# Sparse3D: Distilling Multiview-Consistent Diffusion for Object Reconstruction from Sparse Views

**Zixin Zou**[1], **Weihao Cheng**[2], **Yan-Pei Cao**[2], **Shi-Sheng Huang**[3], **Ying Shan**[2], **Song-Hai Zhang**[1†]

[1]BNRist, Tsinghua University
[2]ARC Lab, Tencent PCG
[3]Beijing Normal University
{zouzx19@mails.,shz@}tsinghua.edu.cn, airrafer@hotmail.com, {caoyanpei,shishenghuang.net}@gmail.com, yingsshan@tencent.com

## Abstract

Reconstructing 3D objects from extremely sparse views is a long-standing and challenging problem. While recent techniques employ image diffusion models for generating plausible images at novel viewpoints or for distilling pre-trained diffusion priors into 3D representations using score distillation sampling (SDS), these methods often struggle to simultaneously achieve high-quality, consistent, and detailed results for both novel-view synthesis (NVS) and geometry. In this work, we present *Sparse3D*, a novel 3D reconstruction method tailored for sparse view inputs. Our approach distills robust priors from a multiview-consistent diffusion model to refine a neural radiance field. Specifically, we employ a controller that harnesses epipolar features from input views, guiding a pre-trained diffusion model, such as Stable Diffusion, to produce novel-view images that maintain 3D consistency with the input. By tapping into 2D priors from powerful image diffusion models, our integrated model consistently delivers high-quality results, even when faced with open-world objects. To address the blurriness introduced by conventional SDS, we introduce category-score distillation sampling (C-SDS) to enhance detail. We conduct experiments on CO3DV2 which is a multi-view dataset of real-world objects. Both quantitative and qualitative evaluations demonstrate that our approach outperforms previous state-of-the-art works on the metrics regarding NVS and geometry reconstruction.

## Introduction

Reconstructing 3D objects from sparse-view images remains a pivotal challenge in the realms of computer graphics and computer vision. This technique has a wide range of applications such as Augmented and Virtual Reality (AR/VR). The advent of the Neural Radiance Field (NeRF) and its subsequent variants has catalyzed significant strides in geometry reconstruction and novel-view synthesis, as delineated in recent studies (Mildenhall et al. 2020; Wang et al. 2021a; Yariv et al. 2021). However, NeRFs exhibit limitations when operating on extremely sparse views, specifically with as few as 2 or 3 images. In these scenarios, the synthesized novel views often suffer in quality due to the limited input observations.

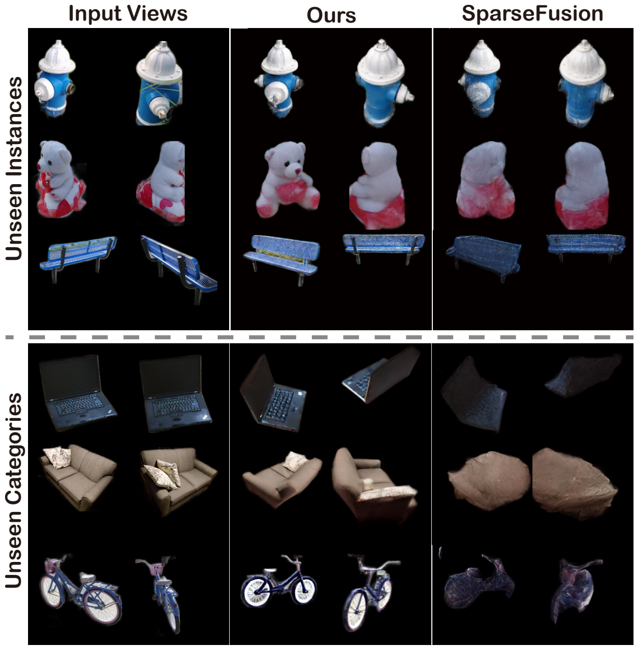

Figure 1: Novel-view synthesis from two input views using our Sparse3D and SparseFusion. Our approach can achieve higher-quality images with more details for unseen instances, especially for the unobserved regions of them (e.g., the left face of the teddybear). Furthermore, our approach can generalize to some unseen categories without any further finetuning, while SparseFusion fails.

Existing methods for sparse-view reconstruction typically leverage a generalizable NeRF model, pre-trained on multiview datasets, to infer 3D representations from projected image features (Yu et al. 2021; Chibane et al. 2021). However, these approaches tend to regress to the mean, failing to produce perceptually sharp outputs, especially in intricate details. To produce plausible results, either in terms of geometry or appearance, from limited observations, sev-

[†] corresponding author

ArXiv version with supplementary materials is available at https://arxiv.org/abs/2308.14078 eral studies have turned to image generation models, such as the diffusion model (Rombach et al. 2022), to "imagine" unseen views based on provided images (Chan et al. 2023; Zhou and Tulsiani 2023). For example, Zero123 (Liu et al. 2023) trains a view-conditioned diffusion model on a large synthetic dataset and achieves impressive results. However, their generated images across different views may not be consistent. Thus, while these view-conditioned diffusion models can produce satisfactory images, their quality and generalization ability are often constrained by the scarcity of posed image datasets. Large-scale image diffusion models (Ramesh et al. 2021; Saharia et al. 2022; Rombach et al. 2022), which are pre-trained on billions of 2D images (Schuhmann et al. 2022), excel in generating high-quality and diverse images. However, despite the diverse, general capability of such models, in 3D reconstruction tasks, users need to synthesize specific instances that are coherent with user-provided input images. Even with recent model customization methods (Kumari et al. 2023; Ruiz et al. 2023; Gal et al. 2022), they prove unwieldy and often fail to produce the specific concept with sufficient fidelity. Consequently, the potential of merging the capabilities of pre-trained large image diffusion models with the viewpoint and appearance perception of specific instances remains an open avenue of exploration.

In contrast to directly generating images at novel views, some recent works explore distilling the priors of pre-trained diffusion models into a NeRF (neural radiance field) framework. This approach facilitates 3D-consistent novel-view synthesis and allows for mesh extraction from the NeRF. Notable works such as DreamFusion (Poole et al. 2023) and SJC (Wang et al. 2023a) employ score distillation sampling (SDS) to harness off-the-shelf diffusion models for text-to-3D generation. However, a persistent challenge with SDS is the production of blurry and oversaturated outputs, attributed to noisy gradients, which in turn compromises the quality of NeRF reconstructions.

In this work, we present *Sparse3D*, a novel 3D reconstruction approach designed to reconstruct high-fidelity 3D objects from sparse and posed input views. Our method hinges on two pivotal components: **(1)** a diffusion model that ensures both multiview consistency and fidelity to user-provided input images while retaining the powerful generalization capabilities of Stable Diffusion (Rombach et al. 2022), and **(2)** a category-score distillation sampling (C-SDS) strategy. At its core, we distill the priors from our fidelity-preserving, multiview-consistent diffusion model into the NeRF reconstruction using an enhanced category-score distillation sampling. Specifically, for the multiview-consistent diffusion model, we propose to utilize an epipolar controller to guide the off-the-shelf Stable Diffusion model to generate novel-view images that are 3D consistent with the content of input images. Notably, by fully harnessing the 2D priors present in Stable Diffusion, our model exhibits robust generalization capabilities, producing high-quality images even when confronted with open-world, unseen objects. To overcome the problem of blurry, oversaturated, and non-detailed results caused by SDS during NeRF reconstruction, we draw inspiration from VSD (Wang et al. 2023b) and propose a category-score distillation sampling strategy (C-SDS).

We evaluate *Sparse3D* on the Common Object in 3D (CO3DV2) dataset and benchmark it against existing approaches. The results show that our approach outperforms state-of-the-art techniques in terms of the quality of both synthesized novel views and reconstructed geometry. Importantly, *Sparse3D* exhibits superior generalization capabilities, particularly for object categories not present in the training domain.

# Related Works

## Multi-view 3D Reconstruction

Multi-view 3D reconstruction is a long-standing problem with impressive works such as traditional Structure-from-Motion (S*f*M) (Schönberger and Frahm 2016) or Multi-view-Stereo (MVS) (Schönberger et al. 2016), and recent learning based approaches (Yao et al. 2018; Yu and Gao 2020). The success of NeRF (Mildenhall et al. 2020; Müller et al. 2022) has led to impressive outcomes in novel-view synthesis and geometric reconstruction. However, these methods still struggle to produce satisfactory results for extremely sparse view scenarios. Subsequent works proposed to use regularization (semantic (Jain, Tancik, and Abbeel 2021), frequency (Yang, Pavone, and Wang 2023), geometry and appearance (Niemeyer et al. 2022)) and geometric priors (e.g. depth (Deng et al. 2022; Roessle et al. 2022) or normal (Yu et al. 2022)) but remain to be inadequate for view generation in unobserved regions, due to the essential lack of scene priors.

## Generalizable Novel-view Synthesis

For generalizable novel-view synthesis using NeRF, some approaches utilize projected features of the sampling points in volumetric rendering (Yu et al. 2021; Wang et al. 2021b; Chibane et al. 2021), or new neural scene representations, such as Light Field Network (Suhail et al. 2022b,a) or Scene Representation Transformer (Sajjadi et al. 2022) for better generalizable novel-view synthesis. Subsequent researches (Kulhánek et al. 2022; Chan et al. 2023; Yoo et al. 2023) propose to further utilize generative models (e.g. VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017) and diffusion model (Rombach et al. 2022)) to generate unseen images. However, these methods didn't have any 3D-aware scene priors, which limits their potential applications. In this paper, we leverage the feature map from a generalizable renderer to guide a pre-trained diffusion model to generate multiview-consistent images, and then distill the diffusion prior into NeRF reconstruction for both novel-view synthesis and geometry reconstruction.

## 3D Generation with 2D Diffusion Model

Diffusion-denoising probabilistic models have brought a boom of generation tasks for 2D images and 3D contents in recent years. Inspired by early works which use CLIP embedding (Jain, Tancik, and Abbeel 2021; Wang et al. 2022; Jain et al. 2022) or GAN (Pan et al. 2021) to regularize the NeRF, DreamFusion (Poole et al. 2023) and SJC (Wang

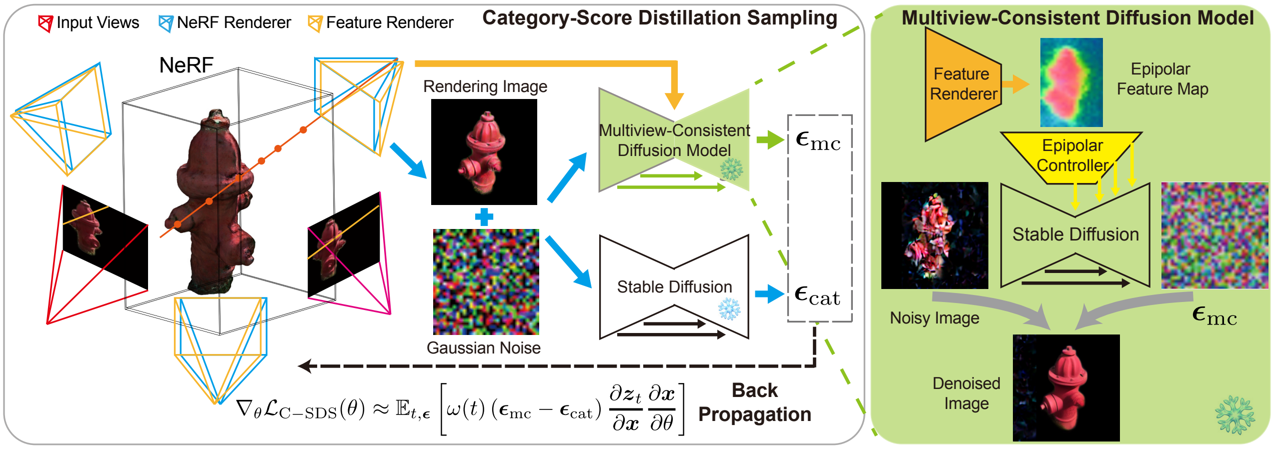
Input Views
NeRF Renderer
Feature Renderer
Category-Score Distillation Sampling
NeRF
Rendering Image
Gaussian Noise
Multiview-Consistent Diffusion Model
Stable Diffusion
$\epsilon_{\mathrm{mc}}$
$\epsilon_{\mathrm{cat}}$
$\nabla_\theta \mathcal{L}_{\mathrm{C-SDS}}(\theta) \approx \mathbb{E}_{t,\epsilon}\left[\omega(t)\left(\epsilon_{\mathrm{mc}} - \epsilon_{\mathrm{cat}}\right)\frac{\partial z_t}{\partial x}\frac{\partial x}{\partial \theta}\right]$
Back Propagation
Multiview-Consistent Diffusion Model
Feature Renderer
Epipolar Feature Map
Epipolar Controller
Stable Diffusion
Noisy Image
$\epsilon_{\mathrm{mc}}$
Denoised Image

Figure 2: Overview of Sparse3D. Our approach consists of two key components: a multiview-consistent diffusion model and a category-score distillation sampling. We utilize epipolar feature map to control the Stable Diffusion model to generate images consistent with the content of input images, serving as a multiview-consistent diffusion model. Based on such a model, we propose a category-score distillation sampling (C-SDS) strategy to achieve more detailed results during NeRF reconstruction.

et al. 2023a) propose a score distillation sampling (SDS) strategy to guide the NeRF optimization for impressive text-to-3D generation. ProlificDreamer (Wang et al. 2023b) proposes variational score distillation (VSD) for more high-fidelity and diverse text-to-3D generation. Magic3D (Lin et al. 2023) improves the 3D generation quality by a two-stage coarse-to-fine strategy. To generate 3D results consistent with the input image observation, subsequent works leverage textual-inversion (Melas-Kyriazi et al. 2023) or denoised-CLIP loss with depth prior (Tang et al. 2023). When additional geometry prior are available (e.g. point clouds from Point-E (Nichol et al. 2022)), some works (Seo et al. 2023; Yu et al. 2023) can produce more 3D consistent creation. In addition to lifting a pre-trained diffusion model, Zero123 (Liu et al. 2023), SparseFusion (Zhou and Tulsiani 2023) and NerfDiff (Gu et al. 2023) train a viewpoint-conditioned diffusion model and achieve impressive results. Instead of training a diffusion model or directly lifting a pre-trained diffusion model, our approach leverages both the advantages of them to train a multiview-consistent diffusion model, with a category-score distillation sampling to improve the results of SDS for more details.

## Method

Given $N$ input images $\{I_n\}$ of an object with corresponding camera poses $\{T_n\}$, where $N$ can be as few as 2, our goal is to reconstruct a neural radiance field (NeRF), enabling generalizable novel view synthesis and high-quality surface reconstruction. To realize this goal, we propose Sparse3D, which distills a multiview-consistent diffusion model prior into the NeRF representation of an object, using a category-score distillation sampling (C-SDS) strategy. Figure 2 shows the overview of our approach. The multiview-consistent diffusion model extracts epipolar features from sparse input views and uses a control network to guide the Stable Diffusion model to generate novel-view images that are faithful to the object shown in the images. A NeRF is then reconstructed with the guidance of the diffusion model. To overcome the blurry problem that occurred in SDS, we propose C-SDS. Benefiting from it, the gradients conditioned on the category prior maintain the optimization with a tightened region of the search space, leading to more detailed results. Finally, our method achieves more consistent and high-quality results of novel-view synthesis and geometry reconstruction.

## Multiview-Consistent Diffusion Model

Our diffusion model consists of a feature renderer, an epipolar controller, and a Stable Diffusion model, where the epipolar controller and the Stable Diffusion model together constitute the noise predictor $\epsilon_\beta$, as shown in Figure 3. The feature renderer $g_\psi$ takes a set of posed images and viewpoint $\pi$ as input, subsequently outputting an epipolar feature map $f_c = g_\psi(\pi, I_1, ..., I_n, T_1, ..., T_n)$, which serves as the input for the epipolar controller. To unify the pre-trained diffusion model and multiview-consistent perception ability for a specific object, we draw inspiration from ControlNet (Zhang and Agrawala 2023). ControlNet enables image generation controlled by conditional inputs (such as depth maps). Instead, we use the epipolar feature map to guide a pre-trained diffusion model to generate images consistent with the content of input images from various viewpoints.

**Feature Renderer.** Previous works acquire the feature map $f_c$ through rendering from Triplane (Gu et al. 2023), 3D Volume (Chan et al. 2023) or epipolar feature transformer (Zhou and Tulsiani 2023). In this paper, we adapt epipolar feature transformer (EFT) following  (Zhou and Tulsiani 2023). The EFT, derived from GPNR (Suhail et al. 2022a), learns a network $g_\psi$ to predict color of given ray $\boldsymbol{r}$ from input images. The rendering process primarily involves three transformers, which output attention weights used to blend colors over input views and epipolar lines for the final prediction. We implement two modifications to the EFT for improved results: (1) a mask embedding and a relative cam-

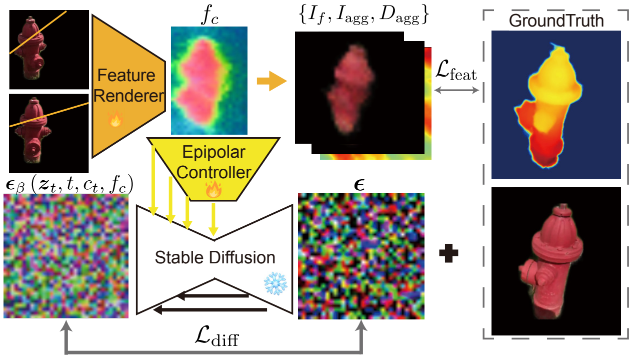
Feature Renderer
$f_c$
$\{I_f, I_{\text{agg}}, D_{\text{agg}}\}$
$\mathcal{L}_{\text{feat}}$
GroundTruth
Epipolar Controller
$\epsilon_\beta(z_t, t, c_t, f_c)$
$\epsilon$
Stable Diffusion
$\mathcal{L}_{\text{diff}}$

Figure 3: Multiview-consistent diffusion model. Our multiview-consistent diffusion model comprises a feature renderer, an epipolar controller, and a Stable Diffusion model.

era transformation embedding are concatenated with other transformer token features. (2) To enhance generalizability and achieve better geometry awareness, we also obtain the aggregated color $I_{\text{agg}}$ and depth images $D_{\text{agg}}$ by attention weights of transformers to compute loss.

**Epipolar Controller.** Given feature maps $f_c$ rendered at arbitrary viewpoints, we propose to learn an epipolar controller to guide a pre-trained diffusion model to generate multiview-consistent images with high quality. Our epipolar controller takes epipolar feature map $f_c$ and category text prompt $c_t$ as input, subsequently outputting the latent features that are fused with the latent features of Stable Diffusion. Rather than training a new diffusion model, we hope to retain the rich 2D priors from Stable Diffusion. Consequently, we jointly train our epipolar controller and feature renderer, while keeping the parameters of Stable Diffusion fixed. On the one hand, by utilizing the feature map, which contains implicit information about the appearance of the specific object and perception of the observation viewpoint, we can control a pre-trained diffusion model to generate images consistent with the content of input images from different viewpoints. On the other hand, our diffusion model inherits the high-quality image generation capabilities from Stable Diffusion, and the additional category prior in the text domain can also enhance the multiview consistency. Furthermore, these priors also enable our model to generalize to open-world unseen categories.

**Training.** Finally, we jointly train the feature renderer and the epipolar controller by the following objective function:

$$\mathcal{L} = \mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{diff}} \tag{1}$$

where $\mathcal{L}_{\text{feat}}$ is the loss for feature renderer and $\mathcal{L}_{\text{diff}}$ is the loss for epipolar controller.

While the feature map primarily serves as input for the controller in our pipeline, we also supervise it with color images and depth images to enhance its perception of appearance, observation viewpoints, and geometry awareness. For a query ray $r$ from novel view when given input images, we decode the color $I_f$ from the feature map and supervise it using ground-truth color values. Additionally, to improve generalizability and geometry awareness, we employ an MSE loss on aggregated color $I_{\text{agg}}$ and depth $D_{\text{agg}}$. We formulate the objective function as follows:

$$\begin{aligned}\mathcal{L}_{\text{feat}} = \sum_r ||I_f(r) - I(r)||^2 + ||I_{\text{agg}}(r) - I(r)||^2 \\ + ||D_{\text{agg}}(r) - D(r)||^2\end{aligned} \tag{2}$$

where $I(r)$ and $D(r)$ are ground-truth color and depth image respectively.

The diffusion model learns a conditional noise predictor to estimate the denoising score by adding Guassian-noise $\epsilon$ to clean data in $T$ timesteps. We minimize the noise prediction error at randomly sampled timestep $t$. The objective of the diffusion model conditioned on text prompt $c_t$ (we use the category name as the conditioned text prompt, e.g. "hydrant") and feature map $f_c$ is given by:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} ||\epsilon - \epsilon_\beta(z_t, t, c_t, f_c)||^2 \tag{3}$$

where $\epsilon_\beta$ is the conditional noise predictor of our diffusion model.

## NeRF Reconstruction with C-SDS

Building on our multiview-consistent diffusion model, we aim to optimize a neural radiance field (NeRF) parameterized with $\theta$, from which more 3D-consistent novel-view synthesis and underlying explicit geometry can be derived. Then to overcome the problem of blurry and non-detailed results in SDS, we propose a category-score distillation sampling (C-SDS) strategy.

**Category-Score Distillation Sampling.** We draw inspiration from VSD (Wang et al. 2023b) and propose a C-SDS for more detailed outcomes as follows:

$$\nabla_\theta \mathcal{L}_{\text{C-SDS}}(\theta) \approx \mathbb{E}_{t,\epsilon}\left[\omega(t)\left(\epsilon_{\text{mc}} - \epsilon_{\text{cat}}\right)\frac{\partial z_t}{\partial x}\frac{\partial x}{\partial \theta}\right] \tag{4}$$

where $\epsilon_{\text{mc}} = \epsilon_\beta(z_t, t, c_t, f_c)$ is the predicted noise by our multiview-consistent diffusion model, $\epsilon_{\text{cat}} = \epsilon_{\text{sd}}(z_t; t, c_t)$ is the predicted noise by Stable Diffusion conditioned text prompt of category $c_t$. And $\omega(t)$ is a weighting function that depends on the timestep $t$.

Instead of employing a Gaussian noise as SDS does, we replace it with an estimation $\epsilon_{\text{cat}}$ incorporating category prior from Stable Diffusion. By providing an approximation of the estimation of the score function of the distribution on rendering images with category prior, our C-SDS can deliver a better gradient with a tightened region of the search space, resulting in more detailed outputs. SDS relies on high classifier-free guidance (CFG, i.e. 100) to achieve a better convergence, but such high CFG may lead to over-saturation and over-smooth problems (Poole et al. 2023). In our experiment, when using a more multiview-consistent diffusion model, it can work with a small CFG (i.e. 7.5). However, the results still suffer from blurry and non-detailed outputs, as the update gradient is not accurate enough. ProlificDreamer utilizes a low-rank adaption (LoRA) of a pre-trained diffusion model to estimate the score function of the distribution

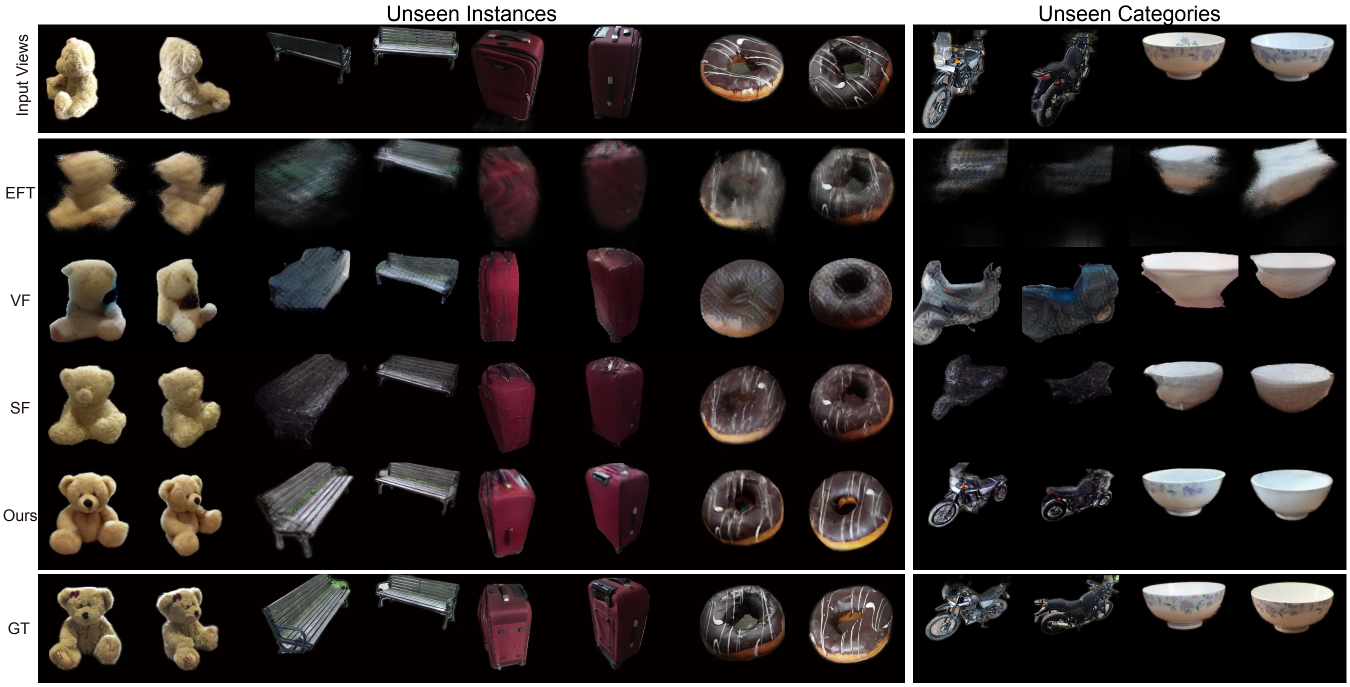

Figure 4: Qualitative comparison of novel-view synthesis when given 2 input views. Our approach achieves both high quality and more details of novel-view images compared to the others (e.g., the face of the teddybear), whenever with unseen instances and unseen categories.

on rendered images. We find that it is hard for LoRA to provide good estimation during our instance-specific optimization. Therefore, our proposed C-SDS offers a simple yet effective way to estimate the score function of the distribution on rendered images for more detailed results.

**One-step Estimation from Diffusion Model.** The predicted noise from the diffusion model can be used not only in C-SDS but also to estimate its one-step denoising image without requiring much extra computation:

$$\begin{aligned} \boldsymbol{z}_{1\text{step}} &= \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\beta\left(\boldsymbol{z}_t, t, c_t, f_c\right)\right), \\ \boldsymbol{x}_{1\text{step}} &= \mathcal{D}(\boldsymbol{z}_{1\text{step}}) \end{aligned} \tag{5}$$

where $\mathcal{D}$ is the decoder of Stable Diffusion. We leverage its one-step estimation to provide an additional regularization term by using perceptual distance and find that perception regularization improves the metrics of results. Specifically, we employ two perceptual losses, which include LPIPS loss (Zhang et al. 2018) and contextual loss (Mechrez, Talmi, and Zelnik-Manor 2018) to formulate the perception regularization from one-step estimation image:

$$\mathcal{L}_{\text{perp}} = \lambda_p \mathcal{L}_{\text{lpips}}(I, \boldsymbol{x}_{1step}) + \lambda_c \mathcal{L}_{\text{contextual}}(I, \boldsymbol{x}_{1step}) \tag{6}$$

**Reference Supervision.** Additionally, we use the reference input images $I$ with their masks $M$ to encourage a consistent appearance with the input images:

$$\mathcal{L}_{\text{ref}} = \lambda_r ||(\hat{I} - I) * \hat{M}||_2^2 + \lambda_m ||\hat{M} - M||_2^2 \tag{7}$$

where $\hat{I}$ and $\hat{M}$ are rendering image and mask, respectively.

**Overall Training.** We combine all of the losses, including $\mathcal{L}_{\mathrm{C-SDS}}, \mathcal{L}_{\text{perp}}, \mathcal{L}_{\text{ref}}$, to formulate the objective function of NeRF reconstruction for a specific object. Once NeRF reconstruction is complete, we can perform volume rendering for novel-view synthesis, and the underlying mesh can be extracted using Marching Cubes (Lorensen and Cline 1987).

## Experiment

In this section, we conduct a qualitative and quantitative evaluation of our approach on the 3D object dataset, CO3Dv2 dataset (Reizenstein et al. 2021), to demonstrate its effectiveness. CO3Dv2 dataset is a real-world dataset, which contains 51 common object categories. We first show the superior quality of novel-view synthesis and 3D reconstruction for unseen object instances in category-specific scenarios with varying numbers of input and then out-of-domain generalization ability for unseen categories.

**Implementation details.** For the feature renderer, we follow SparseFusion (Zhou and Tulsiani 2023) to use three groups of transformer encoders with four 256-dimensional layers to aggregate epipolar features. For the multiview-consistent model, we adopt the Stable Diffusion model v1.5 as our priors. For NeRF reconstruction, we adapt the threestudio (Guo et al. 2023), which is a unified framework for 3D content creation from various inputs, to implement the NeRF reconstruction for specific objects. We set the weights of the losses with $\lambda_p = 100$, $\lambda_c = 10$, $\lambda_r = 1000$ and $\lambda_m = 50$. NeRF optimization runs for 10,000 steps, which

| | Unseen Instances - 2 views | | | | | | Unseen Instance - 3 views | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | LPIPS | FID | CLIP | DISTS | PSNR | SSIM | LPIPS | FID | CLIP | DISTS |
| PN | 15.33 | 0.29 | 0.59 | 371.23 | 0.83 | 0.44 | 15.50 | 0.31 | 0.58 | 363.68 | 0.83 | 0.43 |
| EFT | **21.28** | 0.69 | 0.34 | 293.36 | 0.87 | 0.33 | **22.62** | 0.74 | 0.29 | 242.87 | 0.89 | 0.30 |
| VF | 18.42 | 0.71 | 0.29 | 248.23 | 0.82 | 0.29 | 18.91 | 0.72 | 0.28 | 240.21 | 0.87 | 0.29 |
| SF | **21.28** | 0.76 | 0.23 | 187.22 | 0.91 | 0.26 | 22.31 | 0.78 | 0.22 | 175.02 | 0.92 | 0.24 |
| Ours | 20.95 | **0.77** | **0.22** | **147.65** | **0.93** | **0.23** | 22.06 | **0.79** | **0.20** | **134.22** | **0.94** | **0.21** |
| | Unseen Instances - 6 views | | | | | | Unseen Categories - 2 views | | | | | |
| | PSNR | SSIM | LPIPS | FID | CLIP | DISTS | PSNR | SSIM | LPIPS | FID | CLIP | DISTS |
| PN | 15.65 | 0.33 | 0.55 | 344.58 | 0.85 | 0.42 | 14.82 | 0.31 | 0.50 | 314.45 | 0.81 | 0.44 |
| EFT | **24.47** | 0.80 | 0.23 | 161.78 | 0.93 | 0.25 | **19.31** | 0.56 | 0.41 | 318.64 | 0.87 | 0.38 |
| VF | 19.77 | 0.74 | 0.27 | 232.30 | 0.89 | 0.28 | 15.43 | 0.63 | 0.34 | 301.19 | 0.85 | 0.36 |
| SF | 23.69 | 0.80 | 0.20 | 154.20 | 0.93 | 0.22 | 18.83 | 0.70 | 0.28 | 290.45 | 0.88 | 0.34 |
| Ours | 23.92 | **0.82** | **0.18** | **116.10** | **0.95** | **0.19** | 18.83 | **0.72** | **0.23** | **164.30** | **0.93** | **0.26** |

Table 1: Quantitative comparisons of novel-view synthesis. We evaluate methods on unseen instances with varying numbers of input images, such as 2, 3, and 6, and on unseen categories with 2 input views. We report the average results across categories for each block.

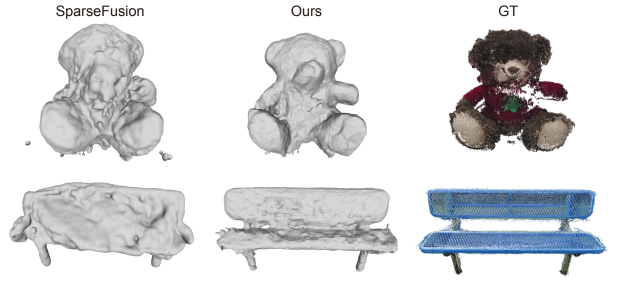

Figure 5: Geometry reconstruction using SparseFusion and Ours. The last column shows the ground-truth point cloud.

takes about 45 minutes on a single 3090 GPU.

## Experimental Settings

**Dataset.** We follow the *fewview-train* and *fewview-dev* splits provided by CO3Dv2 dataset (Reizenstein et al. 2021) for training and evaluation purposes, respectively. For the evaluation of unseen object instances within the same categories, we use the core subset with 10 categories to train the category-specific diffusion model for each category. To assess the out-of-domain generalization ability on unseen categories, we select 10 categories for evaluation and use the remaining 41 categories together for training. Due to the hour-long computation time required for our method, we evaluate only the first 10 object instances of each test split.

**Baselines.** We compare our approach with previous state-of-the-art baselines, including PixelNeRF (PN) (Yu et al. 2021), ViewFormer (VF) (Kulhánek et al. 2022), EFT and SparseFusion (SF) (Zhou and Tulsiani 2023). PixelNeRF and EFT are regression-based methods that deduce images at novel view by projection feature, where EFT is adapted from GPNR for sparse views settings by (Zhou and Tulsiani 2023). ViewFormer is a generative model that employs a VQ-VAE codebook and a transformer module for image generation. SparseFusion is the most relevant baseline to our approach, as it distills the diffusion model prior into NeRF reconstruction.

| | Unseen Instances | | Unseen Categories | |
|---|---|---|---|---|
| | CD ↓ | F-score ↑ | CD ↓ | F-score ↑ |
| SF | 0.27 | 0.23 | 0.37 | 0.18 |
| Ours | **0.21** | **0.32** | **0.27** | **0.28** |

Table 2: Quantitative comparison of geometry reconstruction. Since other baselines only produce images at novel views without 3D representation, we only report the results of ours and SparseFusion.

**Metrics.** We adopt several popular image quality assessments (IQA) to evaluate the quality of novel-view synthesis, including PSNR, SSIM, LPIPS (Zhang et al. 2018), FID (Heusel et al. 2017) and DISTS (Ding et al. 2022). Additionally, since our method can generate plausible results for unobserved regions, the evaluation between them and GT images may not be fair. Thus, we also adopt CLIP embedding similarity (Radford et al. 2021) of generated images with input images. Additionally, we evaluate the most commonly used 3D reconstruction quality metrics, including Chamfer Distance and F-score.

## Qualitative and Quantitative Evaluation

**Unseen Instances: 2 Views.** We first evaluate our approach with extremely sparse views (i.e. 2 views) for unseen object instances within the same categories. Table 1 demonstrates the quantitative comparison of ours and other baselines, with metrics averaged across 10 categories. Although ours has a slightly lower PSNR compared to the others, due to its formulation of pixel-wise MSE which favors mean color rendering results (e.g., blurry images), our approach outperforms all of the others in perception metrics (e.g. LPIPS, FID, etc.). As the qualitative results are shown in Figure 4, benefiting from two proposed key components, our approach achieves both high-quality and more detailed results with 3D consistency. In addition to novel-view synthesis, we evaluate the quality of geometry reconstruction by extracting underlying mesh from NeRF. We only com-

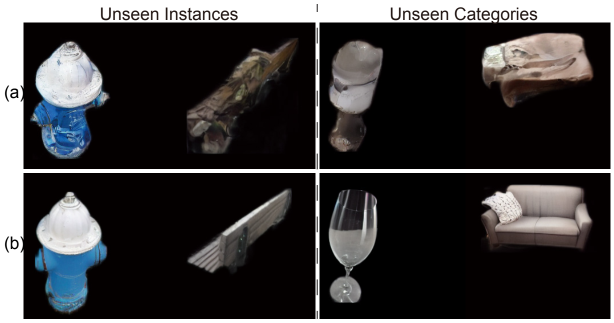

Figure 6: Effect of Stable Diffusion priors. (a) diffusion model from SparseFusion; (b) our diffusion model with Stable Diffusion priors.

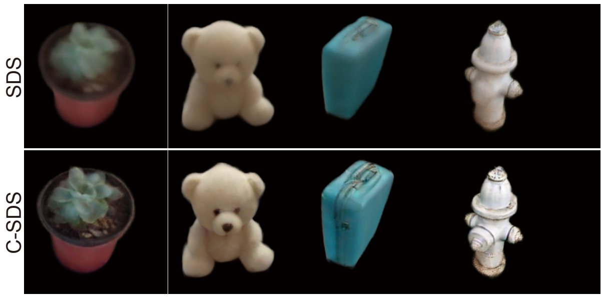

Figure 7: Effect of C-SDS to the quality of NVS from NeRF reconstruction. We can find the results of SDS are blurry and non-detailed in unobserved regions, while ours can generate more details with the same diffusion model.

pare ours with SparseFusion, while the others lack 3D representation. Table 2 shows that our approach significantly outperforms SparseFusion by a wide margin. Figure 5 also illustrates the mesh extracted from NeRF, where our results achieve sharper geometry with more details.

**Unseen Instances: Varying Views.** It's obvious that as the number of input views increases, the results of novel-view synthesis and geometry reconstruction improve. Table 1 shows the comparison of novel-view synthesis on 3 and 6 input views, which demonstrates that our approach consistently outperforms the others with varying input views. More detailed evaluation results for each category and more qualitative results of novel-view synthesis and explicit geometry can be found in supplementary materials.

**Unseen Categories.** We experiment to evaluate the generalization ability to unseen categories. Table 1 and Table 2 show the quantitative results of novel-view synthesis and geometry reconstruction. When confronted with the unseen categories that are out of the training domain, the performance of the other methods has a significant drop, while ours still maintains good performance, achieving the best results among them. The priors from Stable Diffusion enable our diffusion model to faithfully generate images of unseen categories. The last two columns of Figure 4 show the novel-view synthesis of these methods. Our approach still can achieve high-quality images with more details, while the others are blurry and somewhat meaningless. More evaluation of unseen categories can be found in supplementary materials.

## Ablation Studies

**Stable Diffusion Priors.** To evaluate the effect of Stable Diffusion priors, we compare ours and SparseFusion in directly generating novel view images without performing NeRF reconstruction, as shown in Figure 6. In unseen instances scenario, the diffusion model of SparseFusion can generate images at novel viewpoints consistent with the appearance of input images in a certain way (e.g. the blue hydrant with white head) but fails to achieve high-quality image generation. When the feature map is not reliable in some views, SparseFusion fails to generate a multiview-consistent image (e.g. the bench). However, our diffusion model can achieve higher-quality image generation. In the unseen categories scenario, the diffusion model of SparseFusion fails to generate meaningful images, while our method can be generalized to these objects (the last two columns in Figure 6).

**C-SDS.** We also investigate the effect of our distillation strategy on the quality of NeRF reconstruction, by implementing a version of using SDS. When using our multiview-consistent diffusion model with SDS, which can provide a more accurate gradient update direction, there is no need for a large CFG, but it's still not enough for detailed results. In our experiment with setting the CFG value as 7.5, it can achieve plausible results with successful convergence, but the blur problem is still unsolved, as shown in the first row of Figure 7. When applying our proposed C-SDS with the same CFG, it's evident that the results show more details, which demonstrates the effectiveness of the method.

## Limitations

The primary failure cases include (1) extremely partial observation of an object in input views; (2) the Janus problem and (3) sometimes thin structures or self-occlusion parts. Furthermore, our approach relies on accurate camera poses, which can be challenging to estimate directly from extremely sparse views, resulting in noisy estimates.

## Conclusion

In this paper, we introduce Sparse3D, a new approach to reconstructing high-quality 3D objects from sparse input views with camera poses. We utilize an epipolar controller to guide a pre-trained diffusion model to generate high-quality images that are 3D consistent with the content of input images, leading to a multiview-consistent diffusion model. Then, we distill the diffusion priors into NeRF optimization in a better way by using a category-score distillation sampling (C-SDS) strategy, resulting in more detailed results. Experiments demonstrate that our approach can achieve state-of-the-art results with higher quality and more details, even when confronted with open-world, unseen objects.

## Acknowledgments

We sincerely thank the reviewers for their valuable comments. This work was supported by the National Key Research and Development Program of China (No. 2023YFF0905104), the Natural Science Foundation of China (No. 62132012), Beijing Municipal Science and Technology Project (No. Z221100007722001) and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology. Shi-Sheng Huang was supported by the Natural Science Foundation of China (Project Number 62202057), State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2022B03).

## References

Chan, E. R.; Nagano, K.; Chan, M. A.; Bergman, A. W.; Park, J. J.; Levy, A.; Aittala, M.; Mello, S. D.; Karras, T.; and Wetzstein, G. 2023. Generative Novel View Synthesis with 3D-Aware Diffusion Models. *CoRR*, abs/2304.02602.

Chibane, J.; Bansal, A.; Lazova, V.; and Pons-Moll, G. 2021. Stereo Radiance Fields (SRF): Learning View Synthesis from Sparse Views of Novel Scenes. In *IEEE (CVPR)*.

Deng, K.; Liu, A.; Zhu, J.; and Ramanan, D. 2022. Depth-supervised NeRF: Fewer Views and Faster Training for Free. In *IEEE CVPR*, 12872–12881.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2022. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE TPAMI.*, 44(5): 2567–2581.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gu, J.; Trevithick, A.; Lin, K.; Susskind, J. M.; Theobalt, C.; Liu, L.; and Ramamoorthi, R. 2023. NerfDiff: Single-image View Synthesis with NeRF-guided Distillation from 3D-aware Diffusion. *CoRR*, abs/2302.10109.

Guo, Y.-C.; Liu, Y.-T.; Shao, R.; Laforte, C.; Voleti, V.; Luo, G.; Chen, C.-H.; Zou, Z.-X.; Wang, C.; Cao, Y.-P.; and Zhang, S.-H. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio. Accessed: 2023-05-01.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*, 6626–6637.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. In *IEEE CVPR*, 857–866.

Jain, A.; Tancik, M.; and Abbeel, P. 2021. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *ICCV*, 5865–5874.

Kulhánek, J.; Derner, E.; Sattler, T.; and Babuska, R. 2022. ViewFormer: NeRF-Free Neural Rendering from Few Images Using Transformers. In *ECCV*, volume 13675, 198–216.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE CVPR*.

Liu, R.; Wu, R.; Hoorick, B. V.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. arXiv:2303.11328.

Lorensen, W. E.; and Cline, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. In Stone, M. C., ed., *SIGGRAPH*, 163–169.

Mechrez, R.; Talmi, I.; and Zelnik-Manor, L. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In *ECCV*, volume 11218, 800–815.

Melas-Kyriazi, L.; Rupprecht, C.; Laina, I.; and Vedaldi, A. 2023. RealFusion: 360 Reconstruction of Any Object from a Single Image. In *IEEE CVPR*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, 405–421.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4): 102:1–102:15.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. abs/2212.08751.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S. M.; Geiger, A.; and Radwan, N. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *IEEE CVPR*, 5470–5480.

Pan, X.; Dai, B.; Liu, Z.; Loy, C. C.; and Luo, P. 2021. Do 2D GANs Know 3D Shape? Unsupervised 3D Shape Reconstruction from 2D Image GANs. In *ICLR*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In *ICML*, volume 139, 8821–8831.

Reizenstein, J.; Shapovalov, R.; Henzler, P.; Sbordone, L.; Labatut, P.; and Novotny, D. 2021. Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. In *ICCV*.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense Depth Priors for Neural Radiance Fields from Sparse Input Views. In *IEEE CVPR*, 12882–12891.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE CVPR*, 10674–10685.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Lopes, R. G.; Ayan, B. K.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Sajjadi, M. S. M.; Meyer, H.; Pot, E.; Bergmann, U.; Greff, K.; Radwan, N.; Vora, S.; Lucic, M.; Duckworth, D.; Dosovitskiy, A.; Uszkoreit, J.; Funkhouser, T. A.; and Tagliasacchi, A. 2022. Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations. In *IEEE CVPR*, 6219–6228.

Schönberger, J. L.; and Frahm, J. 2016. Structure-from-Motion Revisited. In *IEEE CVPR*, 4104–4113.

Schönberger, J. L.; Zheng, E.; Frahm, J.; and Pollefeys, M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*, volume 9907, 501–518.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Seo, J.; Jang, W.; Kwak, M.; Ko, J.; Kim, H.; Kim, J.; Kim, J.; Lee, J.; and Kim, S. 2023. Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation. abs/2303.07937.

Suhail, M.; Esteves, C.; Sigal, L.; and Makadia, A. 2022a. Generalizable Patch-Based Neural Rendering. In *ECCV*, volume 13692, 156–174.

Suhail, M.; Esteves, C.; Sigal, L.; and Makadia, A. 2022b. Light Field Neural Rendering. In *IEEE CVPR*, 8259–8269.

Tang, J.; Wang, T.; Zhang, B.; Zhang, T.; Yi, R.; Ma, L.; and Chen, D. 2023. Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior. *arXiv preprint arXiv:2303.14184*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *NeurIPS*, 6306–6315.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022. CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. In *IEEE CVPR*, 3825–3834.

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2023a. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. *IEEE CVPR*.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021a. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *NeurIPS*, 27171–27183.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. A. 2021b. IBRNet: Learning Multi-View Image-Based Rendering. In *IEEE CVPR*, 4690–4699.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *CoRR*, abs/2305.16213.

Yang, J.; Pavone, M.; and Wang, Y. 2023. FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization. In *IEEE CVPR*.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In *ECCV*, 785–801.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume Rendering of Neural Implicit Surfaces. In *NeurIPS*.

Yoo, P.; Guo, J.; Matsuo, Y.; and Gu, S. S. 2023. DreamSparse: Escaping from Plato's Cave with 2D Frozen Diffusion Model Given Sparse Views. *CoRR*, abs/2306.03414.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelNeRF: Neural Radiance Fields From One or Few Images. In *IEEE CVPR*, 4578–4587.

Yu, C.; Zhou, Q.; Li, J.; Zhang, Z.; Wang, Z.; and Wang, F. 2023. Points-to-3D: Bridging the Gap between Sparse Points and Shape-Controllable Text-to-3D Generation. abs/2307.13908.

Yu, Z.; and Gao, S. 2020. Fast-MVSNet: Sparse-to-Dense Multi-View Stereo With Learned Propagation and Gauss-Newton Refinement. In *IEEE CVPR*, 1946–1955.

Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. In *NeurIPS*.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE CVPR*, 586–595.

Zhou, Z.; and Tulsiani, S. 2023. SparseFusion: Distilling View-conditioned Diffusion for 3D Reconstruction. In *CVPR*.