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

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

[1]Stanford University, [2]Google Research

Abstract

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

1. Introduction

Models for single-image, 360-degree novel view synthesis (NVS) should produce *realistic* and *diverse* results: the synthesized images should look natural and 3D-consistent to humans, and they should also capture the many possible explanations of unobservable regions. This challenging problem has typically been studied in the context of single objects without backgrounds, where the requirements on both realism and diversity are simplified. Recent progress relies on large 3D datasets like Objaverse-XL [8] which have enabled training conditional diffusion [18] models to perform photorealistic and 3D-consistent NVS via Score Distillation Sampling [SDS; 26]. Meanwhile, since image diversity mostly lies in the background, not the object, the ignorance of background significantly lowers the expectation of synthesizing diverse images–in fact, most object-centric methods do not consider diversity metrics [18, 21, 27].

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

We address both issues with our new model, ZeroNVS. Inspired by previous object-centric methods [18, 21, 27], ZeroNVS also trains a 2D conditional diffusion model followed by 3D distillation. But unlike them, ZeroNVS works well on scenes due to two technical innovations: a new camera parametrization and normalization scheme for conditioning, which allows training the diffusion model on diverse scene datasets, and an "SDS anchoring" mechanism, improving the background diversity over standard SDS.

To overcome the key challenge of limited training data, we propose training the diffusion model on a massive mixed dataset comprised of all scenes from CO3D [30], RealEstate10K [44], and ACID [16], so that the model may potentially handle complex in-the-wild scenes. The mixed data of such scale and diversity are captured with a variety of camera settings and have several different types of 3D ground truth, e.g., computed with COLMAP [32] or ORB-SLAM [23]. We show that while the camera conditioning representations from prior methods [18] are too ambiguous or inexpressive to model in-the-wild scenes, our new camera parametrization and normalization scheme allows exploiting such diverse data sources and leads to superior NVS on real-world scenes.

Building a 2D conditional diffusion model that works effectively for in-the-wild scenes enables us to then study the limitations of SDS in the scene setting. In particular, we observe limited diversity from SDS in the generated scene backgrounds when synthesizing long-range (e.g., 180-degree) novel views. We therefore propose "SDS anchoring" to ameliorate the issue. In SDS anchoring, we propose to first sample several "anchor" novel views using the standard Denoising Diffusion Implicit Model (DDIM) sampling [36]. This yields a collection of pseudo-ground-truth novel views with diverse contents, since DDIM is not prone

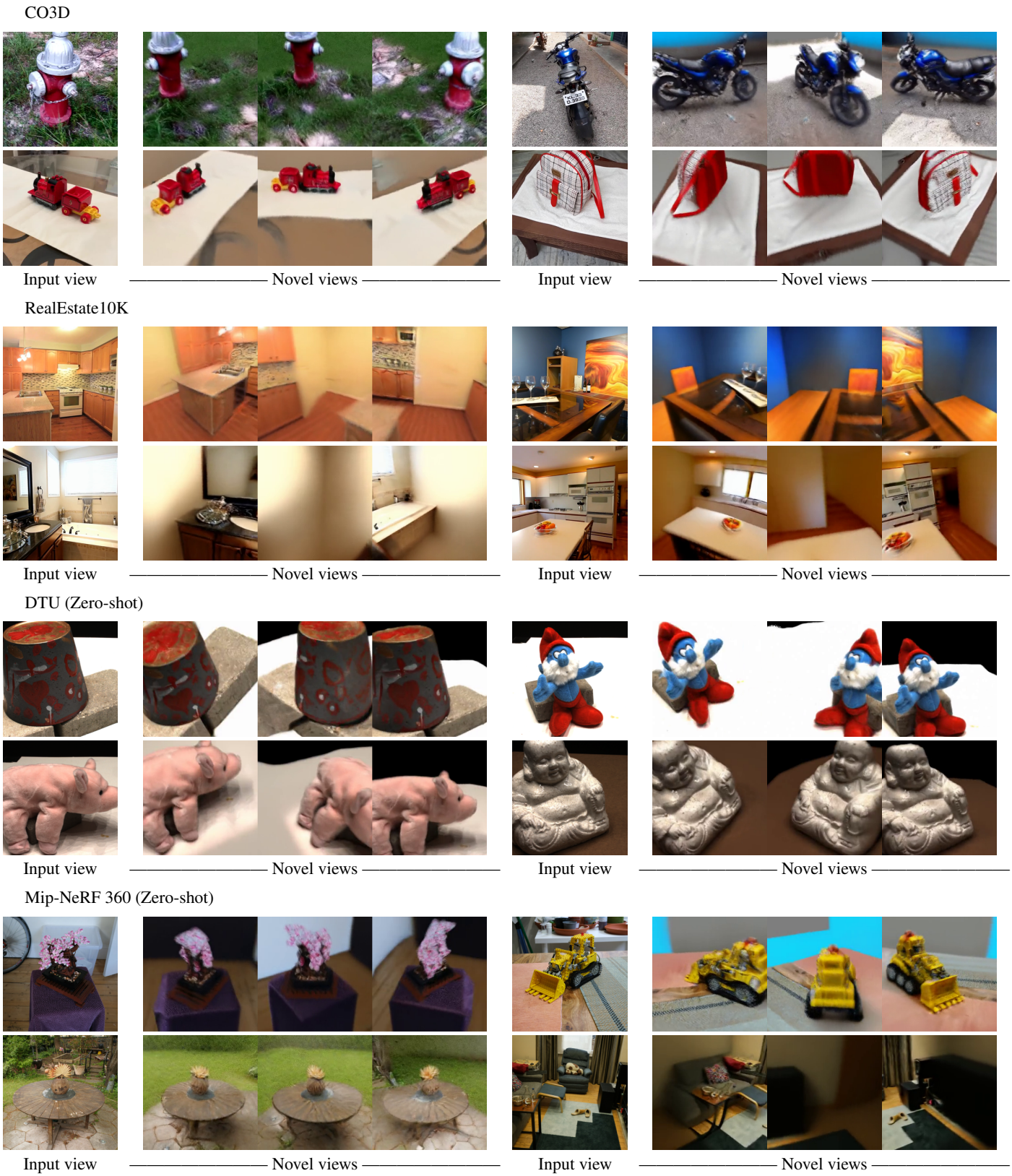

Figure 1. Results for view synthesis from a single image. All NeRFs are predicted by the *same* model.

to mode collapse like SDS. Then, rather than using these views as RGB supervision, we sample from them randomly as conditions for SDS, which enforces diversity while still ensuring 3D-consistent view synthesis.

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

To summarize, we make the following contributions:

- We propose ZeroNVS, which enables full-scene NVS from real images. ZeroNVS first demonstrates that SDS distillation can be used to lift scenes that are not object-centric and may have complex backgrounds to 3D.
- We show that the formulations on handling cameras and scene scale in prior work are either inexpressive or ambiguous for in-the-wild scenes. We propose a new camera conditioning parameterization and a scene normalization scheme. These enable us to train a single model on a large collection of diverse training data consisting of CO3D, RealEstate10K and ACID, allowing strong zero-shot generalization for NVS on in-the-wild images.
- We study the limitations of SDS distillation as applied to scenes. Similar to prior work, we identify a diversity issue, which manifests in this case as novel view predictions with monotone backgrounds. We propose SDS anchoring to ameliorate the issue.
- We show state-of-the-art LPIPS results on DTU *zero-shot*, surpassing prior methods finetuned on this dataset. Furthermore, we introduce the Mip-NeRF 360 dataset as a scene-level single-image novel view synthesis benchmark and analyze the performances of our and other methods. Finally, we show that our proposed SDS anchoring is preferred via a user study.

2. Related Work

3D generation. DreamFusion [26] proposed Score Distillation Sampling (SDS) as a way of leveraging a diffusion model to extract a NeRF given a user-provided text prompt. After DreamFusion, follow-up works such as Magic3D [15], ATT3D [20], ProlificDreamer [38], and Fantasia3D [6] improved the quality, diversity, resolution, or run-time. Other types of 3D generative models include GAN-based 3D generative models, which are primarily restricted to single object categories [3, 4, 12, 24, 25, 33] or to synthetic data [11]. Recently, 3DGP [34] adapted the GAN approach to train 3D generative models on ImageNet. VQ3D [31] and IVID [40] leveraged vector quantization and diffusion, respectively, to learn generative models on ImageNet. One critical critical challenge for scene-based 3D-aware methods 360-degree viewpoint change. Both VQ3D and 3DGP demonstrate only limited camera motion, while IVID generally focuses on small camera motion but can achieve 360-degree views for a small subset of scenes.

Single-image novel view synthesis. PixelNeRF [43] and DietNeRF [14] learn to infer NeRFs from sparse views via training an image-based 3D feature extractor or semantic consistency losses, respectively. However, these approaches do not produce renderings resembling crisp natural images from a single image. Several recent diffusion-based approaches achieve high quality results by separating the problem into two stages. First, a (potentially 3D-aware) diffusion model is trained, and second, the diffusion model is used to distill 3D-consistent scene representations given an input image via techniques like score distillation sampling [26], score Jacobian chaining [37], textual inversion or semantic guidance leveraging the diffusion model [9, 21], or explicit 3D reconstruction from multiple sampled views of the diffusion model [17, 19]. Another diffusion-based work, GeNVS [5], achieves 360 camera motion but only for specific object categories such as fire hydrants. By contrast, ZeroNVS generates 360-degree camera motion by default for a variety of scene categories. This is enabled by innovations such as novel camera conditioning representations and SDS anchoring, which enable us to train on massive real scene datasets and then to perform scene-level NVS with up to 360-degree viewpoint change on diverse scene types.

3. Approach

We consider the problem of scene-level novel view synthesis from a single real image. Similar to prior work [18, 27], we first train a diffusion model $\mathbf{p}_\theta$ to perform novel view synthesis, and then leverage it to perform 3D SDS distillation. Unlike prior work, we focus on scenes rather than objects. Scenes present several unique challenges. First, prior works use representations for cameras and scale which are either ambiguous or insufficiently expressive for scenes. Second, the inference procedure of prior works is based on SDS, which has a known mode collapse issue and which manifests in scenes through greatly reduced background diversity in predicted views. We will attempt to address these challenges through improved representations and inference procedures for scenes compared with prior work [18, 27].

We shall begin by introducing some general notation. Let a scene S consist of a set of images $X = \{X_i\}_{i=1}^n$, depth maps $D = \{D_i\}_{i=1}^n$, extrinsics $E = \{E_i\}_{i=1}^n$, and a shared field-of-view f. We note that an extrinsics matrix E_i can be identified with its rotation and translation components, defined by $E_i = (E_i^R, E_i^T)$. We preprocess our data

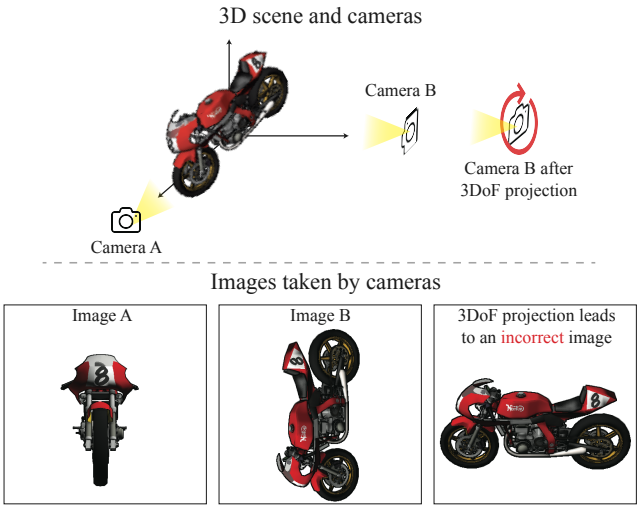

Figure 2. A 3DoF camera pose captures elevation, azimuth, and radius but is incapable of representing a camera's roll (pictured) or cameras positioned and oriented arbitrarily in space.

to consist of square images and assume intrinsics are shared within a given scene, and that there is no skew, distortion, or off-center principal point.

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

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

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

3.1. Representing Objects for View Synthesis

Zero-1-to-3 [18] represents poses with 3 degrees of freedom, given by an elevation θ, azimuth ϕ, and radius z. Let $\mathbf{P} : \mathrm{SE}(3) \to \mathbb{R}^3$ be the projection to (θ, ϕ, z), then

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

is the camera conditioning representation used by Zero-1-to-3. For object mesh datasets such as Objaverse [7] and Objaverse-XL [8], this representation is appropriate because the data is known to consist of single objects without

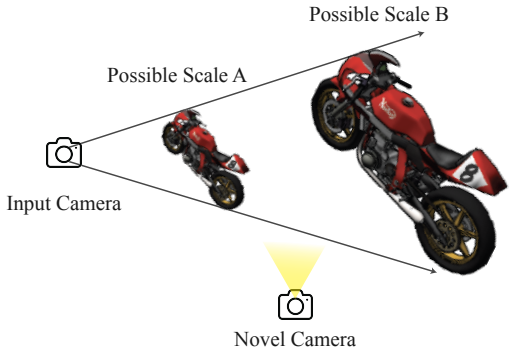

Figure 3. To a monocular camera, a small object close to the camera (left) and a large object at a distance (right) appear identical, despite representing different scenes. Scale ambiguity in input view leads to multiple plausible novel views.

backgrounds, aligned and centered at the origin and imaged from training cameras generated with three degrees of freedom. However, such a parameterization limits the model's ability to generalize to non-object-centric images, and to real-world data. In real-world data, poses can have six degrees of freedom, incorporating both rotation (pitch, roll, yaw) and 3D translation. An illustration of a failure of the 3DoF camera representation is shown in Figure 2.

3.2. Representing Scenes for View Synthesis

For scenes, we should use a camera representation with six degrees of freedom that can capture all possible positions and orientations. One straightforward choice is the relative pose parameterization [39]. We propose to also include the field of view as an additional degree of freedom. We term this combined representation "6DoF+1". This gives us

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

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

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

This is useful given the arbitrary nature of the poses for our datasets which are determined by COLMAP or ORB-SLAM. The poses discovered via these algorithms are not related to any meaningful alignment of the scene, such as a rigid transformation and scale transformation which align the scene to some canonical frame and unit of scale. Although we have seen that $\mathbf{M}_{\mathrm{6DoF+1}}$ is invariant to rigid transformations of the scene, it is not invariant to scale. The scene scales determined by COLMAP and ORB-SLAM are also arbitrary and may vary significantly. One solution is to directly normalize the camera locations. Let $\mathbf{R}(E, \lambda) : \mathrm{SE}(3) \times \mathbb{R} \to \mathrm{SE}(3)$ be a function that scales the translation

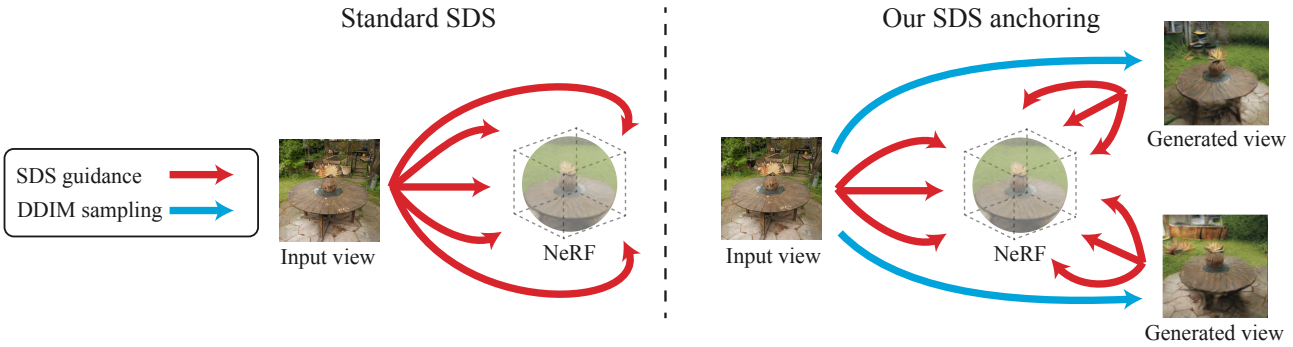

Figure 4. SDS-based NeRF distillation (left) uses the same guidance image for all 360 degrees of novel views. Our "SDS anchoring" (right) first samples novel views via DDIM [35], and then uses the nearest image (whether the input or a sampled novel view) for guidance.

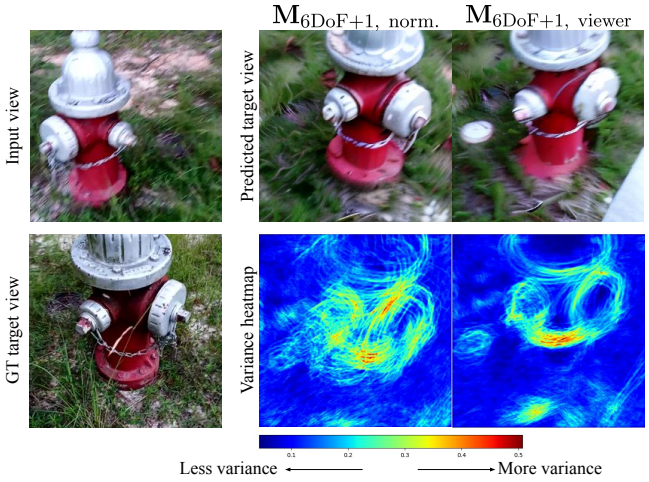

Figure 5. Samples and variance heatmaps of the Sobel edges of multiple samples from ZeroNVS. $\mathbf{M}_{\text{6DoF+1, viewer}}$ reduces randomness from scale ambiguity.

component of E by λ. Then we define

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

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

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

3.3. Addressing Scale Ambiguity with a New Normalization Scheme

The representation $\mathbf{M}_{\text{6DoF+1, norm.}}$ achieves reasonable performance on real scenes by addressing issues in prior representations with limited degrees of freedom and handling of scale. However, a normalization scheme that better addresses scale ambiguity may lead to improved performance. Scene scale is ambiguous given a monocular input image [29, 42]. This complicates NVS, as we illustrate

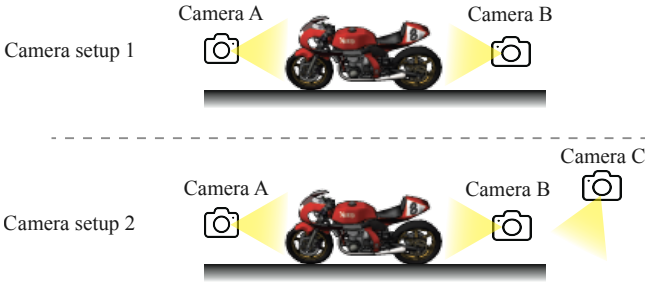

Figure 6. Top: A scene with two cameras facing the object. Bottom: The same scene with a new camera added facing the ground. Addition of Camera C under $\mathbf{M}_{\text{6DoF+1, agg.}}$ drastically changes the scene's scale. $\mathbf{M}_{\text{6DoF+1, viewer.}}$ avoids this.

in Figure 3. We therefore choose to introduce information about the scale of the visible content to our conditioning embedding function $\mathbf{M}$. Rather than normalize by camera locations, Stereo Magnification [44] takes the 5-th quantile of each depth map of the scene, and then takes the 10-th quantile of this aggregated set of numbers, and declares this as the scene scale. Let $\mathbf{Q}_k$ be a function which takes the k-th quantile of a set of numbers, then we define

$$q = \mathbf{Q}_{10}(\{\mathbf{Q}_5(D_i)\}_{i=1}^{n}) \,,$$

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

where in $\mathbf{M}_{\text{6DoF+1, agg.}}$, q is the scale applied to the translation component of the scene's cameras before computing the relative pose. In this way $\mathbf{M}_{\text{6DoF+1, agg.}}$ is different from $\mathbf{M}_{\text{6DoF+1, norm.}}$ because the camera conditioning representation contains information about the scale of the visible content from the depth maps D_i. Although conditioning with $\mathbf{M}_{\text{6DoF+1, agg.}}$ improves performance, there are two issues. The first arises from aggregating the quantiles over all the images. In Figure 6, adding an additional Camera C to the scene changes the value of $\mathbf{M}_{\text{6DoF+1, agg.}}$ despite nothing else having changed about the scene. This makes the view synthesis task from either Camera A or Camera B more ambiguous. To ensure this is impossible, we can simply eliminate the aggregation step over the quantiles of

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

Table 1. **Comparison with the state of the art.** We set a new state-of-the-art for LPIPS on DTU despite being the only method not fine-tuned on DTU. † =Performance reported in Xu et al. [41].

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

Table 2. **Zero-shot comparison**. Comparison with baselines re-trained on our mixture dataset. ZeroNVS outperforms Zero-1-to-3 even when Zero-1-to-3 is trained on the same scene data. Extensive video comparisons are in the supplementary.

all depth maps in the scene. The second issue arises from different depth statistics within the mixture of datasets we use for training. ORB-SLAM generally produces sparser depth maps than COLMAP, and therefore the value of $\mathbf{Q}_k$ may have different meanings for each. We therefore use an off-the-shelf depth estimator [28] to fill holes in the depth maps. We denote the depth D_i infilled in this way as $\bar{D}_i$. We then apply $\mathbf{Q}_k$ to dense depth maps $\bar{D}_i$ instead. We emphasize that the depth estimator is *not* used during inference or distillation. Its purpose is only for the model to learn a consistent definition of scale during training. These two fixes lead to our proposed normalization, which is fully viewer-centric. We define it as

$$q_i = \mathbf{Q}_{20}(\bar{D}_i)\ ,$$
$$\mathbf{M}_{\text{6DoF+1, viewer}}(D, f, E, i, j) = \left[\mathbf{R}\left(E_i^{-1}E_j, \frac{1}{q_i}\right), f\right]\ ,$$

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

3.4. Improving Diversity with SDS Anchoring

Diffusion models trained with the improved camera conditioning representation $\mathbf{M}_{\text{6DoF+1, viewer}}$ achieve superior

NVS on DTU	LPIPS ↓	PSNR ↑	SSIM ↑
All datasets	**0.421**	**12.2**	**0.444**
-ACID	0.446	11.5	0.405
-CO3D	0.456	10.7	0.407
-RealEstate10K	0.435	12.0	0.429

Table 3. **Ablation study on training data.** Training on all datasets improves performance.

view synthesis results via 3D SDS distillation. However, for large viewpoint changes, novel view synthesis is also a generation problem, and it may be desirable to generate diverse and plausible contents rather than contents that are only optimal on average for metrics such as PSNR, SSIM, and LPIPS. However, Poole et al. [26] noted that even when the underlying generative model produces diverse images, SDS distillation of that model tends to seek a single mode. For novel view synthesis of scenes via SDS, we observe a unique manifestation of this diversity issue: lack of diversity is especially apparent in inferred backgrounds. Often, SDS distillation predicts a gray or monotone background for regions not observed by the input camera.

To remedy this, we propose "SDS anchoring" (Figure 4). With SDS anchoring, we first directly sample k novel views $\hat{\boldsymbol{X}}_k = \{\hat{X}_j\}_{j=1}^k$ with $\hat{X}_j \sim p(X_j|X_i, \mathbf{M}(D, f, E, i, j))$ from poses evenly spaced in azimuth for maximum scene coverage. We sample the novel views via DDIM [35], which does not have the mode collapse issues of SDS. Each novel view is generated conditional on the input view. Then, when optimizing the SDS objective, we condition the diffusion model not on the input view, but on the nearest view. As shown quantitatively in a user study in Section 4 and qualitatively in Figure 9, SDS anchoring produces more diverse background contents. We provide more details about the setup of SDS anchoring in the supplementary.

4. Experiments

4.1. Setup

Datasets. Our models are trained on a mixture dataset consisting of CO3D [30], ACID [16], and RealEstate10K [44]. Each example is sampled uniformly at random from the three datasets. We train at 256×256 resolution, center-cropping and adjusting the intrinsics for each image and scene as necessary. We train using our representation $\mathbf{M}_{\text{6DoF+1, viewer}}$ unless otherwise specified. We provide more training details in the supplementary.

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

Input view | GT novel view | ZeroNVS (ours) PSNR=10.8, SSIM=0.22 | PixelNeRF PSNR=12.2, SSIM=0.30

Figure 7. Limitations of PSNR and SSIM for view synthesis evaluation. Misalignments can lead to worse PSNR and SSIM values for predictions that are more semantically sensible.

GT novel view | Zero-1-to-3 | NerDi | ZeroNVS (ours)

Figure 8. Qualitative comparison between baseline methods and our method.

400×300 resolution to be comparable to prior art.

Implementation details. Our diffusion model training code is written in PyTorch and based on the public code for Zero-1-to-3 [18]. We initialize from the pretrained Zero-1-to-3-XL, swapping out the conditioning module to accommodate our novel parameterizations. Our distillation code is implemented in Threestudio [13]. We use a custom NeRF network combining features of Mip-NeRF 360 with Instant-NGP [22]. The noise schedule is annealed following Wang et al. [38]. For details please see the supplementary.

4.2. Main Results

We evaluate all methods using the standard set of novel view synthesis metrics: PSNR, SSIM, and LPIPS. We weigh LPIPS more heavily in the comparison due to the well-known issues with PSNR and SSIM as discussed in [5, 9]. We confirm that PSNR and SSIM do not correlate well with performance in our setting, as illustrated in Figure 7.

The results are shown in Table 1. We first compare against baseline methods DS-NeRF [10], PixelNeRF [43], SinNeRF [41], DietNeRF [14], and NeRDi [9] on DTU. All these methods are trained on DTU, but we achieve a state-of-the-art LPIPS despite being fully zero-shot. We show visual comparisons in Figure 8.

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

Limited diversity is a known issue with SDS-based methods, but the long run time of SDS-based methods makes typical generation-based metrics such as FID cost-prohibitive. Therefore, we quantify the improved diversity

Figure 9. Whereas standard SDS (left) tends to predict monotonous backgrounds, our SDS anchoring (right) generates more diverse background contents. Additionally, SDS anchoring generates noticeably different results depending on the seed.

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

Table 4. **Ablation study on the conditioning representation M.** Our $\mathbf{M}_{\text{6DoF+1, viewer}}$ matches or outperforms other representations.

from SDS anchoring via a user study of 21 users on the Mip-NeRF 360 dataset. Users were asked to compare scenes predicted with and without SDS anchoring along three dimensions: Realism, Creativity, and Overall Preference. The preferences for SDS anchoring were: Realism (78%), Creativity (82%), and Overall Preference (80%). The supplementary provides more details about the setup of the study. Figure 9 includes qualitative examples that show the advantages of SDS anchoring, and the supplementary webpage contains the videos which were shown in the study.

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

5. Conclusion

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

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

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. 6

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 3, 6

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

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2021. 3

[5] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *ICCV*, 2023. 3, 7

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

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. *arXiv preprint arXiv:2212.08051*, 2022. 4

[8] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. *arXiv preprint arXiv:2307.05663*, 2023. 1, 4

[9] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. NeRDi: Single-view NeRF synthesis with language-guided diffusion as general image priors. In *CVPR*, 2022. 3, 7

[10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022. 7

[11] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. GET3D: A generative model of high quality 3D textured shapes learned from images. In *NeurIPS*, 2022. 3

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

[13] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3D content generation, 2023. 7

[14] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021. 3, 7

[15] 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. 3

[16] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021. 1, 6

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

[18] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *CVPR*, 2023. 1, 3, 4, 7

[19] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

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

[21] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. RealFusion: 360° reconstruction of any object from a single image. In *CVPR*, 2023. 1, 3

[22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 7

[23] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 1

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

[25] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 3

[26] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2022. 1, 3, 6

[27] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 1, 3

[28] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 6

[29] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(3), 2022. 5

[30] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*, 2021. 1, 6

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

[32] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

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

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

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv::2010.02502*, 2020. 5, 6

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

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

[38] 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. 3, 7

[39] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 4

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

[41] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. SinNeRF: Training neural radiance fields on complex scenes from a single image. In *ECCV*, 2022. 6, 7

[42] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3D scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 5

[43] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 3, 7

[44] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 1, 5, 6