST) 实习生

https://www.xxlong.site/Wonder3D/

Figure 1. Wonder3D reconstructs highly-detailed textured meshes from a single-view image in only $2 \sim 3$ minutes. Wonder3D first generates consistent multi-view normal maps with corresponding color images via a cross-domain diffusion model, and then leverages a novel normal fusion method to achieve fast and high-quality reconstruction. Compared to the prior works, Wonder3D has achieved a leading-level of geometric details with high efficiency.

图 1. Wonder3D 仅需 $2 \sim 3$ 分钟即可从单视图图像中重建出高度精细的纹理网格。Wonder3D 首先通过跨域扩散模型生成具有相应彩色图像的一致多视图法线贴图，然后利用一种新颖的法线融合方法实现快速且高质量的重建。与以往的工作相比，Wonder3D 以高效率实现了领先水平的几何细节。

Abstract

摘要

In this work, we introduce Wonder3D, a novel method for efficiently generating high-fidelity textured meshes from single-view images. Recent methods based on Score Distillation Sampling (SDS) have shown the potential to recover 3D geometry from 2D diffusion priors, but they typically suffer from time-consuming per-shape optimization and inconsistent geometry. In contrast, certain works directly produce $3D$ information via fast network

inferences, but their results are often of low quality and lack geometric details. To holistically improve the quality, consistency, and efficiency of single-view reconstruction tasks, we propose a cross-domain diffusion model that generates multiview normal maps and the corresponding color images. To ensure the consistency of generation, we employ a multiview cross-domain attention mechanism that facilitates information exchange across views and modalities. Lastly, we introduce a geometry-aware normal fusion algorithm that extracts high-quality surfaces from the multi-view 2D representations in only $2 \sim 3$ minutes. Our extensive evaluations demonstrate that our method achieves high-quality reconstruction results, robust generalization, and good efficiency compared to prior works.

在这项工作中，我们介绍了 Wonder3D，这是一种从单视图图像中高效生成高保真纹理网格的新方法。最近基于分数蒸馏采样 (Score Distillation Sampling，SDS) 的方法显示了从二维扩散先验中恢复三维几何形状的潜力，但它们通常存在每个形状优化耗时以及几何形状不一致的问题。相比之下，某些工作通过快速网络推理直接生成 $3D$ 信息，但它们的结果往往质量较低且缺乏几何细节。为了全面提高单视图重建任务的质量、一致性和效率，我们提出了一种跨域扩散模型，该模型可以生成多视图法线贴图和相应的彩色图像。为了确保生成的一致性，我们采用了一种多视图跨域注意力机制，该机制有助于跨视图和跨模态的信息交换。最后，我们引入了一种几何感知法线融合算法，该算法仅需 $2 \sim 3$ 分钟即可从多视图二维表示中提取高质量的表面。我们的大量评估表明，与以往的工作相比，我们的方法实现了高质量的重建结果、强大的泛化能力和良好的效率。

1. Introduction

1. 引言

Reconstructing 3D geometry from a single image [13, 26 , 33, 38, 40, 42, 45] stands as a fundamental task in computer graphics and 3D computer vision, benefiting a wide range of versatile applications such as novel view synthesis [7, 25, 35] , 3D content creation [36, 48] , and robotics grasping [29, 79] . However, this task is notably challenging since it is ill-posed and demands the ability to discern the 3D geometry of both visible and invisible parts. This ability requires extensive knowledge of the 3D world.

从单张图像中重建三维几何形状 [13, 26 ， 33, 38, 40, 42, 45] 是计算机图形学和三维计算机视觉中的一项基础任务，它有利于广泛的应用，如新颖视图合成 [7, 25, 35] , 3D 、内容创作 [36, 48] 和机器人抓取 [29, 79] 。然而，这项任务极具挑战性，因为它是不适定问题，需要具备识别可见和不可见部分 3D 几何形状的能力。这种能力需要对 3D 世界有广泛的了解。

Recently, the field of 3D generation has experienced rapid and flourishing development with the introduction of diffusion models. A growing body of research [5, 31, 47, 63, 67], such as DreamField [24], DreamFusion [47], and Magic3D [31], resort to distilling prior knowledge of 2D image diffusion models or vision language models to create 3D models from text or images via Score Distillation Sampling (SDS) [47]. Despite their compelling results, these methods suffer from two main limitations: efficiency and consistency. The per-shape optimization process typically entails tens of thousands of iterations, involving full-image volume rendering and inferences of the diffusion models. Consequently, it often consumes tens of minutes or even hours on per-shape optimization. Moreover, the 2D prior model operates by considering only a single view at each iteration and strives to make every view resemble the input image. This often results in the generation of 3D shapes exhibiting inconsistencies, thus, often leading to the generation of 3D shapes with inconsistencies such as multiple faces (i.e., the Janus problem [47]).

最近，随着扩散模型的引入，三维 (3D) 生成领域经历了快速而蓬勃的发展。越来越多的研究 [5, 31, 47, 63, 67]，如 DreamField [24]、DreamFusion [47] 和 Magic3D [31]，借助分数蒸馏采样 (Score Distillation Sampling，SDS)[47] 方法，从二维 (2D) 图像扩散模型或视觉语言模型中提取先验知识，以根据文本或图像创建三维模型。尽管这些方法取得了令人瞩目的成果，但它们存在两个主要局限性: 效率和一致性。每个形状的优化过程通常需要进行数万次迭代，涉及全图像体渲染和扩散模型的推理。因此，每个形状的优化过程通常需要花费数十分钟甚至数小时。此外，二维先验模型在每次迭代时仅考虑单个视图，并试图使每个视图都与输入图像相似。这通常会导致生成的三维形状出现不一致的情况，例如出现多张面孔 (即双面神问题 [47])。

There exists another group of works that endeavor to directly produce 3D geometries like point clouds [41, 45, 75, 80], meshes [16, 37], neural fields [1, 4, 8, 15, 17, 21, 26- 28, 44, 46, 65, 76] via network inference to avoid time-consuming per-shape optimization. Most of them attempt to train 3D generative diffusion models from scratch on 3D assets. However, due to the limited size of publicly available 3D datasets, these methods demonstrate poor generalizability, most of which can only generate shapes on specific categories.

还有另一类工作致力于通过网络推理直接生成三维几何图形，如点云 [41, 45, 75, 80]、网格 [16, 37]、神经场 [1, 4, 8, 15, 17, 21, 26 - 30]，以避免耗时的每个形状优化过程。它们中的大多数试图在三维资产上从头开始训练三维生成扩散模型。然而，由于公开可用的三维数据集规模有限，这些方法的泛化能力较差，其中大多数只能生成特定类别的形状。

More recently, several methods have emerged that directly generate multi-view 2D images, with representative works including SyncDreamer [36] and MVDream [55]. By enhancing the multi-view consistency of image generation, these methods can recover 3D shapes from the generated multi-view images. Our method also adopts a multi-view generation scheme to favor the flexibility and efficiency of 2D representations. However, due to only relying on color images, the fidelity of the generated shapes is not well-maintained, and they struggle to recover geometric details or come with enormous computational costs.

最近，出现了几种直接生成多视图二维图像的方法，代表性的工作包括 SyncDreamer [36] 和 MVDream [55]。通过增强图像生成的多视图一致性，这些方法可以从生成的多视图图像中恢复三维形状。我们的方法也采用了多视图生成方案，以提高三维表示的灵活性和效率。然而，由于仅依赖彩色图像，生成形状的保真度无法得到很好的保持，并且难以恢复几何细节，或者会产生巨大的计算成本。

To better address the issues of fidelity, consistency, generalizability and efficiency in the aforementioned works, in this paper, we introduce a new approach to the task of single-view 3D reconstruction by generating multi-view consistent normal maps and their corresponding color images with a cross-domain diffusion model. The key idea is to extend diffusion frameworks to model the joint distribution of two different domains, i.e., normals and colors. The normal map characterizes the undulations and variations presented on the surface of the shape, thus encoding rich detailed geometric information. This allows for the high-fidelity extraction of 3D geometry from 2D normal maps. Adopting the 2D representations enables Wonder3D to be built on the pre-trained Stable Diffusion model [49], where strong priors facilitate zero-shot generalization ability.

为了更好地解决上述工作中在保真度、一致性、泛化能力和效率方面的问题，在本文中，我们提出了一种新的单视图三维重建方法，通过跨域扩散模型生成多视图一致的法线贴图及其对应的彩色图像。关键思想是将扩散框架扩展到对两个不同领域 (即法线和颜色) 的联合分布进行建模。法线贴图表征了形状表面呈现的起伏和变化，从而编码了丰富的详细几何信息。这使得能够从二维法线贴图中高保真地提取三维几何形状。采用二维表示使 Wonder3D 能够基于预训练的 Stable Diffusion 模型 [49] 构建，其中强大的先验知识有助于实现零样本泛化能力。

The following technical designs of Wonder3D make it a robust and efficient tool to create 3D shapes from single images. 1) Cross-domain switcher. The introduced domain switcher allows the diffusion model to generate either normal maps or color images without significantly modifying the original model. 2) Cross-domain attentions. We further leverage cross-domain attention mechanisms to assist in the information exchange between the two domains, ultimately improving consistency and quality. This mechanism facilitates information perception across different domains, enabling our method to recover high-fidelity geometry. 3) Geometry-aware normal fusion. In order to stably extract surfaces from the generated views, we propose a geometry-aware normal fusion algorithm that is robust to subtle inaccurate generations and capable of reconstructing clean and high-quality geometries (see Figure 1).

Wonder3D 的以下技术设计使其成为一个从单张图像创建三维形状的强大而高效的工具。1) 跨域切换器。引入的域切换器允许扩散模型在不显著修改原始模型的情况下生成法线贴图或彩色图像。2) 跨域注意力机制。我们进一步利用跨域注意力机制来辅助两个领域之间的信息交换，最终提高一致性和质量。这种机制促进了不同领域之间的信息感知，使我们的方法能够恢复高保真的几何形状。3) 几何感知法线融合。为了从生成的视图中稳定地提取表面，我们提出了一种几何感知法线融合算法，该算法对细微的不准确生成具有鲁棒性，并且能够重建干净、高质量的几何形状 (见图 1)。

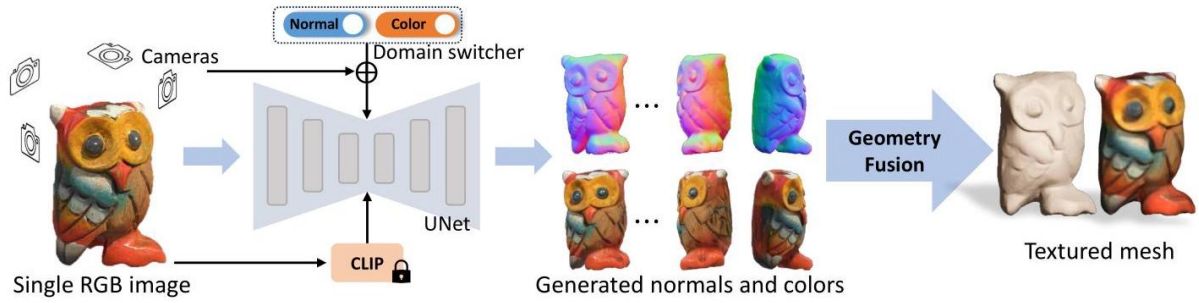

Figure 2. Overview of Wonder3D. Given a single image, Wonder3D takes the input image, the text embedding produced by CLIP model [49], the camera parameters of multiple views, and a domain switcher as conditioning to generate consistent multi-view normal maps and color images. Subsequently, Wonder3D employs an innovative normal fusion algorithm to robustly reconstruct high-quality 3D geometry from the 2D representations, yielding high-fidelity textured meshes.

图 2. Wonder3D 概述。给定单张图像，Wonder3D 以输入图像、CLIP 模型 [49] 生成的文本嵌入、多个视图的相机参数和一个域切换器作为条件，生成一致的多视图法线贴图和彩色图像。随后，Wonder3D 采用一种创新的法线融合算法，从二维表示中稳健地重建高质量的三维几何形状，生成高保真的纹理网格。

We conduct extensive experiments on the Google Scanned Object dataset [14] and various 2D images with

different styles. The experiments validate that Wonder3D achieves a leading level of geometric details with high efficiency and strong generalization among current zero-shot single-view reconstruction methods.

我们在 Google 扫描对象数据集 [14] 和各种不同风格的二维图像上进行了广泛的实验。实验验证了 Wonder3D 在当前零样本单视图重建方法中，以高效率和强泛化能力实现了领先水平的几何细节。

2. Related Works

2. 相关工作

2.1.2D Diffusion Models for 3D Generation

2.1. 用于三维生成的二维扩散模型

Recent compelling successes in 2D diffusion models [9, 22, 51] and large vision language models (e.g., CLIP model [49]) provide new possibilities for generating 3D assets using the strong priors of 2D diffusion models. Pioneering works DreamFusion [47] and SJC [63] propose to distill a 2D text-to-image generation model to generate 3D shapes from texts, and many follow-up works follow such per-shape optimization scheme. For the task of text-to-3D [2, 5, 6, 23, 31, 52, 53, 61, 67, 69, 73, 82] or image-to-3D synthesis [42, 48, 50, 54, 58, 71], these methods typically optimize a 3 D representation (i.e., NeRF, mesh, or SDF), and then leverage neural rendering to generate 2D images from various viewpoints. The images are then fed into the 2D diffusion models or CLIP model for calculating SDS [47] losses, which guide the 3D shape optimization. However, most of these methods always suffer from low efficiency and multi-face problem, where a per-shape optimization consumes tens of minutes and the optimized geometry tends to produce multiple faces due to the lack of explicit 3D supervision.

近期，二维扩散模型 [9, 22, 51] 和大型视觉语言模型 (例如 CLIP 模型 [49]) 取得了令人瞩目的成功，这为利用二维扩散模型的强大先验知识生成三维资产提供了新的可能性。开创性工作 DreamFusion [47] 和 SJC [63] 提出提炼二维文本到图像生成模型，以从文本中生成三维形状，许多后续工作都采用了这种逐形状优化方案。对于文本到三维 [2, 5, 6, 23, 31, 52, 53, 61, 67, 69, 73, 82] 或图像到三维合成 [42, 48, 50, 54, 58, 71] 任务，这些方法通常会优化一个 3 D 表示 (即神经辐射场 (NeRF)、网格或有向距离场 (SDF))，然后利用神经渲染从不同视角生成二维图像。接着将这些图像输入到二维扩散模型或 CLIP 模型中，以计算随机差分方程 (SDS)[47] 损失，该损失用于指导三维形状的优化。然而，这些方法大多存在效率低下和多面问题，即每个形状的优化需要花费数十分钟，并且由于缺乏明确的三维监督，优化后的几何体往往会产生多个面。

2.2.3D Generative Models

2.2. 三维生成模型

Instead of performing a time-consuming per-shape optimization guided by 2D diffusion models, some works attempt to directly train 3D diffusion models based on various 3D representations, like point clouds [41, 45, 75, 80], meshes [16, 37], neural fields [1, 4, 8, 15, 17, 21, 26- 28, 44, 46, 65, 76] However, due to the limited size of public available 3D assets dataset, most of the works have only been validated on limited categories of shapes, and

how to scale up on large datasets is still an open problem. On the contrary, our method adopts 2D representations and, thus, can be built upon the 2D diffusion models [51] whose pre-trained priors significantly facilitate zero-shot generalization ability.

一些工作没有采用由 2D 扩散模型引导的耗时的逐形状优化方法，而是尝试基于各种三维表示 (如点云 [41, 45, 75, 80]、网格 [16, 37]、神经场 [1, 4, 8, 15, 17, 21, 26 - 28, 44, 46, 65, 76]) 直接训练三维扩散模型。然而，由于公开可用的三维资产数据集规模有限，大多数工作仅在有限类别的形状上进行了验证，如何在大型数据集上进行扩展仍然是一个悬而未决的问题。相反，我们的方法采用二维表示，因此可以基于二维扩散模型 [51] 构建，其预训练的先验知识显著促进了零样本泛化能力。

2.3. Multi-view Diffusion Models

2.3. 多视图扩散模型

To generate consistent multi-view images, some efforts [3, 11, 18, 30, 34, 57, 59, 60, 62, 68, 70, 72, 74, 77, 81] are made to extend 2D diffusion models from single-view images to multi-view images. However, most of these methods focus on image generation and are not designed for 3D reconstruction. The recent works Viewset Diffusion [57], SyncDreamer [36], and MVDream [55] share a similar idea to produce consistent multi-view color images via attention layers. However, unlike that normal maps explicitly encode geometric information, reconstruction from color images always suffers from texture ambiguity, and, thus, they either struggle to recover geometric details or require huge computational costs. SyncDreamer [36] requires dense views for 3D reconstruction but still suffers from low-quality geometry and blurring textures. MVDream [55] still resorts to a time-consuming optimization using SDS loss for 3D reconstruction, and its multi-view distillation scheme requires 1.5 hours. In contrast, our method can reconstruct high-quality textured meshes in just 2 minutes.

为了生成一致的多视图图像，一些研究 [3, 11, 18, 30, 34, 57, 59, 60, 62, 68, 70, 72, 74, 77, 81] 致力于将二维扩散模型从单视图图像扩展到多视图图像。然而，这些方法大多专注于图像生成，并非为 3D 重建而设计。近期的工作 Viewset Diffusion [57]、SyncDreamer [36] 和 MVDream [55] 有一个相似的思路，即通过注意力层生成一致的多视图彩色图像。然而，与法线贴图明确编码几何信息不同，从彩色图像进行重建往往会受到纹理歧义的影响，因此，它们要么难以恢复几何细节，要么需要巨大的计算成本。SyncDreamer [36] 在进行三维重建时需要密集视图，但仍然存在几何质量低和纹理模糊的问题。MVDream [55] 在进行三维重建时仍然依赖于使用随机差分方程 (SDS) 损失进行耗时的优化，其多视图提炼方案需要 1.5 小时。相比之下，我们的方法仅需 2 分钟即可重建出高质量的纹理网格。

3. Problem Formulation

3. 问题表述

3.1. Diffusion Models

3.1. 扩散模型

Diffusion models [22, 56] are first proposed to gradually recover images from a specifically designed degradation process, where a forward Markov chain and a Reverse Markov chain are adopted. The forward process will be iteratively applied to the target image z until the image becomes complete Gaussian noise at the end. On the contrary, the reverse chain then is employed to iteratively denoise the corrupted image, i.e., recovering z_{t-1} from z_t by predicting the added random noise ϵ . The readers can refer to [22, 56] for more details about image diffusion models.

扩散模型 [22, 56] 最初是为了从一个专门设计的退化过程中逐步恢复图像而提出的，其中采用了一个前向马尔可夫链和一个反向马尔可夫链。前向过程会迭代地应用于目标图像 z，直到图像最终变成完全的高斯噪声。相反，反向链则用于迭代地对受损图像进行去噪，即通过预测添加的随机噪声 ϵ 从 z_t 中恢复 z_{t-1}。读者可以参考 [22, 56] 以获取有关图像扩散模型的更多详细信息。

3.2.The Distribution of 3D Assets

3.2. 三维资产的分布

Unlike that prior works adopt 3D representations like point clouds, tri-planes, or neural radiance fields, we propose that the distribution of 3D assets, denoted as $p_a(\mathbf{z})$, can be modeled as a joint distribution of its corresponding 2D multiview normal maps and color images. Specifically, given a set of cameras $\{\pi_1, \pi_2, \cdots, \pi_K\}$ and a conditional input image y , we have

与之前的工作采用点云、三平面或神经辐射场等三维表示不同，我们提出三维资产的分布 (表示为 $p_a(\mathbf{z})$) 可以建模为其对应的二维多视图法线图和彩色图像的联合分布。具体来说，给定一组相机 $\{\pi_1, \pi_2, \cdots, \pi_K\}$ 和一个条件输入图像 y，我们有

$$p_a(\mathbf{z}) = p_{nc}\left(n^{1:K}, x^{1:K} \mid y\right), \tag{1}$$

where p_{nc} is the distribution of the normal maps $n^{1:K}$ and color images $x^{1:K}$ observed from a 3D asset conditioned on an image y . For simplicity, we omit the symbol y for this equation in the following discussions. Therefore, our goal is to learn a model f that synthesizes multiple normal maps and color images of a set of camera poses denoted as

其中 p_{nc} 是在图像 y 条件下从三维资产观察到的法线图 $n^{1:K}$ 和彩色图像 $x^{1:K}$ 的分布。为简单起见，在以下讨论中我们省略该方程中的符号 y。因此，我们的目标是学习一个模型 f，该模型能合成一组相机位姿的多个法线图和彩色图像，这些相机位姿表示为

$$\left(n^{1:K}, x^{1:K}\right) = f\left(y, \pi_{1:K}\right). \tag{2}$$

Finally, we can formulate this cross-domain joint distribution as a Markov chain within the diffusion scheme:

最后，我们可以将这种跨域联合分布在扩散方案中表述为一个马尔可夫链:

$$p\left(n_T^{(1:K)}, x_T^{(1:K)}\right) \prod_t p_{nc}\left(n_{t-1}^{(1:K)}, x_{t-1}^{(1:K)} \mid n_t^{(1:K)}, x_t^{(1:K)}\right),$$

(3)

where $p\left(n_T^{(1:K)}, x_T^{(1:K)}\right)$ are Gaussian noises. Our key problem is to characterize the distribution p_{nc}, so that we can sample from this Markov chain to generate normal maps and images.

其中 $p\left(n_T^{(1:K)}, x_T^{(1:K)}\right)$ 是高斯噪声。我们的关键问题是刻画分布 p_{nc}，以便我们可以从这个马尔可夫链中采样来生成法线图和图像。

4. Method

4. 方法

As per our problem formulation in Section 3.2, we propose a multi-view cross-domain diffusion scheme, which operates on two distinct domains to generate multi-view consistent normal maps and color images. The overview of our method is presented in Figure 2. First, our method adopts a multi-view diffusion scheme to generate multi-view normal maps and color images, and enforces the consistency across different views using multi-view attentions (see Section 4.1). Second, our proposed domain switcher allows the diffusion model to operate on more than one domain. A cross-domain attention is proposed to propagate information between the normal domain and color image domain ensuring geometric and visual coherence between the two domains (see Section 4.2). Finally, our novel geometry-aware normal fusion reconstructs the high-quality geometry and appearance from the multi-view 2D normal and color images (see Section 4.3).

根据我们在第 3.2 节中的问题表述，我们提出了一种多视图跨域扩散方案，该方案在两个不同的域上操作，以生成多视图一致的法线图和彩色图像。我们方法的概述如图 2 所示。首先，我们的方法采用多视图扩散方案来生成多视图法线图和彩色图像，并使用多视图注意力机制来确保不同视图之间的一致性 (见第 4.1 节)。其次，我们提出的域切换器允许扩散模型在多个域上操作。我们提出了一种跨域注意力机制，用于在法线域和彩色图像域之间传播信息，确保两个域之间的几何和视觉连贯性 (见第 4.2 节)。最后，我们新颖的几何感知法线融合方法从多视图二维法线图和彩色图像中重建出高质量的几何形状和外观 (见第 4.3 节)。

4.1. Consistent Multi-view Generation

4.1. 一致的多视图生成

The prior 2D diffusion models [33, 49] generate each image separately, so that the resulting images are not geometrically and visually consistent across different views. To enhance consistency among different views, similar to prior works such as SyncDreamer [36] and MVDream [55], we utilize attention mechanism to facilitate information propagation across different views, implicitly encoding multi-view dependencies (as illustrated in Figure 3).

先前的二维扩散模型 [33, 49] 分别生成每张图像，因此生成的图像在不同视图之间在几何和视觉上并不一致。为了增强不同视图之间的一致性，与 SyncDreamer [36] 和 MVDream [55] 等先前的工作类似，我们利用注意力机制来促进不同视图之间的信息传播，隐式地编码多视图依赖关系 (如图 3 所示)。

This is achieved by extending the original self-attention layers to be global-aware, allowing connections to other views within the attention layers. Keys and values from different views are connected to each other to facilitate the exchange of information. By sharing information across different views within the attention layers, the diffusion model perceives multi-view correlation and is capable of generating consistent multi-view color images and normal maps.

这是通过将原始的自注意力层扩展为全局感知层来实现的，允许注意力层内与其他视图建立连接。不同视图的键和值相互连接，以促进信息交换。通过在注意力层内跨不同视图共享信息，扩散模型能够感知多视图相关性，并能够生成一致的多视图彩色图像和法线图。

4.2. Cross-Domain Diffusion

4.2. 跨域扩散

Our model is built upon pre-trained stable diffusion models [49] to leverage its strong priors. Given that current 2D diffusion models [33, 49] are designed for a single domain, the main challenge lies in how to effectively extend stable diffusion models to support multi-domain operations.

我们的模型基于预训练的稳定扩散模型 [49] 构建，以利用其强大的先验知识。鉴于当前的二维扩散模型 [33, 49] 是为单域设计的，主要挑战在于如何有效地扩展稳定扩散模型以支持多域操作。

Naive Solutions. To achieve this goal, we explore several possible designs. A straightforward solution is to add four more channels to the output of the UNet module representing the extra domain. Therefore, the diffusion model can simultaneously output normals and color image domains. However, we notice that such a design suffers from low convergence speed and poor generalization. This is because the channel expansion may perturb the pre-trained weights of stable diffusion models and therefore cause catastrophic model forgetting.

简单的解决方案。为了实现这一目标，我们探索了几种可能的设计。一种直接的解决方案是在 UNet 模块的输出中添加四个额外的通道，以表示额外的域。因此，扩散模型可以同时输出法线域和彩色图像域。然而，我们注意到这种设计存在收敛速度慢和泛化能力差的问题。这是因为通道扩展可能会干扰稳定扩散模型的预训练权重，从而导致灾难性的模型遗忘。

Revisiting Eq. 1, it is possible to factor the joint distribution into two conditional distributions:

回顾公式 1，可以将联合分布分解为两个条件分布:

$$q_a\left(\mathbf{z}\right) = q_n\left(n^{1:K}\right) \cdot q_c\left(x^{1:K} \mid n^{1:K}\right). \tag{4}$$

This equation suggests an alternative solution where we could initially train a diffusion model to generate normal maps and then train another diffusion model to produce color images, conditioning on the generated normal maps (or vice versa). Nonetheless, the implementation of this two-stage framework introduces certain complications. It not only substantially increases the computational cost but also results in performance degradation. Please refer to Section 5.6 for an in-depth discussion.

这个方程提出了另一种解决方案，我们可以先训练一个扩散模型来生成法线图，然后在生成的法线图的条件下训练另一个扩散模型来生成彩色图像 (反之亦然)。然而，这种两阶段框架的实现会带来一些复杂性。它不仅会大幅增加计算成本，还会导致性能下降。有关详细讨论请参阅第 5.6 节。

Domain Switcher. To overcome these difficulties mentioned above, we design a cross-domain diffusion scheme via a domain switcher, denoted as s . The switcher s is a one-dimensional vector serving as labels for distinct domains, and we further feed the switcher into the diffusion model as an extra input. Therefore, the formulation of Eq. 2 can be extended as:

域切换器。为了克服上述困难，我们通过一个域切换器设计了一种跨域扩散方案，记为 s 。切换器 s 是一个一维向量，用作不同域的标签，我们进一步将该切换器作为额外输入馈入扩散模型。因此，公式 2 可以扩展为:

$$n^{1:K}, x^{1:K} = f\left(y, \pi_{1:K}, s_n\right), f\left(y, \pi_{1:K}, s_c\right). \tag{5}$$

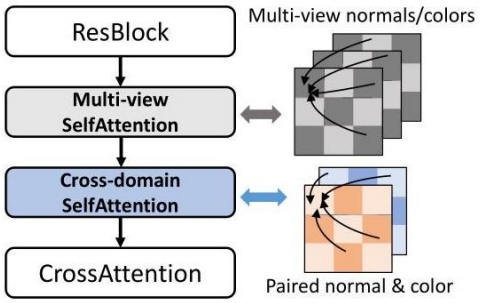

Figure 3. The illustration of the structure of the multi-view cross-domain transformer block.

图 3. 多视图跨域变压器模块结构示意图。

The domain switcher s is first encoded via positional encoding [43] and subsequently concatenated with the time embedding. This combined representation is then injected into the UNet of the stable diffusion models. Interestingly, experiments show that this subtle modification does not significantly alter the pre-trained priors. As a result, it allows for fast convergence and robust generalization, without requiring substantial changes to the stable diffusion models.

域切换器 s 首先通过位置编码 [43] 进行编码，随后与时间嵌入进行拼接。然后将这个组合表示注入到稳定扩散模型的 UNet 中。有趣的是，实验表明这种细微的修改并不会显著改变预训练的先验知识。因此，它能够实现快速收敛和强大的泛化能力，而无需对稳定扩散模型进行重大修改。

Cross-domain Attention. Using the proposed domain switcher, the diffusion model can generate two different domains. However, it is important to note that for a single view, there is no guarantee that the generated color image and the normal map will be geometrically consistent. To address this issue and ensure the consistency between the generated normal maps and color images, we introduce a cross-domain attention mechanism to facilitate the exchange of information between the two domains. This mechanism aims to ensure that the generated outputs align well in terms of geometry and appearance.

跨域注意力机制。利用所提出的域切换器，扩散模型可以生成两个不同的域。然而，需要注意的是，对于单视图而言，无法保证生成的彩色图像和法线图在几何上是一致的。为了解决这个问题并确保生成的法线图和彩色图像之间的一致性，我们引入了一种跨域注意力机制，以促进两个域之间的信息交换。该机制旨在确保生成的输出在几何和外观方面能够很好地对齐。

The cross-domain attention layer maintains the same structure as the original self-attention layer and is integrated before the cross-attention layer in each transformer block of the UNet, as depicted in Figure 3. In the cross-domain attention layer, the keys and values from the normal and color image domains are combined and processed through attention operations. This design ensures that the generations of color images and normal maps are closely correlated, thus promoting geometric consistency between the two domains.

跨域注意力层保持与原始自注意力层相同的结构，并如图 3 所示，集成在 UNet 每个变压器模块的交叉注意力层之前。在跨域注意力层中，法线图和彩色图像域的键和值被组合起来，并通过注意力操作进行处理。这种设计确保了彩色图像和法线图的生成紧密相关，从而促进了两个域之间的几何一致性。

4.3. Textured Mesh Extraction

4.3. 纹理网格提取

To extract explicit 3D geometry from 2D normal maps and color images, we optimize a neural implicit signed distance field (SDF) to amalgamate all 2D generated data. Nonetheless, adopting existing SDF-based reconstruction methods, such as NeuS [64], proves unviable. These methods were tailored for real-captured images and necessitate dense input views. In contrast, our generated views are relatively sparse, and the generated normal maps and color images may exhibit subtle inaccurate predictions of some pixels. Regrettably, these errors accumulate during the geometry optimization, leading to distorted geometries, outliers, and incompleteness. To overcome the challenges above, we propose a novel geometric-aware optimization scheme.

为了从二维法线图和彩色图像中提取显式的三维几何信息，我们优化了一个神经隐式有符号距离场 (SDF)，以融合所有生成的二维数据。然而，采用现有的基于 SDF 的重建方法 (如 NeuS [64]) 是不可行的。这些方法是为真实捕获的图像量身定制的，需要密集的输入视图。相比之下，我们生成的视图相对稀疏，并且生成的法线图和彩色图像可能对某些像素存在细微的不准确预测。遗憾的是，这些误差在几何优化过程中会累积，导致几何形状扭曲、出现离群值和不完整。为了克服上述挑战，我们提出了一种新颖的几何感知优化方案。

Optimization Objectives. With the obtained normal maps $G_{0:N}$ and color images $H_{0:N}$, we first leverage segmentation models to segment the object masks $M_{0:N}$ from the normal maps or color images. Then we optimize SDF field with such an objective function:

优化目标。利用获得的法线图 $G_{0:N}$ 和彩色图像 $H_{0:N}$ ，我们首先利用分割模型从法线图或彩色图像中分割出物体掩码 $M_{0:N}$ 。然后，我们使用这样一个目标函数来优化 SDF 场:

$$\mathcal{L} = \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{mask}} \tag{6}$$

$$+\mathcal{R}_{\text{eik}}+\mathcal{R}_{\text{sparse}}+\mathcal{R}_{\text{smooth}},$$

where $\mathcal{L}_{\text{normal}}$ denotes the normal loss term that will be discussed later, $\mathcal{L}_{rgb}$ denotes a MSE loss term that calculates the errors between rendered colors $\widehat{h}_k$ and generated colors h_k, $\mathcal{L}_{\text{mask}}$ denotes a binary cross-entropy loss term that calculating errors between the rendered mask $\widehat{m}_k$ and the generated mask m_k, $\mathcal{R}_{\text{eik}}$ denotes eikonal regularization term that encourages the magnitude of the SDF gradients to be unit length, $\mathcal{R}_{\text{sparse}}$ denotes a sparsity regularization term that avoid floaters of SDF, and $\mathcal{R}_{\text{smooth}}$ denotes a 3D smoothness regularization term that enforces the SDF gradients to be smooth in 3D space (see details in the supp.).

其中 $\mathcal{L}_{\text{normal}}$ 表示法线损失项 (稍后将进行讨论)，$\mathcal{L}_{rgb}$ 表示均方误差 (MSE) 损失项，用于计算渲染颜色 $\widehat{h}_k$ 和生成颜色 h_k, $\mathcal{L}_{\text{mask}}$ 之间的误差；$\widehat{m}_k$ 表示二元交叉熵损失项，用于计算渲染掩码 $\widehat{m}_k$ 和生成掩码 m_k, $\mathcal{R}_{\text{eik}}$ 之间的误差；$\mathcal{R}_{\text{sparse}}$ 表示 eikonal 正则化项，用于促使 SDF 梯度的模长为单位长度；$\mathcal{R}_{\text{sparse}}$ 表示稀疏性正则化项，用于避免 SDF 出现浮动点；$\mathcal{R}_{\text{smooth}}$ 表示 3D 平滑性正则化项，用于强制 SDF 梯度在三维空间中保持平滑 (详见补充材料)。

Geometry-aware Normal Loss. Thanks to the differentiable nature of SDF representation, we can easily extract normal values $\widehat{g}$ of the optimized SDF via calculating the gradients of SDF. We maximize the similarity of the normal of SDF $\widehat{g}$ and our generated normal g to provide 3D geometric supervision. To tolerate trivial inaccuracies of the generated normals from different views, we introduce a geometry-aware normal loss:

几何感知法线损失。由于 SDF 表示具有可微性，我们可以通过计算 SDF 的梯度轻松提取优化后的 SDF 的法线值 $\widehat{g}$。我们最大化 SDF 的法线 $\widehat{g}$ 和我们生成的法线 g 之间的相似度，以提供 3D 几何监督。为了容忍不同视图生成的法线存在的细微不准确，我们引入了一种几何感知法线损失:

$$\mathcal{L}_{\text{normal}}=\frac{1}{\sum w_k}\sum w_k\cdot\left(1-\cos\left(\widehat{g}_k,g_k\right)\right) \tag{7}$$

where we measure the error between the normal of SDF $\widehat{g}_k$ and the generated normal g_k for the k_{th} sampled ray, $\cos(\cdot,\cdot)$ denotes cosine function, and w_k is a geometric-aware weight defined as

我们在此测量采样光线 k_{th} 的有符号距离函数 (SDF) 法线 $\widehat{g}_k$ 与生成法线 g_k 之间的误差，$\cos(\cdot,\cdot)$ 表示余弦函数，w_k 是一个几何感知权重，定义为

$$w_k=\begin{cases}0, & \cos\left(\mathbf{v}_k,\mathbf{g}_k\right)>\epsilon\\ \exp\left(\left|\cos\left(\mathbf{v}_k,\mathbf{g}_k\right)\right|\right), & \cos\left(\mathbf{v}_k,\mathbf{g}_k\right)\leq\epsilon\end{cases}. \tag{8}$$

Here $\exp(\cdot)$ denotes exponential function, $|\cdot|$ denotes absolute function, ϵ is a negative threshold closing to zero, and we measure the cosine value of the angle between the generated normal g_k and the k_{th} ray' s viewing direction $\mathbf{v}_k$.

这里 $\exp(\cdot)$ 表示指数函数，$|\cdot|$ 表示绝对值函数，ϵ 是一个接近零的负阈值，我们测量生成法线 g_k 与光线 k_{th} 的观察方向 $\mathbf{v}_k$ 之间夹角的余弦值。

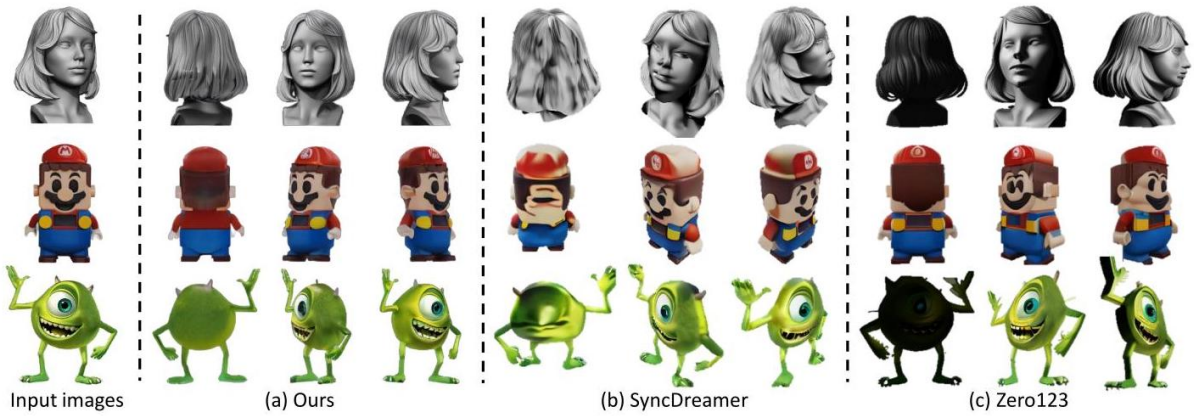

Figure 4. The qualitative comparisons with baseline models on synthesized multi-view color images. We present the picked best results of Zero123 [33] after multiple runs.

图 4. 在合成多视图彩色图像上与基线模型的定性比较。我们展示了 Zero123 [33] 多次运行后挑选出的最佳结果。

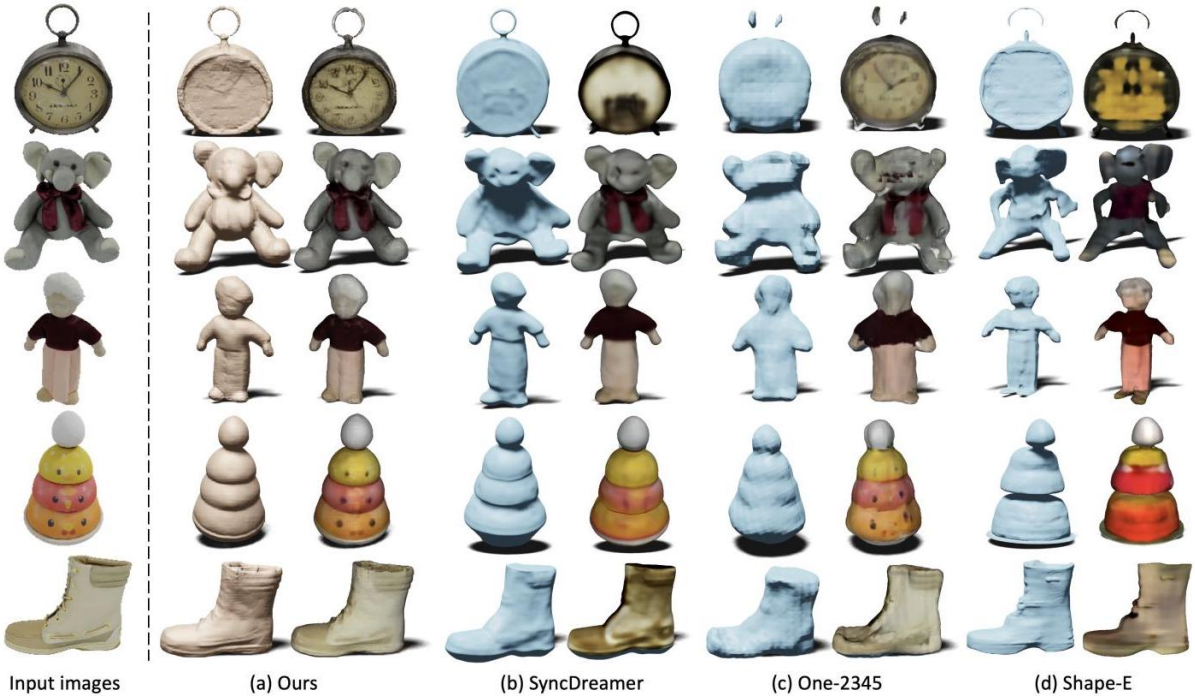

Figure 5. The qualitative comparisons with baseline methods on GSO [14] dataset in terms of the reconstructed textured meshes.

图 5. 在 GSO [14] 数据集上，就重建的纹理网格而言，与基线方法的定性比较。

The design rationale behind this approach lies in the orientation of normals, which are deliberately set to face outward, while the viewing direction is inward-facing. This configuration ensures that the angle between the normal vector and the viewing ray remains not less than $90°$. A deviation from this criterion would imply inaccuracies in the generated normals. Furthermore, rather than treating normals from different views equally, we introduce a

weighting mechanism. We assign higher weights to normals that form larger angles with the viewing rays. This prioritization enhances the accuracy of our geometric supervision process.

这种方法背后的设计原理在于法线的方向，法线被特意设置为朝外，而观察方向是朝内的。这种配置确保了法线向量与观察光线之间的夹角不小于 90° 。若偏离此标准，则意味着生成的法线存在误差。此外，我们没有对不同视图的法线一视同仁，而是引入了一种加权机制。我们为与观察光线形成较大夹角的法线分配更高的权重。这种优先级设置提高了我们几何监督过程的准确性。

Outlier Dropping. Instead of directly summing up the color errors of all sampled rays at each iteration, we first sort these errors in a descending order and then discard the top largest errors according to a predefined percentage. This enables the optimization process to circumvent the outlier regions with poor quality or inconsistency, and instead leverage the regularization capability of MLP to learn the geometry. This strategy effectively eliminates inaccurate isolated surfaces and distorted textures.

异常值剔除。在每次迭代中，我们并非直接对所有采样光线的颜色误差进行求和，而是首先将这些误差按降序排列，然后根据预定义的百分比舍弃最大的误差。这使得优化过程能够避开质量不佳或不一致的异常区域，转而利用多层感知机 (MLP) 的正则化能力来学习几何形状。这种策略能有效消除不准确的孤立表面和扭曲的纹理。

5. Experiments

5. 实验

5.1. Implementation Details

5.1. 实现细节

We train our model on the LVIS subset of the Objaverse dataset [10], which comprises approximately 30k+ objects following a cleanup process. To create the rendered multiview dataset, we first normalized each object to be centered and of unit scale. Then we render normal maps and color images from six views, including the front, back, left, right, front-right, and front-left views, using Blenderproc [12].

我们在 Objaverse 数据集 [10] 的 LVIS 子集中训练我们的模型，该子集在清理后大约包含 30k+ 个对象。为了创建渲染的多视图数据集，我们首先将每个对象进行归一化处理，使其居中且具有单位尺度。然后，我们使用 Blenderproc [12] 从六个视图 (包括前、后、左、右、右前和左前视图) 渲染法线贴图和彩色图像。

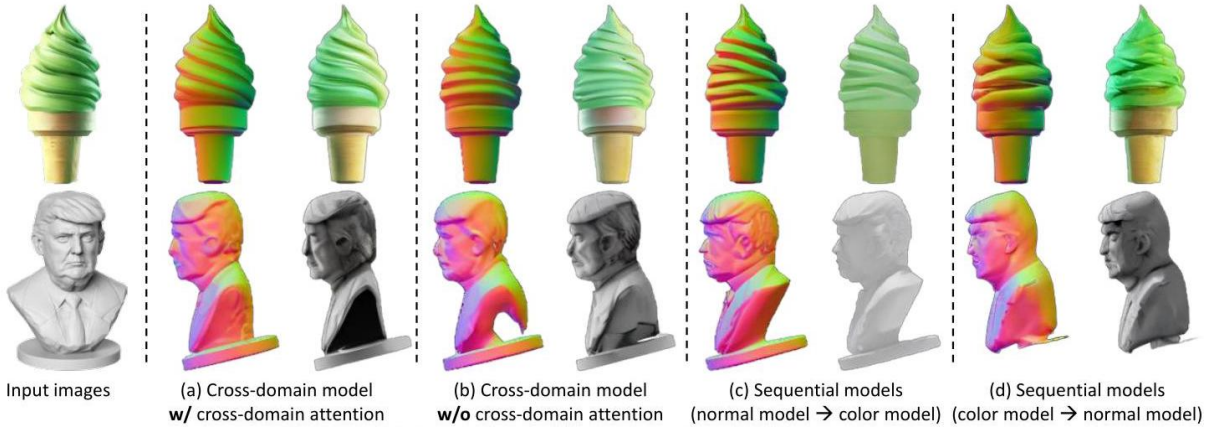

Figure 6. Ablation studies on different cross-domain diffusion schemes.

图 6. 不同跨域扩散方案的消融研究。

We fine-tune our model starting from the Stable Diffusion Image Variations Model, which has previously been fine-tuned with image conditions. During fine-tuning, we use a reduced image size of 256×256 and a total batch size of 512 for training. The fine-tuning process involves training the model for 30,000 steps. This entire training procedure typically requires approximately 3 days on a cluster of 8 Nvidia Tesla A800 GPUs. To reconstruct 3D geometry from the 2D representations, our normal fusion method is built on the instant-NGP SDF reconstruction method [19].

我们从 Stable Diffusion 图像变体模型开始微调我们的模型，该模型此前已根据图像条件进行过微调。在微调过程中，我们使用缩小后的图像尺寸 256×256，并以 512 的总批量大小进行训练。微调过程包括对模型进行 30,000 步的训练。在由 8 块英伟达特斯拉 A800 GPU 组成的集群上，整个训练过程通常大约需要 3 天。为了从二维表示中重建三维几何形状，我们的法线融合方法基于即时神经图形原语 (instant-NGP) 的有符号距离函数 (SDF) 重建方法 [19] 构建。

5.2. Baselines

5.2. 基线方法

We adopt Zero123 [33], RealFusion [42], Magic123 [48], One-2-3-45 [32], Point-E [45], Shap-E [26] and a recent work SyncDreamer [36] as baseline methods. Given an input image, Zero123 [33] is capable of generating novel views of arbitrary viewpoints, and it can be incorporated with SDS loss [47] for 3D reconstruction (we adopt the implementation of ThreeStudio [20]). RealFusion [42] and Magic123 [48] leverage Stable Diffusion [51] and SDS loss for single-view reconstruction. One-2-3-45 [32] directly predict SDFs via SparseNeuS [39] by taking the generated multiple images of Zero123 [33]. Point-E [45] and Shap-E [26] are 3D generative models trained on a large internal OpenAI 3D dataset, both of which are able to convert a single-view image into a point cloud or an implicit representation. SyncDreamer[36] generates multi-view consistent images from a single image for deriving 3D geometry.

我们采用 Zero123 [33]、RealFusion [42]、Magic123 [48]、One-2-3-45 [32]、Point-E [45]、Shap-E [26] 和近期的工作 SyncDreamer [36] 作为基线方法。给定一张输入图像，Zero123 [33] 能够生成任意视点的新视图，并且可以结合稳定扩散监督 (SDS) 损失 [47] 进行三维重建 (我们采用 ThreeStudio [20] 的实现)。RealFusion [42] 和 Magic123 [48] 利用 Stable Diffusion [51] 和 SDS 损失进行单视图重建。One-2-3-45 [32] 通过稀疏神经表面 (SparseNeuS)[39] 直接根据 Zero123 [33] 生成的多张图像预测有符号距离函数 (SDF)。Point-E [45] 和 Shap-E [26] 是在 OpenAI 内部大型三维数据集上训练的三维生成模型，它们都能够将单视图图像转换为点云或隐式表示。SyncDreamer [36] 从单张图像生成多视图一致的图像，以推导三维几何形状。

5.3. Evaluation Protocol

5.3. 评估协议

Evaluation Datasets. Following prior research [33, 36] , we adopt the Google Scanned Object dataset [14] for our evaluation, which includes a wide variety of common everyday objects. Our evaluation dataset matches that of Sync-Dreamer [36], comprising 30 objects that span from everyday items to animals. For each object in the evaluation set, we render an image with a size of 256×256 , which serves as the input. Additionally, we include some images with diverse styles collected from the internet or text2image models in our evaluation.

评估数据集。遵循先前的研究 [33, 36] ，我们采用谷歌扫描物体数据集 [14] 进行评估，该数据集包含各种各样常见的日常物品。我们的评估数据集与 Sync-Dreamer [36] 的数据集相匹配，包含 30 种物体，从日常用品到动物都有涉及。对于评估集中的每个物体，我们渲染一张尺寸为 256×256 的图像作为输入。此外，我们还在评估中纳入了一些从互联网或文本到图像模型收集的不同风格的图像。

Method	Chamfer Dist.↓	Volume IoU↑
Realfusion [42]	0.0819	0.2741
Magic123 [48]	0.0516	0.4528
One-2-3-45 [32]	0.0629	0.4086
Point-E [45]	0.0426	0.2875
Shap-E [26]	0.0436	0.3584
Zero123 [33]	0.0339	0.5035
SyncDreamer [36]	0.0261	0.5421
Ours	0.0199	0.6244

方法	倒角距离 (Chamfer Dist.)↓	体积交并比 (Volume IoU)↑
Realfusion [42]	0.0819	0.2741
Magic123 [48]	0.0516	0.4528
One-2-3-45 [32]	0.0629	0.4086
Point-E [45]	0.0426	0.2875
Shap-E [26]	0.0436	0.3584
Zero123 [33]	0.0339	0.5035
SyncDreamer [36]	0.0261	0.5421
我们的方法	0.0199	0.6244

Table 1. Quantitative comparison with baseline methods. We report Chamfer Distance and Volume IoU on the GSO [14] dataset.

表 1. 与基线方法的定量比较。我们报告了在 GSO [14] 数据集上的倒角距离 (Chamfer Distance) 和体积交并比 (Volume IoU)。

Metrics. To evaluate the quality of the single-view reconstructions, we adopt two metrics: Chamfer Distances (CD) and Volume IoU between ground-truth shapes and reconstructed shapes. Since different methods adopt various canonical systems, we first align the generated shapes to the ground-truth shapes before calculating the two metrics. Moreover, we adopt the metrics PSNR, SSIM [66] and LPIPS [78] for evaluating the generated color images.

指标。为了评估单视图重建的质量，我们采用了两个指标: 真实形状与重建形状之间的倒角距离 (Chamfer Distances，CD) 和体积交并比 (Volume IoU)。由于不同的方法采用了不同的规范系统，我们在计算这两个指标之前，首先将生成的形状与真实形状进行对齐。此外，我们采用峰值信噪比 (PSNR)、结构相似性指数 (SSIM [66]) 和学习感知图像块相似度 (LPIPS [78]) 指标来评估生成的彩色图像。

5.4. Single View Reconstruction

5.4. 单视图重建

We evaluate the quality of the reconstructions of different methods. The quantitative results are summarized in Table 1, and the qualitative comparisons are presented in Fig. 5. Shap-E [26] tends to produce incomplete and distorted meshes. SyncDreamer [36] generates shapes that are roughly aligned with the input image but lack detailed geometries, and the texture quality is subpar. One-2-3- 45 [32] attempts to reconstruct meshes from the multiview-inconsistent outputs of Zero123 [33]. While it can capture coarse geometries, it loses important details in the process. In comparison, our method stands out by achieving the highest quality, both in terms of geometry and textures.

我们评估了不同方法的重建质量。定量结果总结在表 1 中，定性比较展示在图 5 中。Shap - E [26] 倾向于生成不完整且扭曲的网格。SyncDreamer [36] 生成的形状大致与输入图像对齐，但缺乏详细的几何信息，并且纹理质量较差。One - 2 - 3 - 45 [32] 尝试从 Zero123 [33] 的多视图不一致输出中重建网格。虽然它可以捕捉到粗略的几何形状，但在此过程中丢失了重要的细节。相比之下，我们的方法在几何形状和纹理方面都达到了最高质量，脱颖而出。

Figure 7. Ablation study on the strategies in the mesh extraction module: geometry-aware normal loss and outlier-dropping strategy.

图 7. 网格提取模块中策略的消融研究: 几何感知法线损失和离群点剔除策略。

Method	PSNR↑	SSIM↑	LPIPS↓
Realfusion [42]	15.26	0.722	0.283
Zero123 [33]	18.93	0.779	0.166
SyncDreamer [36]	20.05	0.798	0.146
Ours	26.07	0.924	0.065

方法	峰值信噪比 (PSNR)↑	结构相似性指数 (SSIM)↑	学习感知图像块相似度 (LPIPS)↓
真实融合法 (Realfusion) [42]	15.26	0.722	0.283
零一到三法 (Zero123) [33]	18.93	0.779	0.166
同步梦想家法 (SyncDreamer) [36]	20.05	0.798	0.146
我们的方法	26.07	0.924	0.065

Table 2. The quantitative comparison in novel view synthesis. We report PSNR, SSIM [66], LPIPS [78] on the GSO [14] dataset.

表 2. 新视角合成的定量比较。我们报告了 GSO [14] 数据集上的峰值信噪比 (PSNR)、结构相似性指数 (SSIM)[66] 和学习感知图像块相似度 (LPIPS)[78]。

5.5. Novel View Synthesis

5.5 新视角合成

We evaluate the quality of novel view synthesis for different methods. The quantitative results are presented in Table 2, and the qualitative results can be found in Figure 4. Zero123 [33] produces visually reasonable images, but they lack multi-view consistency since it operates on each view independently. Although SyncDreamer [33]

introduces a volume attention scheme to enhance the consistency of multi-view images, their model is sensitive to the elevation degrees of the input images and tends to produce unreasonable results. In contrast, our method generates images that not only exhibit semantic consistency with the input image but also maintain high consistency across multiple views in terms of both colors and geometry.

我们评估了不同方法的新视角合成质量。定量结果见表 2，定性结果见图 4。Zero123 [33] 生成的图像在视觉上合理，但由于它独立处理每个视角，因此缺乏多视角一致性。尽管 SyncDreamer [33] 引入了体注意力机制来增强多视角图像的一致性，但其模型对输入图像的仰角敏感，容易产生不合理的结果。相比之下，我们的方法生成的图像不仅与输入图像具有语义一致性，而且在颜色和几何形状方面的多视角之间也保持高度一致。

5.6. Discussions

5.6 讨论

In this section, we perform a series of experiments to validate the effectiveness of the designs in our method.

在本节中，我们进行了一系列实验，以验证我们方法中设计的有效性。

Cross-Domain Diffusion. To validate the effectiveness of our proposed cross-domain diffusion scheme, we study the following settings: (a) cross-domain model with cross-domain attention; (b) cross-domain model without cross-domain attention; (c) sequential models rgb-to-normal: first train a multi-view color diffusion model then train a multiview normal diffusion model conditioned on the previously generated color images; (d) sequential models normal-to-rgb: first train a multi-view normal diffusion model then train a multi-view color diffusion model conditioned on the previously generated normal images.

跨域扩散。为了验证我们提出的跨域扩散方案的有效性，我们研究了以下设置:(a) 具有跨域注意力的跨域模型；(b) 没有跨域注意力的跨域模型；(c) 顺序模型 (RGB 到法线): 先训练一个多视角颜色扩散模型，然后基于先前生成的彩色图像训练一个多视角法线扩散模型；(d) 顺序模型 (法线到 RGB): 先训练一个多视角法线扩散模型，然后基于先前生成的法线图像训练一个多视角颜色扩散模型。

As shown in (a) and (b) of Figure 6, it's evident that the cross-domain attentions significantly enhance the consistency between color images and normals, particularly in terms of the detailed geometries of objects like the ice-cream and the sculpture. From (c) and (d) of Figure 6, while the normals and color images generated by sequential models maintain some consistency, their results suffer from performance drops, like color shifts of (c) and wrong geometries of (d).

如图 6 (a) 和 (b) 所示，很明显，跨域注意力显著增强了彩色图像和法线之间的一致性，特别是在冰淇淋和雕塑等物体的详细几何形状方面。从图 6 (c) 和 (d) 可以看出，虽然顺序模型生成的法线和彩色图像保持了一定的一致性，但它们的结果存在性能下降的问题，如 (c) 中的颜色偏移和 (d) 中的错误几何形状。

Normal Fusion. To assess the efficacy of our normal fusion algorithm, we conducted experiments using the complex lion model, which is rich in geometric details, as illustrated in Figure 7. The baseline model's surfaces exhibited numerous holes and noises. Utilizing either the geometry-aware normal loss or the outlier-dropping loss

helps mitigate the noisy surfaces. Finally, combining both strategies yields the best performance, resulting in clean surfaces while preserving detailed geometries.

法线融合。为了评估我们的法线融合算法的有效性，我们使用了具有丰富几何细节的复杂狮子模型进行实验，如图 7 所示。基线模型的表面有许多孔洞和噪声。使用几何感知法线损失或异常值去除损失有助于减少表面噪声。最后，结合这两种策略可获得最佳性能，在保留详细几何形状的同时使表面干净。

6. Conclusions and Future Works

6 结论与未来工作

In this paper, we present Wonder3D, an innovative approach designed for efficiently generating high-fidelity textured meshes from single-view images. Experimental results demonstrate that our method upholds good efficiency and robust generalization, and delivers high-quality geometry. However, limited by computational resources, the current implementation of Wonder3D only produces normals and color images from six views. This limited number of views makes it challenging for our method to accurately reconstruct objects with very thin structures and severe occlusions. To address this issue, Wonder3D may benefit from leveraging more efficient multi-view attention mechanisms to handle a greater number of views effectively.

在本文中，我们提出了 Wonder3D，这是一种创新方法，旨在从单视角图像高效生成高保真纹理网格。实验结果表明，我们的方法保持了良好的效率和强大的泛化能力，并能提供高质量的几何形状。然而，受计算资源的限制，Wonder3D 目前的实现仅能从六个视角生成法线和彩色图像。这种有限的视角数量使得我们的方法难以准确重建具有非常薄的结构和严重遮挡的物体。为了解决这个问题，Wonder3D 可以利用更高效的多视角注意力机制来有效处理更多的视角。

Acknowledgements

致谢

This research is partially supported by the Innovation and Technology Commission of the HKSAR Government under the InnoHK initiative and Ref. T45-205/21-N of Hong Kong RGC. Song-Hai Zhang is supported by the National Key Research and Development Program of China (No. 2023YFF0905104), the Natural Science Foundation of China (No. 62132012), and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

本研究部分得到了香港特别行政区政府创新科技署 InnoHK 计划以及香港研究资助局参考编号 T45 - 205/21 - N 的支持。张松海得到了中国国家重点研发计划 (编号 2023YFF0905104)、中国自然科学基金 (编号 62132012) 以及清华 - 腾讯互联网创新技术联合实验室的支持。

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Ren-derdiffusion: Image diffusion for 3 d reconstruction, inpaint-ing and generation. In $CVPR$, 2023.2, 3

Titas Anciukevičius、Zexiang Xu、Matthew Fisher、Paul Henderson、Hakan Bilen、Niloy J Mitra 和 Paul Guerrero。Renderdiffusion: 用于 3 d 重建、修复和生成的图像扩散。见 $CVPR$, 2023.2, 3

[2] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Reimagine the negative prompt algorithm: Transform 2d diffusion into 3 d , alleviate janus problem and beyond. arXiv preprint arXiv:2304.04968, 2023. 3

Mohammadreza Armandpour、Huangjie Zheng、Ali Sadeghian、Amir Sadeghian 和 Mingyuan Zhou。重新构想负提示算法: 将二维扩散转换为 3 d，缓解双面问题及其他。预印本 arXiv:2304.04968，2023 年。3

[3] 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. Generative novel view synthesis with 3d-aware diffusion models. In $ICCV$, 2023.3

Eric R Chan、Koki Nagano、Matthew A Chan、Alexander W Bergman、Jeong Joon Park、Axel Levy、Miika Aittala、Shalini De Mello、Tero Karras 和 Gordon Wetzstein。使用 3D 感知扩散模型进行生成式新视角合成。见 $ICCV$, 2023.3

[4] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3 d generation and reconstruction. In $ICCV$, 2023.2, 3

陈汉生 (Hansheng Chen)、顾佳涛 (Jiatao Gu)、陈安沛 (Anpei Chen)、田伟 (Wei Tian)、涂卓文 (Zhuowen Tu)、刘灵杰 (Lingjie Liu) 和苏浩 (Hao Su)。单阶段扩散神经辐射场 (NeRF): 一种统一的 3 d 生成与重建方法。见 $ICCV$, 2023.2, 3

[5] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873, 2023. 2, 3

陈锐 (Rui Chen)、陈永伟 (Yongwei Chen)、焦宁欣 (Ningxin Jiao) 和贾奎 (Kui Jia)。Fantasia3D: 解耦几何与外观以实现高质量文本到三维内容创作。预印本 arXiv:2303.13873，2023 年。2, 3

[6] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. arXiv preprint arXiv:2308.11473, 2023. 3

陈艺文 (Yiwen Chen)、张弛 (Chi Zhang)、杨晓峰 (Xiaofeng Yang)、蔡中昂 (Zhongang Cai)、余刚 (Gang Yu)、杨磊 (Lei Yang) 和林国生 (Guosheng Lin)。It3D: 通过显式视图合成改进文本到三维生成。预印本 arXiv:2308.11473，2023 年。3

[7] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. arXiv preprint arXiv:2402.14650, 2024. 2

程凯 (Kai Cheng)、龙笑笑 (Xiaoxiao Long)、杨开智 (Kaizhi Yang)、姚瑶 (Yao Yao)、尹伟 (Wei Yin)、马悦欣 (Yuexin Ma)、王文平 (Wenping Wang) 和陈学进 (Xuejin Chen)。Gaussianpro: 具有渐进传播的三维高斯 splatting。预印本 arXiv:2402.14650，2024 年。2

[8] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3 d shape completion, reconstruction, and generation. In CVPR, 2023. 2, 3

郑彦吉 (Yen - Chi Cheng)、李欣莹 (Hsin - Ying Lee)、谢尔盖·图利亚科夫 (Sergey Tulyakov)、亚历山大·G·施温 (Alexander G Schwing) 和桂良彦 (Liang - Yan Gui)。Sdfusion: 多模态 3 d 形状补全、重建与生成。见计算机视觉与模式识别会议 (CVPR)，2023 年。2, 3

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. T-PAMI, 2023. 3

弗洛里内尔 - 阿林·克罗伊托鲁 (Florinel - Alin Croitoru)、弗拉德·洪德鲁 (Vlad Hondru)、拉杜·图多尔·约内斯库 (Radu Tudor Ionescu) 和穆巴拉克·沙阿 (Mubarak Shah)。视觉中的扩散模型: 综述。《模式分析与机器智能汇刊》(T - PAMI)，2023 年。3

[10] 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 $CVPR$, 2023.6

马特·戴特克 (Matt Deitke)、达斯汀·施文克 (Dustin Schwenk)、霍尔迪·萨尔瓦多 (Jordi Salvador)、卢卡·魏斯 (Luca Weihs)、奥斯卡·米歇尔 (Oscar Michel)、伊莱·范德比尔 (Eli VanderBilt)、路德维希·施密特 (Ludwig Schmidt)、基安娜·埃斯哈尼 (Kiana Ehsani)、阿尼尔 uddha·肯巴维 (Aniruddha Kembhavi) 和阿里·法尔哈迪 (Ali Farhadi)。Objaverse: 一个带注释的三维物体宇宙。见 $CVPR$, 2023.6

[11] 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 $CVPR$, 2023.3

邓聪悦 (Congyue Deng)、蒋驰宇 (Chiyu Jiang)、查尔斯·R·齐 (Charles R Qi)、闫鑫晨 (Xinchen Yan)、周银 (Yin Zhou)、莱昂尼达斯·吉巴斯 (Leonidas Guibas)、德拉戈米尔·安格洛夫 (Dragomir Anguelov) 等。Nerdi: 以语言引导的扩散作为通用图像先验的单视图神经辐射场合成。见 $CVPR$, 2023.3

[12] Maximilian Denninger, Dominik Winkelbauer, Martin Sun-dermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. Journal of Open Source Software, 8(82):4901, 2023. 7

马克西米利安·登宁格 (Maximilian Denninger)、多米尼克·温克尔鲍尔 (Dominik Winkelbauer)、马丁·桑德迈尔 (Martin Sundermeyer)、沃特·博尔德伊克 (Wout Boerdijk)、马库斯·克瑙尔 (Markus Knauer)、克劳斯·H·斯特罗布 (Klaus H. Strobl)、马蒂亚斯·胡姆特 (Matthias Humt) 和鲁道夫·特里贝尔 (Rudolph Triebel)。Blenderproc2: 用于逼真渲染的程序化管道。《开源软件杂志》，8(82):4901，2023 年。7

[13] Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping Wang. Tore: Token reduction for efficient human mesh recovery with transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15143- 15155, 2023. 2

窦智洋 (Zhiyang Dou)、吴清璇 (Qingxuan Wu)、林程 (Cheng Lin)、曹泽宇 (Zeyu Cao)、吴强强 (Qiangqiang Wu)、万伟霖 (Weilin Wan)、小村拓 (Taku Komura) 和王文平 (Wenping Wang)。Tore: 基于 Transformer 的高效人体网格恢复的令牌缩减方法。见电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集，第 15143 - 15155 页，2023 年。2

[14] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kin-man, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3 d scanned household items. In $ICRA$, 2022.3, 6, 7, 8

劳拉·唐斯 (Laura Downs)、安东尼·弗朗西斯 (Anthony Francis)、内特·柯尼格 (Nate Koenig)、布兰登·金曼 (Brandon Kinman)、瑞安·希克曼 (Ryan Hickman)、克里斯塔·雷曼 (Krista Reymann)、托马斯·B·麦克休 (Thomas B McHugh) 和文森特·范霍克 (Vincent Vanhoucke)。谷歌扫描物体: 一个高质量的 3 d 扫描家居用品数据集。见 $ICRA$ ，2022 年。3, 6, 7, 8

[15] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. arXiv preprint arXiv:2303.17015, 2023. 2, 3

齐亚·埃尔科奇 (Ziya Erkoç)、马方昌 (Fangchang Ma)、单琦 (Qi Shan)、马蒂亚斯·尼斯纳 (Matthias Nießner) 和安吉拉·戴 (Angela Dai)。Hyperdiffusion: 通过权重空间扩散生成隐式神经场。预印本 arXiv:2303.17015，2023 年。2, 3

[16] 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 3 d textured shapes learned from images. NeurIPS, 2022. 2, 3

高军 (Jun Gao)、沈天畅 (Tianchang Shen)、王梓安 (Zian Wang)、陈文正 (Wenzheng Chen)、尹康雪 (Kangxue Yin)、李戴清 (Daiqing Li)、奥尔·利塔尼 (Or Litany)、赞·戈伊契奇 (Zan Gojcic) 和桑贾·菲德勒 (Sanja Fidler)。Get3D: 从图像中学习的高质量 3 d 纹理形状生成模型。神经信息处理系统大会 (NeurIPS)，2022 年。2, 3

[17] Jiatao Gu, Qingzhe Gao, Shuangfei Zhai, Baoquan Chen, Lingjie Liu, and Josh Susskind. Learning controllable 3d diffusion models from single-view images. arXiv preprint arXiv:2304.06700, 2023. 2, 3

顾佳涛 (Jiatao Gu)、高庆哲 (Qingzhe Gao)、翟双飞 (Shuangfei Zhai)、陈宝权 (Baoquan Chen)、刘凌杰 (Lingjie Liu) 和约什·萨斯金德 (Josh Susskind)。从单视图图像中学习可控的 3D 扩散模型。预印本 arXiv:2304.06700，2023 年。2, 3

[18] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In $ICML$, 2023.3

顾佳涛 (Jiatao Gu)、亚历克斯·特里维西克 (Alex Trevithick)、林恺恩 (Kai-En Lin)、约书亚·M·萨斯金德 (Joshua M Susskind)、克里斯蒂安·特奥巴尔 (Christian Theobalt)、刘凌杰 (Lingjie Liu) 和拉维·拉马穆尔蒂 (Ravi Ramamoorthi)。Nerfdiff: 通过从 3D 感知扩散中进行神经辐射场 (NeRF) 引导的蒸馏实现单图像视图合成。见 $ICML$, 2023.3

[19] Yuan-Chen Guo. Instant neural surface reconstruction, 2022. https://github.com/bennyguo/instant-nsr-pl.7

郭元晨 (Yuan-Chen Guo)。即时神经表面重建，2022 年。https://github.com/bennyguo/instant-nsr-pl.7

[20] 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 3 d content generation. https://github.com/threestudio-project/ threestudio,2023. 7

郭元晨 (Yuan-Chen Guo)、刘英田 (Ying-Tian Liu)、邵睿智 (Ruizhi Shao)、克里斯蒂安·拉福尔特 (Christian Laforte)、维克拉姆·沃莱蒂 (Vikram Voleti)、罗冠 (Guan Luo)、陈家豪 (Chia-Hao Chen)、邹子鑫 (Zi-Xin Zou)、王晨 (Chen Wang)、曹彦培 (Yan-Pei Cao) 和张松海 (Song-Hai Zhang)。threestudio: 用于 3 d 内容生成的统一框架。https://github.com/threestudio-project/ threestudio，2023 年。7

[21] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Bar-las Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. arXiv preprint arXiv:2303.05371, 2023. 2, 3

安奇特·古普塔 (Anchit Gupta)、熊文瀚 (Wenhan Xiong)、聂一新 (Yixin Nie)、伊恩·琼斯 (Ian Jones) 和巴拉斯·奥古兹 (Bar-las Oğuz)。3dgen: 用于纹理网格生成的三平面潜在扩散。预印本 arXiv:2303.05371，2023 年。2, 3

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020. 3, 4

乔纳森·霍 (Jonathan Ho)、阿贾伊·贾恩 (Ajay Jain) 和彼得·阿贝贝尔 (Pieter Abbeel)。去噪扩散概率模型。见神经信息处理系统大会 (NeurIPS)，2020 年。3, 4

[23] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. arXiv preprint arXiv:2306.12422, 2023. 3

黄玉坤 (Yukun Huang)、王佳楠 (Jianan Wang)、史宇凯 (Yukai Shi)、齐贤彪 (Xianbiao Qi)、查正军 (Zheng-Jun Zha) 和张磊 (Lei Zhang)。Dreamtime: 用于文本到 3D 内容创建的改进优化策略。预印本 arXiv:2306.12422，2023 年。3

[24] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 867-876, 2022. 2

阿贾伊·贾恩 (Ajay Jain)、本·米尔登霍尔 (Ben Mildenhall)、乔纳森·T·巴伦 (Jonathan T Barron)、彼得·阿贝贝尔 (Pieter Abbeel) 和本·普尔 (Ben Poole)。通过梦境场实现零样本文本引导的物体生成。见电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议 (IEEE/CVF Conference on Computer Vision and Pattern Recognition) 论文集，第 867 - 876 页，2022 年。2

[25] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaox-iao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3 d gaussian splatting with shading functions for reflective surfaces. arXiv preprint arXiv:2311.17977, 2023. 2

蒋英文琪 (Yingwenqi Jiang)、涂佳东 (Jiadong Tu)、刘源 (Yuan Liu)、高希峰 (Xifeng Gao)、龙晓晓 (Xiaox-iao Long)、王文平 (Wenping Wang) 和马跃新 (Yuexin Ma)。Gaussianshader: 用于反射表面的带有着色函数的 3 d 高斯散点法。预印本 arXiv:2311.17977，2023 年。2

[26] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3 d implicit functions. arXiv preprint arXiv:2305.02463, 2023. 2, 3, 7

许宇佑 (Heewoo Jun) 和亚历克斯·尼科尔 (Alex Nichol)。Shap-e: 生成条件 3 d 隐式函数。预印本 arXiv:2305.02463，2023 年。2, 3, 7

[27] Animesh Karnewar, Niloy J Mitra, Andrea Vedaldi, and David Novotny. Holofusion: Towards photo-realistic 3d generative modeling. In $ICCV$, 2023 .

阿尼梅什·卡尔内瓦尔 (Animesh Karnewar)、尼洛伊·J·米特拉 (Niloy J Mitra)、安德里亚·韦尔达利 (Andrea Vedaldi) 和大卫·诺沃特尼 (David Novotny)。Holofusion: 迈向照片级真实感 3D 生成式建模。见 $ICCV$, 2023 。

[28] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In CVPR, 2023.2,3

金承旭 (Seung Wook Kim)、布拉德利·布朗 (Bradley Brown)、尹康学 (Kangxue Yin)、卡斯滕·克雷斯 (Karsten Kreis)、卡特娅·施瓦茨 (Katja Schwarz)、李戴清 (Daiqing Li)、罗宾·隆巴赫 (Robin Rombach)、安东尼奥·托拉尔巴 (Antonio Torralba) 和桑贾·菲德勒 (Sanja Fidler)。Neuralfield-ldm: 使用分层潜在扩散模型进行场景生成。见计算机视觉与模式识别会议 (CVPR)，2023 年。2, 3

[29] Kilian Kleeberger, Richard Bormann, Werner Kraus, and Marco F Huber. A survey on learning-based robotic grasping. Current Robotics Reports, 1:239-249, 2020. 2

基利安·克莱伯格 (Kilian Kleeberger)、理查德·博尔曼 (Richard Bormann)、维尔纳·克劳斯 (Werner Kraus) 和马尔科·F·胡贝尔 (Marco F Huber)。基于学习的机器人抓取研究综述。《当前机器人报告》，1:239 - 249，2020 年。2

[30] Jiabao Lei, Jiapeng Tang, and Kui Jia. Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In $CVPR$, 2022.3

雷家宝 (Jiabao Lei)、唐佳鹏 (Jiapeng Tang) 和贾奎 (Kui Jia)。通过使用 RGB - D 扩散模型的增量视图修复进行生成式场景合成。见 $CVPR$, 2022.3

[31] 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 $CVPR$, 2023.2, 3

林陈轩 (Chen-Hsuan Lin)、高军 (Jun Gao)、唐鲁明 (Luming Tang)、托瓦基·高川 (Towaki Takikawa)、曾晓辉 (Xiaohui Zeng)、黄勋 (Xun Huang)、卡斯滕·克雷斯 (Karsten Kreis)、桑贾·菲德勒 (Sanja Fidler)、刘明宇 (Ming-Yu Liu) 和林宗毅 (Tsung-Yi Lin)。Magic3d: 高分辨率文本到 3D 内容创建。见 $CVPR$, 2023.2, 3

[32] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu , and Hao Su. One-2-3-45: Any single image to 3 d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928, 2023. 7

刘明华 (Minghua Liu)、徐超 (Chao Xu)、金海安 (Haian Jin)、陈凌浩 (Linghao Chen)、泽祥 Xu 以及苏浩 (Hao Su)。One-2-3-45: 无需逐形状优化，在 45 秒内将任意单张图像转换为 3 d 网格。预印本 arXiv:2306.16928，2023 年 7 月。

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tok-makov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In $ICCV$, 2023.2, 4, 6, 7 , 8

刘若诗 (Ruoshi Liu)、吴润迪 (Rundi Wu)、巴塞尔·范·胡里克 (Basile Van Hoorick)、帕维尔·托卡科夫 (Pavel Tok-makov)、谢尔盖·扎哈罗夫 (Sergey Zakharov) 以及卡尔·冯德里克 (Carl Vondrick)。Zero-1-to-3: 零样本单图像生成 3D 对象。见 $ICCV$, 2023.2, 4, 6, 7 ， 8

[34] Xinhang Liu, Shiu-hong Kao, Jiaben Chen, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf: Enhancing nerf reconstruction using pseudo-observations from diffusion models. arXiv preprint arXiv:2305.15171, 2023. 3

刘新航 (Xinhang Liu)、高秀虹 (Shiu-hong Kao)、陈家本 (Jiaben Chen)、戴宇荣 (Yu-Wing Tai) 以及唐启康 (Chi-Keung Tang)。Deceptive-nerf: 利用扩散模型的伪观测增强神经辐射场 (NeRF) 重建。预印本 arXiv:2305.15171，2023 年 3 月。

[35] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7824-7833, 2022. 2

刘渊 (Yuan Liu)、彭思达 (Sida Peng)、刘凌杰 (Lingjie Liu)、王倩倩 (Qianqian Wang)、王鹏 (Peng Wang)、克里斯蒂安·特奥博尔特 (Christian Theobalt)、周小伟 (Xiaowei Zhou) 以及王文平 (Wenping Wang)。用于遮挡感知的基于图像渲染的神经光线。见《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 7824 - 7833 页，2022 年 2 月。

[36] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453, 2023. 2, 3, 4, 7, 8

刘渊 (Yuan Liu)、林程 (Cheng Lin)、曾子娇 (Zijiao Zeng)、龙笑笑 (Xiaoxiao Long)、刘凌杰 (Lingjie Liu)、小村拓 (Taku Komura) 以及王文平 (Wenping Wang)。Syncdreamer: 从单视图图像生成多视图一致的图像。预印本 arXiv:2309.03453，2023 年 2 月、3 月、4 月、7 月、8 月。

[37] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. In ICLR, 2023.2,3

刘震 (Zhen Liu)、冯瑶 (Yao Feng)、迈克尔·J·布莱克 (Michael J Black)、德里克·诺鲁泽扎赫赖 (Derek Nowrouzezahrai)、利亚姆·保尔 (Liam Paull) 以及刘炜阳 (Weiyang Liu)。Meshdiffu-sion: 基于分数的生成式 3D 网格建模。见国际学习表征会议 (ICLR)，2023 年 2 月、3 月。

[38] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 12849-12858, 2021. 2

龙笑笑 (Xiaoxiao Long)、林程 (Cheng Lin)、刘凌杰 (Lingjie Liu)、李伟 (Wei Li)、克里斯蒂安·特奥博尔特 (Christian Theobalt)、杨瑞刚 (Ruigang Yang) 以及王文平 (Wenping Wang)。用于深度估计的自适应表面法线约束。见《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议 (ICCV) 论文集》，第 12849 - 12858 页，2021 年 2 月。

[39] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In European Conference on Computer Vision, pages 210-227. Springer, 2022. 7

龙笑笑 (Xiaoxiao Long)、林程 (Cheng Lin)、王鹏 (Peng Wang)、小村拓 (Taku Komura) 以及王文平 (Wenping Wang)。Sparseneus: 从稀疏视图进行快速可泛化的神经表面重建。见《欧洲计算机视觉会议论文集》，第 210 - 227 页。施普林格出版社，2022 年 7 月。

[40] Xiaoxiao Long, Yuhang Zheng, Yupeng Zheng, Beiwen Tian, Cheng Lin, Lingjie Liu, Hao Zhao, Guyue Zhou, and Wenping Wang. Adaptive surface normal constraint for geometric estimation from monocular images. arXiv preprint arXiv:2402.05869, 2024. 2

龙笑笑 (Xiaoxiao Long)、郑宇航 (Yuhang Zheng)、郑宇鹏 (Yupeng Zheng)、田贝文 (Beiwen Tian)、林程 (Cheng Lin)、刘凌杰 (Lingjie Liu)、赵浩 (Hao Zhao)、周谷雨 (Guyue Zhou) 以及王文平 (Wenping Wang)。用于单目图像几何估计的自适应表面法线约束。预印本 arXiv:2402.05869，2024 年 2 月。

[41] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2837-2845, 2021. 2, 3

罗世通 (Shitong Luo) 和胡伟 (Wei Hu)。用于 3D 点云生成的扩散概率模型。见《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 2837 - 2845 页，2021 年 2 月、3 月。

[42] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In $CVPR$, 2023.2, 3, 7, 8

卢克·梅拉斯 - 基里亚齐 (Luke Melas-Kyriazi)、伊罗·莱娜 (Iro Laina)、克里斯蒂安·鲁普雷希特 (Christian Rupprecht) 以及安德里亚·韦尔达利 (Andrea Vedaldi)。Realfusion: 从单张图像进行任意物体的 360 度重建。见 $CVPR$, 2023.2, 3, 7, 8

[43] 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 $ECCV$, 2020.5

本·米尔登霍尔 (Ben Mildenhall)、普拉图尔·P·斯里尼瓦桑 (Pratul P Srinivasan)、马修·坦西克 (Matthew Tancik)、乔纳森·T·巴伦 (Jonathan T Barron)、拉维·拉马穆尔蒂 (Ravi Ramamoorthi) 以及任 Ng。神经辐射场 (NeRF): 将场景表示为神经辐射场以进行视图合成。见 $ECCV$, 2020.5

[44] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In $CVPR$, 2023.2, 3

诺曼·米勒 (Norman Müller)、亚瓦尔·西迪基 (Yawar Siddiqui)、洛伦佐·波尔齐 (Lorenzo Porzi)、塞缪尔·罗塔·布洛 (Samuel Rota Bulo)、彼得·孔奇德 (Peter Kontschieder) 以及马蒂亚斯·尼斯纳 (Matthias Nießner)。Diffrf: 渲染引导的 3D 辐射场扩散。见 $CVPR$, 2023.2, 3

[45] 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. 2, 3, 7

亚历克斯·尼科尔 (Alex Nichol)、许熙宇 (Heewoo Jun)、普拉富拉·达里瓦尔 (Prafulla Dhariwal)、帕梅拉·米什金 (Pamela Mishkin) 以及马克·陈 (Mark Chen)。Point-e: 一个从复杂提示生成 3D 点云的系统。预印本 arXiv:2212.08751，2022 年 2 月、3 月、7 月。

[46] Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc Van Gool, and Sergey Tulyakov. Autodecoding latent 3 d diffusion models. arXiv preprint arXiv:2307.05445, 2023. 2, 3

埃万杰洛斯·恩塔维利斯 (Evangelos Ntavelis)、阿利亚克桑德尔·西亚罗欣 (Aliaksandr Siarohin)、凯尔·奥尔谢夫斯基 (Kyle Olszewski)、王朝阳 (Chaoyang Wang)、吕克·范·古尔 (Luc Van Gool) 和谢尔盖·图利亚科夫 (Sergey Tulyakov)。自动解码潜在 3 d 扩散模型。预印本 arXiv:2307.05445，2023 年。2, 3

[47] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Milden-hall. Dreamfusion: Text-to-3d using 2d diffusion. In ICLR, 2023. 2, 3, 7

本·普尔 (Ben Poole)、阿杰伊·贾恩 (Ajay Jain)、乔纳森·T·巴伦 (Jonathan T Barron) 和本·米尔登霍尔 (Ben Milden - hall)。Dreamfusion: 使用 2D 扩散的文本到 3D 转换。收录于 ICLR，2023 年。2, 3, 7

[48] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Sko-rokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3 d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843, 2023. 2, 3, 7

钱国成 (Guocheng Qian)、麦金杰 (Jinjie Mai)、阿卜杜拉·哈姆迪 (Abdullah Hamdi)、任健 (Jian Ren)、阿利亚克桑德尔·西亚罗欣 (Aliaksandr Siarohin)、李冰 (Bing Li)、李欣莹 (Hsin - Ying Lee)、伊万·斯科罗霍多夫 (Ivan Skorokhodov)、彼得·翁卡 (Peter Wonka)、谢尔盖·图利亚科夫 (Sergey Tulyakov) 等。Magic123: 使用 2D 和 3D 扩散先验从单张图像生成高质量 3 d 物体。预印本 arXiv:2306.17843，2023 年。2, 3, 7

[49] 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. 2, 3, 4

亚历克·拉德福德 (Alec Radford)、郑旭 (Jong Wook Kim)、克里斯·哈拉西 (Chris Hallacy)、阿迪蒂亚·拉梅什 (Aditya Ramesh)、加布里埃尔·戈 (Gabriel Goh)、桑迪尼·阿加瓦尔 (Sandhini Agarwal)、吉里什·萨斯特里 (Girish Sastry)、阿曼达·阿斯凯尔 (Amanda Askell)、帕梅拉·米什金 (Pamela Mishkin)、杰克·克拉克 (Jack Clark) 等。从自然语言监督中学习可迁移的视觉模型。收录于 ICML，2021 年。2, 3, 4

[50] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aber-man, Michael Rubinstein, Jonathan Barron, et al. Dream-booth3d: Subject-driven text-to-3d generation. arXiv preprint arXiv:2303.13508, 2023. 3

阿米特·拉杰 (Amit Raj)、斯里尼瓦斯·卡扎 (Srinivas Kaza)、本·普尔 (Ben Poole)、迈克尔·尼迈耶 (Michael Niemeyer)、纳塔尼尔·鲁伊斯 (Nataniel Ruiz)、本·米尔登霍尔 (Ben Mildenhall)、希兰·扎达 (Shiran Zada)、基弗·阿伯曼 (Kfir Aberman)、迈克尔·鲁宾斯坦 (Michael Rubinstein)、乔纳森·巴伦 (Jonathan Barron) 等。Dream - booth3d: 基于主题驱动的文本到 3D 生成。预印本 arXiv:2303.13508，2023 年。3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In $CVPR$, 2022.3, 7

罗宾·龙巴赫 (Robin Rombach)、安德烈亚斯·布拉特曼 (Andreas Blattmann)、多米尼克·洛伦茨 (Dominik Lorenz)、帕特里克·埃瑟 (Patrick Esser) 和比约恩·奥默 (Björn Ommer)。使用潜在扩散模型进行高分辨率图像合成。收录于 $CVPR$, 2022.3, 7

[52] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. Ditto-nerf: Diffusion-based iterative text to omnidirectional 3 d model. arXiv preprint arXiv:2304.02827, 2023. 3

徐惠基 (Hoigi Seo)、金河妍 (Hayeon Kim)、金光贤 (Gwanghyun Kim) 和千世英 (Se Young Chun)。Ditto - nerf: 基于扩散的迭代文本到全向 3 d 模型转换。预印本 arXiv:2304.02827，2023 年。3

[53] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. arXiv preprint arXiv:2303.07937, 2023. 3

徐俊英 (Junyoung Seo)、张宇锡 (Wooseok Jang)、郭民燮 (Min - Seop Kwak)、高在勋 (Jaehoon Ko)、金贤秀 (Hyeonsu Kim)、金俊浩 (Junho Kim)、金镇华 (Jin - Hwa Kim)、李智英 (Jiyoung Lee) 和李承龙 (Seungryong Kim)。让 2D 扩散模型了解 3D 一致性以实现稳健的文本到 3D 生成。预印本 arXiv:2303.07937，2023 年。3

[54] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything- 3d: Towards single-view anything reconstruction in the wild. arXiv preprint arXiv:2304.10261, 2023. 3

沈秋红 (Qiuhong Shen)、杨兴义 (Xingyi Yang) 和王新潮 (Xinchao Wang)。Anything - 3d: 实现野外单视图任意物体重建。预印本 arXiv:2304.10261，2023 年。3

[55] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512, 2023. 2, 3, 4

施怡春 (Yichun Shi)、王鹏 (Peng Wang)、叶江龙 (Jianglong Ye)、龙麦 (Mai Long)、李可杰 (Kejie Li) 和杨晓 (Xiao Yang)。Mvdream: 用于 3D 生成的多视图扩散。预印本 arXiv:2308.16512，2023 年。2, 3, 4

[56] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In $ICML$, 2015.3, 4

雅沙·索尔 - 迪克斯坦 (Jascha Sohl - Dickstein)、埃里克·韦斯 (Eric Weiss)、尼鲁·马赫什瓦拉纳坦 (Niru Maheswaranathan) 和苏里亚·甘古利 (Surya Ganguli)。使用非平衡热力学进行深度无监督学习。收录于 $ICML$, 2015.3, 4

[57] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea

斯坦尼斯瓦夫·希马诺维奇 (Stanislaw Szymanowicz)、克里斯蒂安·鲁普雷希特 (Christian Rupprecht) 和安德里亚

Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. arXiv preprint arXiv:2306.07881, 2023. 3

韦代尔迪 (Vedaldi)。视图集扩散: 基于 (0 -) 图像条件的从 2D 数据生成 3D 生成模型。预印本 arXiv:2306.07881，2023 年。3

[58] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In $ICCV$, 2023. 3

唐俊树 (Junshu Tang)、王腾飞 (Tengfei Wang)、张博 (Bo Zhang)、张婷 (Ting Zhang)、易冉 (Ran Yi)、马利庄 (Lizhuang Ma) 和陈东 (Dong Chen)。Make - it - 3d: 利用扩散先验从单张图像进行高保真 3D 创建。收录于 $ICCV$ ，2023 年。3

[59] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multiview image generation with correspondence-aware diffusion. arXiv preprint arXiv:2307.01097, 2023. 3

唐世涛 (Shitao Tang)、张富阳 (Fuyang Zhang)、陈家成 (Jiacheng Chen)、王鹏 (Peng Wang) 和古川安孝 (Yasutaka Furukawa)。Mvdiffusion: 通过对应感知扩散实现整体多视图图像生成。预印本 arXiv:2307.01097，2023 年。3

[60] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B Tenenbaum, Frédo Durand, William T Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. arXiv preprint arXiv:2306.11719, 2023. 3

阿尤什·特瓦里 (Ayush Tewari)、尹天威 (Tianwei Yin)、乔治·卡泽纳维特 (George Cazenavette)、西蒙·雷奇科夫 (Semon Rezchikov)、约书亚·B·特南鲍姆 (Joshua B Tenenbaum)、弗雷多·迪朗 (Frédo Durand)、威廉·T·弗里曼 (William T Freeman) 和文森特·西茨曼 (Vincent Sitzmann)。带正向模型的扩散: 在无直接监督的情况下解决随机逆问题。预印本 arXiv:2306.11719，2023 年。3

[61] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3 d meshes from text prompts. arXiv preprint arXiv:2304.12439, 2023. 3

克里斯蒂娜·察利科格鲁 (Christina Tsalicoglou)、法比安·曼哈特 (Fabian Manhardt)、亚历山德罗·托尼奥尼 (Alessio Tonioni)、迈克尔·尼迈耶 (Michael Niemeyer) 和费德里科·通巴里 (Federico Tombari)。Textmesh: 根据文本提示生成逼真的 3 d 网格。预印本 arXiv:2304.12439，2023 年。3

[62] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In $CVPR$, 2023.3

曾宏宇 (Hung-Yu Tseng)、李钦波 (Qinbo Li)、金昌一 (Changil Kim)、苏希布·阿尔西桑 (Suhib Alsisan)、黄家斌 (Jia-Bin Huang) 和约翰内斯·科普夫 (Johannes Kopf)。使用姿态引导的扩散模型进行一致的视图合成。见 $CVPR$, 2023.3

[63] 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, 2023.2,3

王浩辰 (Haochen Wang)、杜晓丹 (Xiaodan Du)、李佳豪 (Jiahao Li)、雷蒙德·A·叶 (Raymond A Yeh) 和格雷格·沙克纳罗维奇 (Greg Shakhnarovich)。分数雅可比链: 提升预训练的二维扩散模型以进行三维生成。见计算机视觉与模式识别会议 (CVPR)，2023 年。2,3

[64] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In NeurIPS, 2021. 5

王鹏 (Peng Wang)、刘凌杰 (Lingjie Liu)、刘渊 (Yuan Liu)、克里斯蒂安·特奥博尔特 (Christian Theobalt)、小村拓 (Taku Komura) 和王文平 (Wenping Wang)。Neus: 通过体渲染学习神经隐式表面以进行多视图重建。见神经信息处理系统大会 (NeurIPS)，2021 年。5

[65] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In CVPR, 2023. 2,3

王腾飞 (Tengfei Wang)、张博 (Bo Zhang)、张婷 (Ting Zhang)、顾书阳 (Shuyang Gu)、包建民 (Jianmin Bao)、塔达斯·巴尔图塞蒂斯 (Tadas Baltrusaitis)、沈晶晶 (Jingjing Shen)、陈东 (Dong Chen)、文芳 (Fang Wen)、陈启峰 (Qifeng Chen) 等。罗丹 (Rodin): 使用扩散技术雕刻三维数字化身的生成模型。见计算机视觉与模式识别会议 (CVPR)，2023 年。2,3

[66] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. TIP, 2004. 7, 8

王周 (Zhou Wang)、艾伦·C·博维克 (Alan C Bovik)、哈米德·R·谢赫 (Hamid R Sheikh) 和埃罗·P·西蒙切利 (Eero P Simoncelli)。图像质量评估: 从误差可见性到结构相似性。《图像处理汇刊》(TIP)，2004 年。7, 8

[67] 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. 2, 3

王正义 (Zhengyi Wang)、卢程 (Cheng Lu)、王逸凯 (Yikai Wang)、包凡 (Fan Bao)、李崇轩 (Chongxuan Li)、苏航 (Hang Su) 和朱军 (Jun Zhu)。多产梦想家 (Prolificdreamer): 通过变分分数蒸馏实现高保真和多样化的文本到三维生成。预印本 arXiv:2305.16213，2023 年。2, 3

[68] 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. 3

丹尼尔·沃森 (Daniel Watson)、威廉·陈 (William Chan)、里卡多·马丁 - 布劳拉 (Ricardo Martin-Brualla)、乔纳森·霍 (Jonathan Ho)、安德里亚·塔利亚萨基 (Andrea Tagliasacchi) 和穆罕默德·诺鲁兹 (Mohammad Norouzi)。使用扩散模型进行新颖视图合成。预印本 arXiv:2210.04628，2022 年。3

[69] Jinbo Wu, Xiaobo Gao, Xing Liu, Zhengyang Shen, Chen Zhao, Haocheng Feng, Jingtuo Liu, and Errui Ding. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. arXiv preprint arXiv:2307.16183, 2023. 3

吴金波 (Jinbo Wu)、高晓波 (Xiaobo Gao)、刘星 (Xing Liu)、沈正阳 (Zhengyang Shen)、赵晨 (Chen Zhao)、冯浩成 (Haocheng Feng)、刘景拓 (Jingtuo Liu) 和丁二锐 (Errui Ding)。高清融合 (Hd-fusion): 利用多重噪声估计进行详细的文本到三维生成。预印本 arXiv:2307.16183，2023 年。3

[70] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. arXiv preprint arXiv:2303.17905, 2023. 3

向剑锋 (Jianfeng Xiang)、杨蛟龙 (Jiaolong Yang)、黄彬彬 (Binbin Huang) 和童欣 (Xin Tong)。使用二维扩散模型进行三维感知图像生成。预印本 arXiv:2303.17905，2023 年。3

[71] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild

徐德佳 (Dejia Xu)、江一帆 (Yifan Jiang)、王培豪 (Peihao Wang)、范志文 (Zhiwen Fan)、王毅 (Yi Wang) 和王章阳 (Zhangyang Wang)。Neurallift - 360: 将野外的

2d photo to a 3d object with 360 views. arXiv e-prints, pages arXiv-2211, 2022. 3

二维照片提升为具有 360 度视图的三维物体。预印本，arXiv - 2211 页，2022 年。3

[72] Paul Yoo, Jiaxian Guo, Yutaka Matsuo, and Shixiang Shane Gu. Dreamsparse: Escaping from plato’s cave with 2 d frozen diffusion model given sparse views. CoRR, 2023. 3

保罗·尤 (Paul Yoo)、郭佳贤 (Jiaxian Guo)、松尾丰 (Yutaka Matsuo) 和顾世翔 (Shixiang Shane Gu)。梦幻稀疏 (Dreamsparse): 在给定稀疏视图的情况下，利用 2 d 冻结扩散模型逃离柏拉图洞穴。计算机研究报告 (CoRR)，2023 年。3

[73] Chaohui Yu, Qiang Zhou, Jingliang Li, Zhe Zhang, Zhibin Wang, and Fan Wang. Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. arXiv preprint arXiv:2307.13908, 2023. 3

余朝晖 (Chaohui Yu)、周强 (Qiang Zhou)、李京亮 (Jingliang Li)、张哲 (Zhe Zhang)、王志斌 (Zhibin Wang) 和王帆 (Fan Wang)。点到三维 (Points - to - 3d): 弥合稀疏点与形状可控的文本到三维生成之间的差距。预印本 arXiv:2307.13908，2023 年。3

[74] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In $ICCV$, 2023 . 3

杰森·J·余 (Jason J. Yu)、费雷什特·福尔加尼 (Fereshteh Forghani)、康斯坦丁诺斯·G·德尔帕尼斯 (Konstantinos G. Derpanis) 和马库斯·A·布鲁贝克 (Marcus A. Brubaker)。使用扩散模型进行长期光度一致的新颖视图合成。见 $ICCV$, 2023 。3

[75] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3 d shape generation. In NeurIPS, 2022.2 ,3

曾晓辉 (Xiaohui Zeng)、阿拉什·瓦赫达特 (Arash Vahdat)、弗朗西斯·威廉姆斯 (Francis Williams)、赞·戈伊西奇 (Zan Gojcic)、奥尔·利塔尼 (Or Litany)、桑贾·菲德勒 (Sanja Fidler) 和卡斯滕·克雷斯 (Karsten Kreis)。Lion: 用于 3 d 形状生成的潜在点扩散模型。发表于《神经信息处理系统大会》(NeurIPS)，2022 年。2 ,3

[76] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. In SIGGRAPH, 2023.2,3

张彪 (Biao Zhang)、唐佳鹏 (Jiapeng Tang)、马蒂亚斯·尼斯纳 (Matthias Niessner) 和彼得·翁卡 (Peter Wonka)。3dshape2vecset: 用于神经场和生成式扩散模型的三维形状表示。发表于《计算机图形学与交互技术年度研讨会》(SIGGRAPH)，2023 年。2,3

[77] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. arXiv preprint arXiv:2305.11588, 2023. 3

张景波 (Jingbo Zhang)、李小雨 (Xiaoyu Li)、万子玉 (Ziyu Wan)、王灿 (Can Wang) 和廖静 (Jing Liao)。Text2nerf: 基于文本驱动和神经辐射场的三维场景生成。预印本 arXiv:2305.11588，2023 年。3

[78] 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, 8

理查德·张 (Richard Zhang)、菲利普·伊索拉 (Phillip Isola)、阿列克谢·A·埃弗罗斯 (Alexei A Efros)、伊莱·谢赫特曼 (Eli Shechtman) 和奥利弗·王 (Oliver Wang)。深度特征作为感知指标的不合理有效性。发表于 $CVPR$, 2018.7, 8

[79] Yuhang Zheng, Xiangyu Chen, Yupeng Zheng, Songen Gu, Runyi Yang, Bu Jin, Pengfei Li, Chengliang Zhong, Zeng-mao Wang, Lina Liu, et al. Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping. arXiv preprint arXiv:2403.09637, 2024. 2

郑宇航 (Yuhang Zheng)、陈翔宇 (Xiangyu Chen)、郑宇鹏 (Yupeng Zheng)、顾松恩 (Songen Gu)、杨润仪 (Runyi Yang)、金步 (Bu Jin)、李鹏飞 (Pengfei Li)、钟成亮 (Chengliang Zhong)、王增茂 (Zeng-mao Wang)、刘莉娜 (Lina Liu) 等。Gaussiangrasper: 用于开放词汇机器人抓取的三维语言高斯 splatting。预印本 arXiv:2403.09637，2024 年。2

[80] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5826-5835, 2021. 2, 3

周林奇 (Linqi Zhou)、杜奕伦 (Yilun Du) 和吴佳俊 (Jiajun Wu)。通过点 - 体素扩散进行三维形状生成与补全。发表于《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集》，第 5826 - 5835 页，2021 年。2, 3

[81] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3 d reconstruction. In CVPR, 2023. 3

周志卓 (Zhizhuo Zhou) 和舒巴姆·图尔西阿尼 (Shubham Tulsiani)。Sparsefusion: 用于 3 d 重建的视图条件扩散蒸馏。发表于《计算机视觉与模式识别会议》(CVPR)，2023 年。3

[82] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. arXiv preprint arXiv:2305.18766, 2023. 3

约瑟夫·朱 (Joseph Zhu) 和庄培烨 (Peiye Zhuang)。Hifa: 具有高级扩散引导的高保真文本到三维转换。预印本 arXiv:2305.18766，2023 年。3