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

Ruoshi Liu[1] Rundi Wu[1] Basile Van Hoorick[1] Pavel Tokmakov[2] Sergey Zakharov[2] Carl Vondrick[1]
[1] Columbia University [2] Toyota Research Institute
zero123.cs.columbia.edu

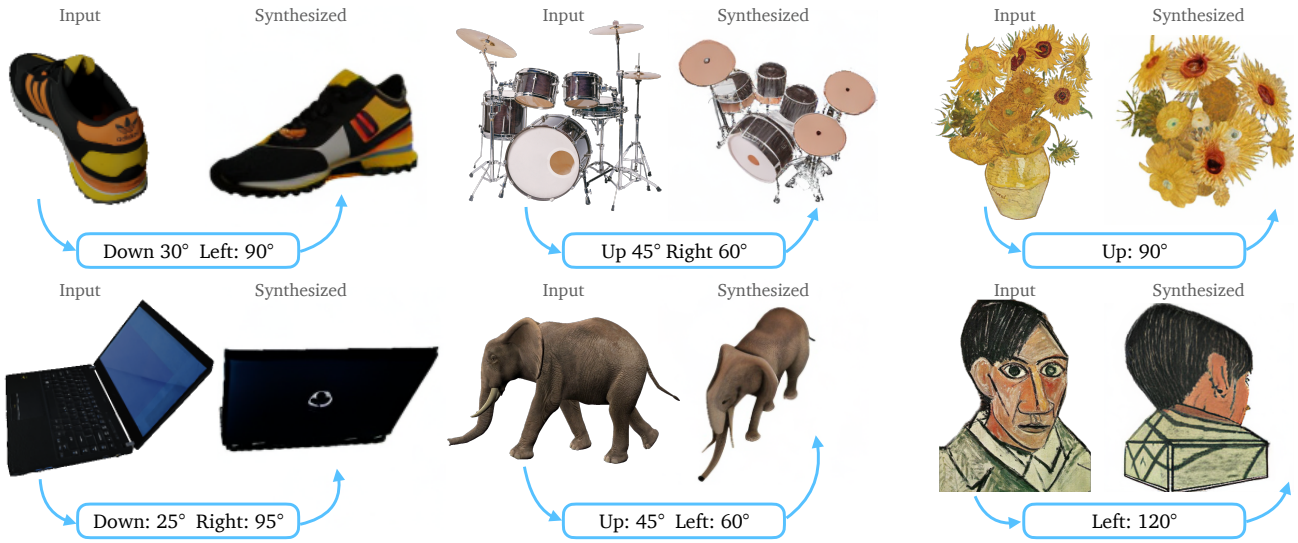

Figure 1: Given a single RGB image of an object, we present ***Zero-1-to-3***, a method to synthesize an image from a specified camera viewpoint. Our approach synthesizes views that contain rich details consistent with the input view for large relative transformations. It also achieves strong zero-shot performance on objects with complex geometry and artistic styles.

Abstract

We introduce Zero-1-to-3, a framework for changing the camera viewpoint of an object given just a single RGB image. To perform novel view synthesis in this under-constrained setting, we capitalize on the geometric priors that large-scale diffusion models learn about natural images. Our conditional diffusion model uses a synthetic dataset to learn controls of the relative camera viewpoint, which allow new images to be generated of the same object under a specified camera transformation. Even though it is trained on a synthetic dataset, our model retains a strong zero-shot generalization ability to out-of-distribution datasets as well as in-the-wild images, including impressionist paintings. Our viewpoint-conditioned diffusion approach can further be used for the task of 3D reconstruction from a single image. Qualitative and quantitative experiments show that our method significantly outperforms state-of-the-art single-view 3D reconstruction and novel view synthesis models by leveraging Internet-scale pre-training.

1. Introduction

From just a single camera view, humans are often able to imagine an object's 3D shape and appearance. This ability is important for everyday tasks, such as object manipulation [17] and navigation in complex environments [7], but is also key for visual creativity, such as painting [32]. While this ability can be partially explained by reliance on geometric priors like symmetry, we seem to be able to generalize to much more challenging objects that break physical and geometric constraints with ease. In fact, we can predict the 3D shape of objects that do not (or even *cannot*) exist in the physical world (see third column in Figure 1). To achieve this degree of generalization, humans rely on prior knowledge accumulated through a lifetime of visual exploration.

In contrast, most existing approaches for 3D image reconstruction operate in a closed-world setting due to their reliance on expensive 3D annotations (e.g. CAD models) or category-specific priors [37, 21, 36, 67, 68, 66, 25, 24]. Very recently, several methods have made major strides in the di-

rection of open-world 3D reconstruction by pre-training on large-scale, diverse datasets such as CO3D [43, 30, 36, 15]. However, these approaches often still require geometry-related information for training, such as stereo views or camera poses. As a result, the scale and diversity of the data they use remain insignificant compared to the recent Internet-scale text-image collections [47] that enable the success of large diffusion models [45, 44, 33]. It has been shown that Internet-scale pre-training endows these models with rich semantic priors, but the extent to which they capture geometric information remains largely unexplored.

In this paper, we demonstrate that we are able to learn control mechanisms that manipulate the camera viewpoint in large-scale diffusion models, such as Stable Diffusion [44], in order to perform zero-shot novel view synthesis and 3D shape reconstruction. Given a single RGB image, both of these tasks are severely under-constrained. However, due to the scale of training data available to modern generative models (over 5 billion images), diffusion models are state-of-the-art representations for the natural image distribution, with support that covers a vast number of objects from many viewpoints. Although they are trained on 2D monocular images without any camera correspondences, we can fine-tune the model to learn controls for relative camera rotation and translation during the generation process. These controls allow us to encode arbitrary images that are decoded to a different camera viewpoint of our choosing. Figure 1 shows several examples of our results.

The primary contribution of this paper is to demonstrate that large diffusion models have learned rich 3D priors about the visual world, even though they are only trained on 2D images. We also demonstrate state-of-the-art results for novel view synthesis and state-of-the-art results for zero-shot 3D reconstruction of objects, both from a single RGB image. We begin by briefly reviewing related work in Section 2. In Section 3, we describe our approach to learn controls for camera extrinsics by fine-tuning large diffusion models. Finally, in Section 4, we present several quantitative and qualitative experiments to evaluate zero-shot view synthesis and 3D reconstruction of geometry and appearance from a single image. We will release all code and models as well as an online demo.

2. Related Work

3D generative models. Recent advancements in generative image architectures combined with large scale image-text datasets [47] have made it possible to synthesize high-fidelity of diverse scenes and objects [33, 40, 45]. In particular, diffusion models have shown to be very effective at learning scalable image generators using a denoising objective [6, 48]. However, scaling them to the 3D domain would require large amounts of expensive annotated 3D data. Instead, recent approaches rely on transferring pre-trained large-scale 2D diffusion models to 3D without using any ground truth 3D data. Neural Radiance Fields or NeRFs [31] have emerged as a powerful representation, thanks to their ability to encode scenes with high fidelity. Typically, NeRF is used for single-scene reconstruction, where many posed images covering the entire scene are provided. The task is then to predict novel views from unobserved angles. DreamFields [22] has shown that NeRF is a more versatile tool that can also be used as the main component in a 3D generative system. Various follow-up works [38, 26, 53] substitute CLIP for a distillation loss from a 2D diffusion model that is repurposed to generate high-fidelity 3D objects and scenes from text inputs.

Figure 2: **Viewpoint bias in text-to-image models.** We show samples from both Dall-E-2 and Stable Diffusion v2 from the prompt *"a chair"*. Most samples show a chair in a forward-facing canonical pose.

Our work explores an unconventional approach to novel-view synthesis, modeling it as a viewpoint-conditioned image-to-image translation task with diffusion models. The learned model can also be combined with 3D distillation to reconstruct 3D shape from a single image. Prior work [56] adopted a similar pipeline but did not demonstrate zero-shot generalization capability. Concurrent approaches [9, 29, 61] proposed similar techniques to perform image-to-3D generation using language-guided priors and textual inversion [14]. In comparison, our method learns control of viewpoints through a synthetic dataset and demonstrates zero-shot generalization to in-the-wild images.

Single-view object reconstruction. Reconstructing 3D objects from a single view is a highly challenging problem that requires strong priors. One line of work builds priors from relying on collections of 3D primitives represented as meshes [58, 62], voxels [16, 60], or point clouds [12, 30], and use image encoders for conditioning. These models are constrained by the variety of the used 3D data collection and show poor generalization capabilities due to the global nature of this type of conditioning. Moreover, they require an additional pose estimation step to ensure alignment between the estimated shape and the input. On the other hand, locally conditioned models [46, 63, 54, 51, 52] aim

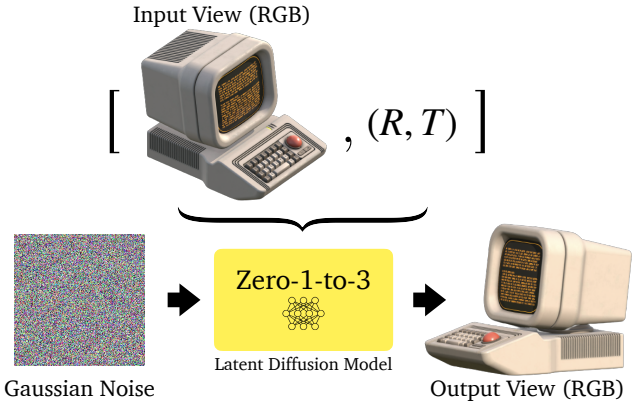

Figure 3: **Zero-1-to-3** is a viewpoint-conditioned image translation model using a conditional latent diffusion architecture. Both the input view and a relative viewpoint transformation are used as conditional information.

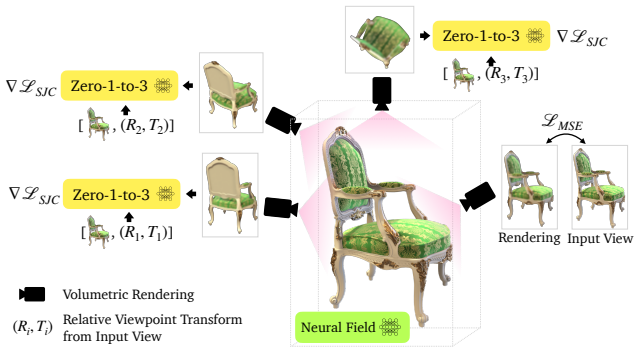

Figure 4: **3D reconstruction with Zero-1-to-3.** Zero-1-to-3 can be used to optimize a neural field for the task of 3D reconstruction from a single image. During training, we randomly sample viewpoints and use Zero-1-to-3 to supervise the 3D reconstruction.

to use local image features directly for scene reconstruction and show greater cross-domain generalization capabilities, though are generally limited to close-by view reconstructions. Recently, MCC [59] learns a general-purpose representation for 3D reconstruction from RGB-D views and is trained on large-scale dataset of object-centric videos.

In our work, we demonstrate that rich geometric information can be extracted directly from a pre-trained Stable Diffusion model, alleviating the need for additional depth information.

3. Method

Given a single RGB image $x \in \mathbb{R}^{H \times W \times 3}$ of an object, our goal is to synthesize an image of the object from a different camera viewpoint. Let $R \in \mathbb{R}^{3\times3}$ and $T \in \mathbb{R}^3$ be the relative camera rotation and translation of the desired viewpoint, respectively. We aim to learn a model f that synthesizes a new image under this camera transformation:

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

where we denote $\hat{x}_{R,T}$ as the synthesized image. We want our estimated $\hat{x}_{R,T}$ to be perceptually similar to the true but unobserved novel view $x_{R,T}$.

Novel view synthesis from monocular RGB image is severely under-constrained. Our approach will capitalize on large diffusion models, such as Stable Diffusion, in order to perform this task, since they show extraordinary zero-shot abilities when generating diverse images from text descriptions. Due to the scale of their training data [47], pre-trained diffusion models are state-of-the-art representations for the natural image distribution today.

However, there are two challenges that we must overcome to create f. Firstly, although large-scale generative models are trained on a large variety of objects in different viewpoints, the representations do not explicitly encode the correspondences between viewpoints. Secondly, generative models inherit viewpoint biases reflected on the Internet. As shown in Figure 2, Stable Diffusion tends to generate images of forward-facing chairs in canonical poses. These two problems greatly hinder the ability to extract 3D knowledge from large-scale diffusion models.

3.1. Learning to Control Camera Viewpoint

Since diffusion models have been trained on internet-scale data, their support of the natural image distribution likely covers most viewpoints for most objects, but these viewpoints cannot be controlled in the pre-trained models. Once we are able to *teach* the model a mechanism to control the camera extrinsics with which a photo is captured, then we unlock the ability to perform novel view synthesis.

To this end, given a dataset of paired images and their relative camera extrinsics $\{(x, x_{(R,T)}, R, T)\}$, our approach, shown in Figure 3, fine-tunes a pre-trained diffusion model in order to learn controls over the camera parameters without destroying the rest of the representation. Following [44], we use a latent diffusion architecture with an encoder $\mathcal{E}$, a denoiser U-Net ϵ_θ, and a decoder $\mathcal{D}$. At the diffusion time step $t \sim [1, 1000]$, let $c(x, R, T)$ be the embedding of the input view and relative camera extrinsics. We then solve for the following objective to fine-tune the model:

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

After the model ϵ_θ is trained, the inference model f can generate an image by performing iterative denoising from a Gaussian noise image [44] conditioned on $c(x, R, T)$.

The main result of this paper is that fine-tuning pre-trained diffusion models in this way enables them to learn a

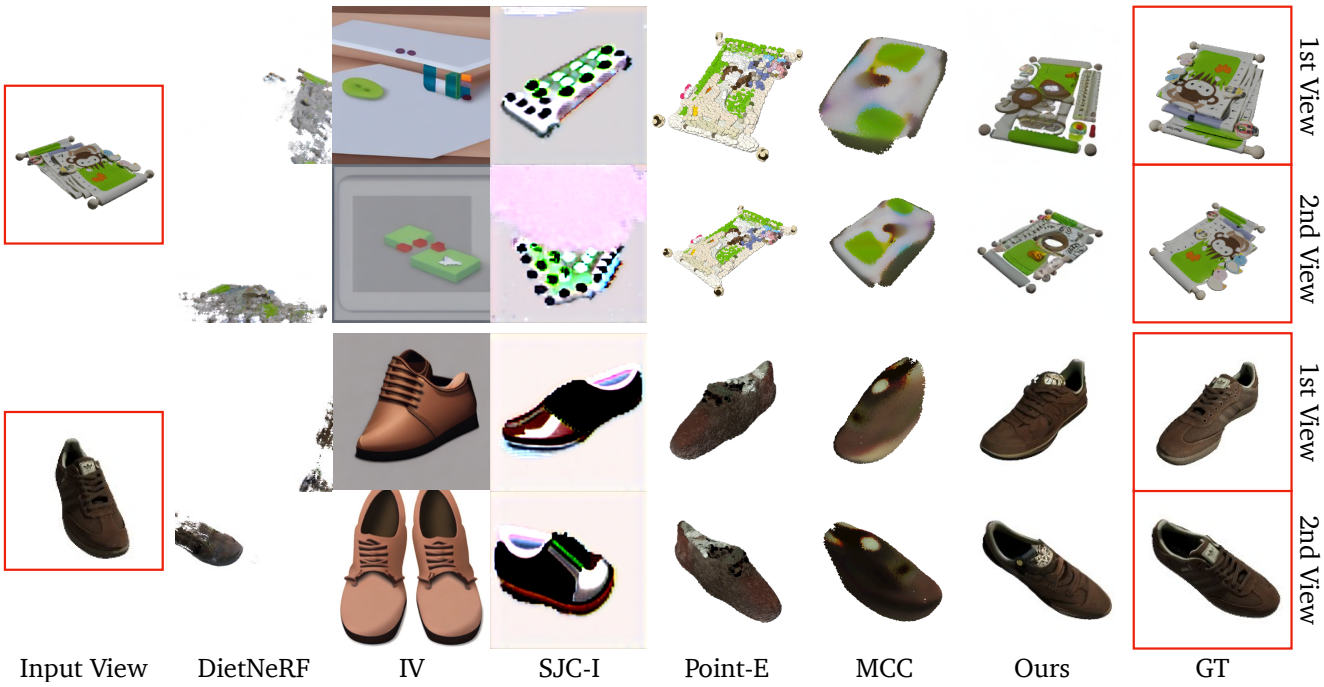

Figure 5: **Novel view synthesis on Google Scanned Objects [10].** The input view shown on the left is used to synthesize two randomly sampled novel views. Corresponding ground truth views are shown on the right. Compared to the baselines, our synthesized novel view contain rich textual and geometric details that are highly consistent with the ground truth, while baseline methods display a significant loss of high-frequency details.

generic mechanism for controlling the camera viewpoints, which extrapolates outside of the objects seen in the fine-tuning dataset. In other words, this fine-tuning allows controls to be "bolted on" and the diffusion model can retain the ability to generate photorealistic images, except now with control of viewpoints. This compositionality establishes zero-shot capabilities in the model, where the final model can synthesize new views for object classes that lack 3D assets and never appear in the fine-tuning set.

3.2. View-Conditioned Diffusion

3D reconstruction from a single image requires both low-level perception (depth, shading, texture, etc.) and high-level understanding (type, function, structure, etc.). Therefore, we adopt a hybrid conditioning mechanism. On one stream, a CLIP [39] embedding of the input image is concatenated with (R, T) to form a "posed CLIP" embedding $c(x, R, T)$. We apply cross-attention to condition the denoising U-Net, which provides high-level semantic information of the input image. On the other stream, the input image is channel-concatenated with the image being denoised, assisting the model in keeping the identity and details of the object being synthesized. To be able to apply classifier-free guidance [19], we follow a similar mechanism proposed in [3], setting the input image and the posed CLIP embedding to a null vector randomly, and scaling the conditional information during inference.

3.3. 3D Reconstruction

In many applications, synthesizing novel views of an object is not enough. A full 3D reconstruction capturing both the appearance and geometry of an object is desired. We adopt a recently open-sourced framework, Score Jacobian Chaining (SJC) [53], to optimize a 3D representation with priors from text-to-image diffusion models. However, due to the probabilistic nature of diffusion models, gradient updates are highly stochastic. A crucial technique used in SJC, inspired by DreamFusion [38], is to set the classifier-free guidance value to be significantly higher than usual. This methodology decreases the diversity of each sample but improves the fidelity of the reconstruction.

As shown in Figure 4, similarly to SJC, we randomly sample viewpoints and perform volumetric rendering. We then perturb the resulting images with Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$, and denoise them by applying the U-Net ϵ_θ conditioned on the input image x, posed CLIP embedding $c(x, R, T)$, and timestep t, in order to approximate the score toward the non-noisy input x_π:

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

where $\nabla \mathcal{L}_{SJC}$ is the PAAS score introduced by [53].

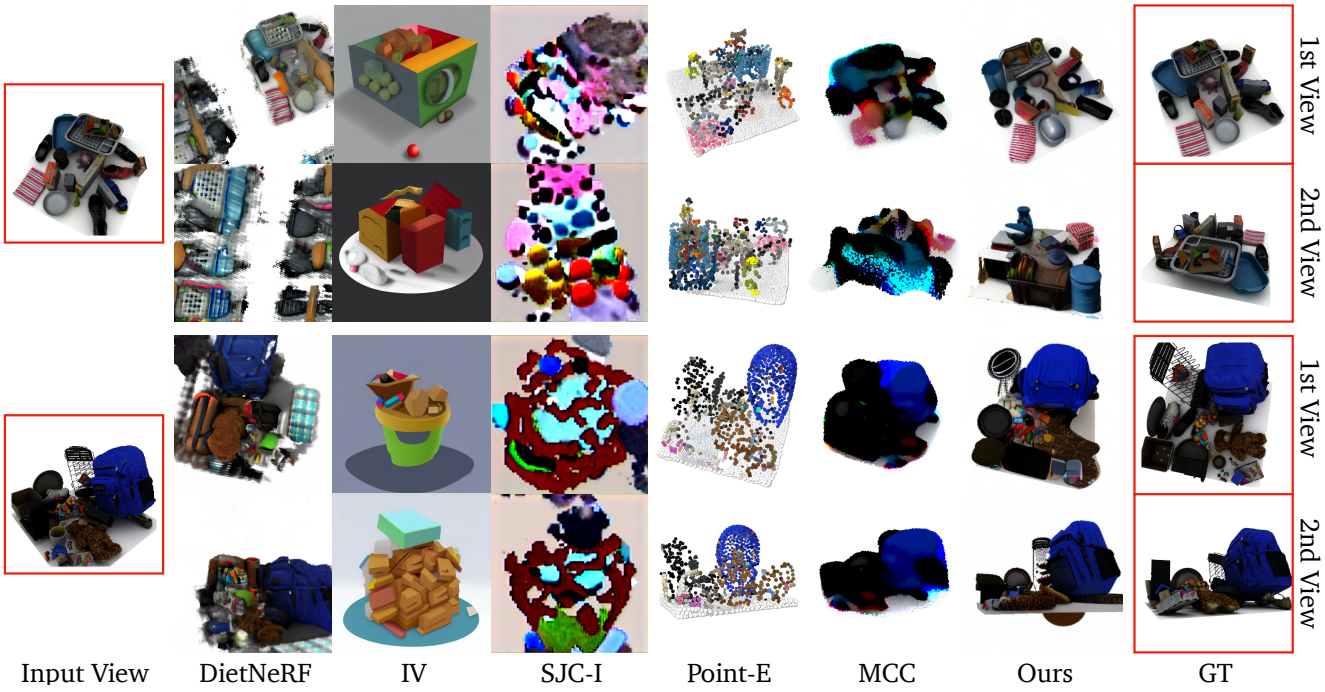

Figure 6: **Novel view synthesis on RTMV [50].** The input view shown on the left is used to synthesize 2 randomly sampled novel views. Corresponding ground truth views are shown on the right. Our synthesized view maintains a high fidelity even under big camera viewpoint changes, while most other methods deteriorate in quality drastically.

In addition, we optimize the input view with an MSE loss. To further regularize the NeRF representation, we also apply a depth smoothness loss to every sampled viewpoint, and a near-view consistency loss to regularize the change in appearance between nearby views.

3.4. Dataset

We use the recently released *Objaverse* [8] dataset for fine-tuning, which is a large-scale open-source dataset containing 800K+ 3D models created by 100K+ artists. While it has no explicit class labels like ShapeNet [4], Objaverse embodies a large diversity of high-quality 3D models with rich geometry, many of them with fine-grained details and material properties. For each object in the dataset, we randomly sample 12 camera extrinsics matrices $\mathcal{M}_{\urcorner}$ pointing at the center of the object and render 12 views with a ray-tracing engine. At training time, two views can be sampled for each object to form an image pair $(x, x_{R,T})$. The corresponding relative viewpoint transformation (R, T) that defines the mapping between both perspectives can easily be derived from the two extrinsic matrices.

4. Experiments

We assess our model's performance on zero-shot novel view synthesis and 3D reconstruction. As confirmed by the authors of Objaverse, the datasets and images we used in this paper are outside of the Objaverse dataset, and can thus be considered zero-shot results. We quantitatively compare our model to the state-of-the-art on synthetic objects and scenes with different levels of complexity. We also report qualitative results using diverse in-the-wild images, ranging from pictures we took of daily objects to paintings.

4.1. Tasks

We describe two closely related tasks that take single-view RGB images as input, and we apply them zero-shot.

Novel view synthesis. Novel view synthesis is a long-standing 3D problem in computer vision that requires a model to learn the depth, texture, and shape of an object implicitly. The extremely limited input information of only a single view requires a novel view synthesis method to leverage prior knowledge. Recent popular methods have relied on optimizing implicit neural fields with CLIP consistency objectives from randomly sampled views [23]. Our approach for view-conditional image generation is orthogonal, however, because we invert the order of 3D reconstruction and novel view synthesis, while still retaining the identity of the object depicted in the input image. This way, the aleatoric uncertainty due to self-occlusion can be modeled by a probabilistic generative model when rotating around objects, and the semantic and geometric priors learned by large diffusion models can be leveraged effectively.

Figure 7: **Novel view synthesis on in-the-wild images.** The 1st, 3rd, and 4th rows show results on images taken by an iPhone, and the 2nd row shows results on an image downloaded from the Internet. Our method works are robust to objects with different surface materials and geometry. We randomly sampled 5 different viewpoints and directly showcase the results without cherry-picking. We include more uncurated results in the supplementary materials.

3D Reconstruction. We can also adapt a stochastic 3D reconstruction framework such as SJC [53] or DreamFusion [38] to create a most likely 3D representation. We parameterize this as a voxel radiance field [5, 49, 13], and subsequently extract a mesh by performing marching cubes on the density field. The application of our view-conditioned diffusion model for 3D reconstruction provides a viable path to channel the rich 2D appearance priors learned by our diffusion model toward 3D geometry.

4.2. Baselines

To be consistent with the scope of our method, we compare only to methods that operate in a zero-shot setting and use single-view RGB images as input.

For novel view synthesis, we compare against several state-of-the-art, single-image algorithms. In particular, we benchmark DietNeRF [23], which regularizes NeRF with a CLIP image-to-image consistency loss across viewpoints. In addition, we compare with Image Variations (IV) [1], which is a Stable Diffusion model fine-tuned to be conditioned on images instead of text prompts and could be seen as a semantic nearest-neighbor search engine with Stable Diffusion. Finally, we adapted SJC [53], a diffusion-based text-to-3D model where the original text-conditioned diffusion model is replaced with an image-conditioned diffusion model, which we termed SJC-I.

For 3D reconstruction, we use two state-of-the-art, single-view algorithms as baselines: (1) Multiview Compressive Coding (MCC) [59], which is a neural field-based approach that completes RGB-D observations into a 3D representation, as well as (2) Point-E [34], which is a diffusion model over colorized point clouds. MCC is trained on CO3Dv2 [43], while Point-E is notably trained on a significantly bigger OpenAI's internal 3D dataset. We also compare against SJC-I.

Since MCC requires depth input, we use MiDaS [42, 41] off-the-shelf for depth estimation. We convert the obtained relative disparity map to an absolute pseudo-metric depth

Figure 8: **Diversity of novel view synthesis.** With an input view, we fix another viewpoint and randomly generate multiple conditional samples. The different results reflect a range of diversity in terms of both geometry and appearance information that is missing in the input view.

	DietNeRF [23]	Image Variation [1]	SJC-I [53]	Ours
PSNR ↑	8.933	5.914	6.573	**18.378**
SSIM ↑	0.645	0.540	0.552	**0.877**
LPIPS ↓	0.412	0.545	0.484	**0.088**
FID ↓	12.919	22.533	19.783	**0.027**

Table 1: **Results for novel view synthesis on Google Scanned Objects.** All metrics demonstrate that our method is able to outperform the baselines by a significant margin.

	DietNeRF [23]	Image Variation [1]	SJC-I [53]	Ours
PSNR ↑	7.130	6.561	7.953	**10.405**
SSIM ↑	0.406	0.442	0.456	**0.606**
LPIPS ↓	0.507	0.564	0.545	**0.323**
FID ↓	5.143	10.218	10.202	**0.319**

Table 2: **Results for novel view synthesis on RTMV.** Scenes in RTMV are out-of-distribution from Objaverse training data, yet our model still outperforms the baselines by a significant margin.

map by assuming standard scale and shift values that look reasonable across the entire test set.

4.3. Benchmarks and Metrics

We evaluate both tasks on Google Scanned Objects (GSO) [10], which is a dataset of high-quality scanned household items, as well as RTMV [50], which consists of complex scenes, each composed of 20 random objects. In all experiments, the respective ground truth 3D models are used for evaluating 3D reconstruction.

For novel view synthesis, we numerically evaluate our method and baselines extensively with four metrics covering different aspects of image similarity: PSNR, SSIM [55], LPIPS [64], and FID [18]. For 3D reconstruction, we measure Chamfer Distance and volumetric IoU.

4.4. Novel View Synthesis Results

We show the numerical results in Tables 1 and 2. Figure 5 shows that our method, as compared to all baselines on GSO, is able to generate highly photorealistic images that are closely consistent with the ground truth. Such a trend can also be found on RTMV in Figure 6, even though the scenes are out-of-distribution compared to the Objaverse dataset. Among our baselines, we observed that Point-E tends to achieve much better results than other baselines, maintaining impressive zero-shot generalizability. However, the small size of the generated point clouds greatly limits the applicability of Point-E for novel view synthesis.

In Figure 7, we further demonstrate the generalization performance of our model to objects with challenging geometry and texture as well as its ability to synthesize high-fidelity viewpoints while maintaining the object type, identity and low-level details.

Diversity across samples. Novel view synthesis from a single image is a severely under-constrained task, which makes diffusion models a particularly apt choice of architecture compared to NeRF in terms of capturing the underlying uncertainty. Because input images are 2D, they always depict only a partial view of the object, leaving many parts unobserved. Figure 8 exemplifies the diversity of plausible, high-quality images sampled from novel viewpoints.

4.5. 3D Reconstruction Results

We show numerical results in Tables 3 and 4. Figure 9 qualitatively shows our method reconstructs high-fidelity

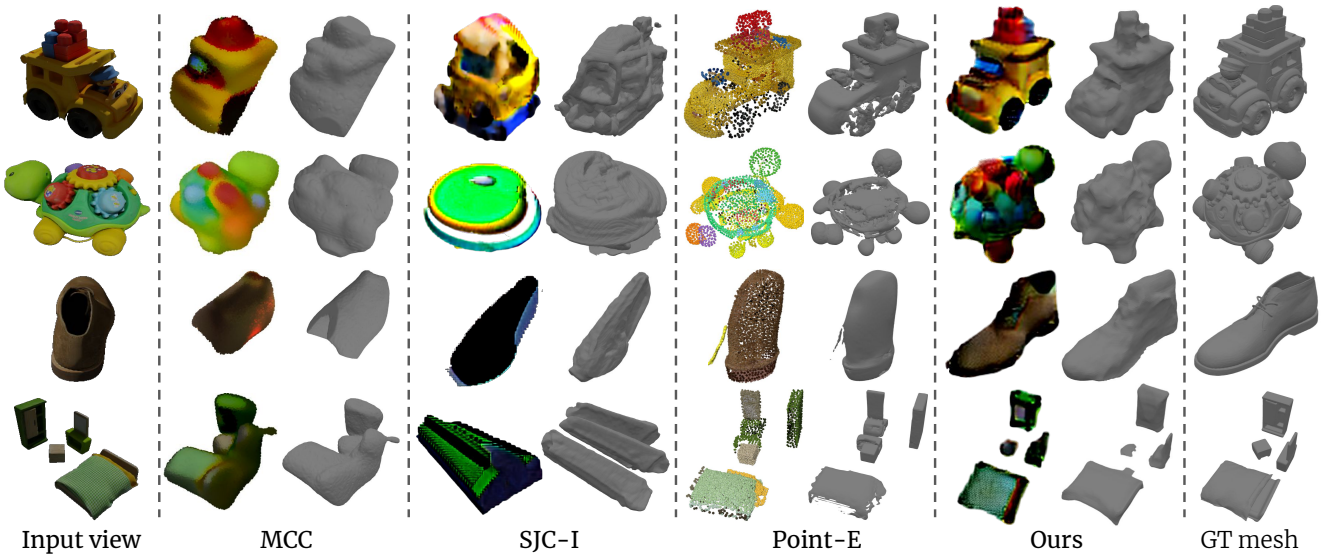

Figure 9: **Qualitative examples of 3D reconstruction.** The input view images are shown on the left. For each method, we show a rendered view from a different angle and the reconstructed 3D mesh. The ground truth meshes are shown on the right.

	MCC [59]	SJC-I [53]	Point-E [34]	Ours
CD ↓	0.1230	0.2245	0.0804	**0.0717**
IoU ↑	0.2343	0.1332	0.2944	**0.5052**

Table 3: **Results for single view 3D reconstruction on GSO.** Note that our volumetric IoU is better than the compared methods by a large margin.

	MCC [59]	SJC-I [53]	Point-E [34]	Ours
CD ↓	0.1578	0.1554	0.1565	**0.1352**
IoU ↑	0.1550	0.1380	0.0784	**0.2196**

Table 4: **Results for single view 3D reconstruction on RTMV.** Because RTMV consists of cluttered scenes with many objects in them, none of the studied approaches seem to perform very well. Our method is still the best one however, despite not being explicitly trained for the 3D reconstruction task.

3D meshes that are consistent with the ground truth. MCC tends to give a good estimation of surfaces that are visible from the input view, but often fails to correctly infer the geometry at the back of the object.

SJC-I is also frequently unable to reconstruct a meaningful geometry. On the other hand, Point-E has an impressive zero-shot generalization ability, and is able to predict a reasonable estimate of object geometry. However, it generates non-uniform sparse point clouds of only 4,096 points, which sometimes leads to holes in the reconstructed surfaces (according to their provided mesh conversion method). Therefore, it obtains a good CD score but falls short of the volumetric IoU. Our method leverages the learned multi-view priors from our view-conditioned diffusion model and combines them with the advantages of a NeRF-style representation. Both factors provide improvements in terms of CD and volumetric IoU over prior works, as indicated by Tables 3 and 4.

4.6. Text to Image to 3D

In addition to in-the-wild images, we also tested our method on images generated by txt2img models such as Dall-E-2 [40]. As shown in Figure 10, our model is able to generate novel views of these images while preserving the identity of the objects. We believe this could be very useful in many text-to-3D generation applications.

5. Discussion

In this work, we have proposed a novel approach, *Zero-1-to-3*, for zero-shot, single-image novel-view synthesis and 3D reconstruction. Our method capitalizes on the Stable Diffusion model, which is pre-trained on internet-scaled data and captures rich semantic and geometric priors. To extract this information, we have fine-tuned the model on synthetic data to learn control over the camera viewpoint. The resulting method demonstrated state-of-the-art results on several benchmarks due to its ability to leverage strong object shape priors learned by Stable Diffusion.

5.1. Future Work

From objects to scenes. Our approach is trained on a dataset of single objects on a plain background. While we have demonstrated a strong degree of generalizations to

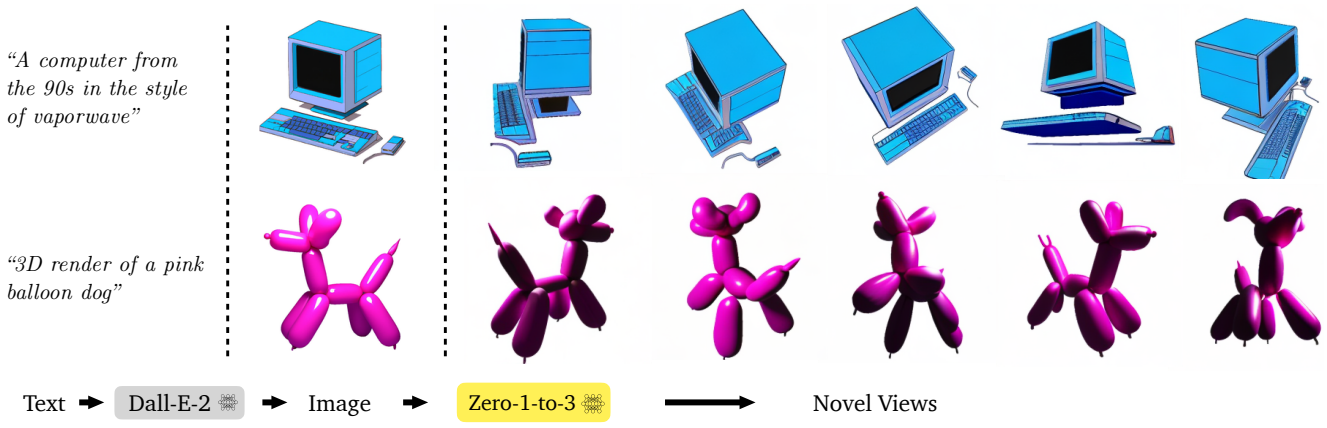

Figure 10: **Novel View Synthesis from Dall-E-2 Generated Images.** The composition of multiple objects (1st row) and the lighting details (2nd row) are preserved in our synthesized novel views.

scenes with several objects on RTMV dataset, the quality still degrades compared to the in-distribution samples from GSO. Generalization to scenes with complex backgrounds thus remains an important challenge for our method.

From scenes to videos. Being able to reason about geometry of dynamic scenes from a single view would open novel research directions, such as understanding occlusions [52, 27] and dynamic object manipulation. A few approaches for diffusion-based video generation have been proposed recently [20, 11], and extending them to 3D would be key to opening up these opportunities.

Combining graphics pipelines with Stable Diffusion. In this paper, we demonstrate a framework to extract 3D knowledge of objects from Stable Diffusion. A powerful natural image generative model like Stable Diffusion contains other implicit knowledge about lighting, shading, texture, etc. Future work can explore similar mechanisms to perform traditional graphics tasks, such as scene relighting.

Acknowledgements: We would like to thank Changxi Zheng and Samir Gadre for their helpful feedback. We would also like to thank the authors of SJC [53], NeRDi [9], SparseFusion [65], and Objaverse [8] for their helpful discussions. This research is based on work partially supported by the Toyota Research Institute, the DARPA MCS program under Federal Agreement No. N660011924032, and the NSF NRI Award #1925157.

References

[1] Stable diffusion image variations - a hugging face space by lambdalabs. 6, 7, 12

[2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam. 12

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023. 4

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial radiance fields. In *ECCV*, 2022. 6

[6] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. In *ICLR*, 2021. 2

[7] Ken Cheng. Reflections on geometry and navigation. *Connection Science*, 17(1-2):5–21, 2005. 1

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. *arXiv preprint arXiv:2212.08051*, 2022. 5, 9, 12

[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. *arXiv preprint arXiv:2212.03267*, 2022. 2, 9

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *ICRA*, 2022. 4, 7

[11] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 9

[12] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 2

[13] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 6

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2

[15] 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. 2

[16] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2

[17] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009. 1

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2022. 4

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 9

[21] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Planes vs. chairs: Category-guided 3D shape learning without any 3D cues. In *ECCV*, 2022. 1

[22] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022. 2

[23] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a Diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021. 5, 6, 7, 13

[24] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 1

[25] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 1

[26] 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. *arXiv preprint arXiv:2211.10440*, 2022. 2

[27] Ruoshi Liu, Sachit Menon, Chengzhi Mao, Dennis Park, Simon Stent, and Carl Vondrick. Shadows shed light on 3D objects. *arXiv preprint arXiv:2206.08990*, 2022. 9

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[29] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. RealFusion: 360° reconstruction of any object from a single image. *arXiv preprint arXiv:2302.10663*, 2023. 2

[30] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, 2019. 2

[31] 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. 2

[32] Gillian M Morriss-Kay. The evolution of human artistic creativity. *Journal of anatomy*, 216(2):158–176, 2010. 1

[33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2

[34] 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. 6, 8

[35] OPHoperHPO. Ophoperhpo/image-background-remove-tool: automated high-quality background removal framework for an image using neural networks. . 12, 13

[36] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 1, 2

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1

[38] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023. 2, 4, 6

[39] 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. 4

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2, 8

[41] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 6, 13

[42] 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*, 44(3):1623–1637, 2020. 6, 13

[43] 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. 2, 6

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image

diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[46] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 2

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2, 3

[48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[49] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 6

[50] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, and Stan Birchfield. RTMV: A ray-traced multi-view synthetic dataset for novel view synthesis. *ECCVW*, 2022. 5, 7

[51] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 2

[52] Basile Van Hoorick, Purva Tendulkar, Dídac Surís, Dennis Park, Simon Stent, and Carl Vondrick. Revealing occlusions with 4D neural fields. In *CVPR*, 2022. 2, 9

[53] 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, 4, 6, 7, 8, 9

[54] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[56] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 2

[57] Wikipedia. Spherical coordinate system — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Spherical%20coordinate%20system&oldid=1142703172, 2023. [Online; accessed 14-March-2023]. 12

[58] Markus Worchel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, and Peter Eisert. Multi-view mesh reconstruction with neural deferred shading. In *CVPR*, 2022. 2

[59] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3D reconstruction. *arXiv:2301.08247*, 2023. 3, 6, 8, 13

[60] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. MarrNet: 3D shape reconstruction via 2.5D sketches. In *NeurIPS*, 2017. 2

[61] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. NeuralLift-360: Lifting an in-the-wild 2D photo to a 3D object with 360° views. In *CVPR*, 2023. 2

[62] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. In *NeurIPS*, 2019. 2

[63] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[64] 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

[65] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction, 2022. 9

[66] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-D safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *ICCV*, 2019. 1

[67] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *CVPR*, 2018. 1

[68] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, 2017. 1

Appendix

A. Coordinate System & Camera Model

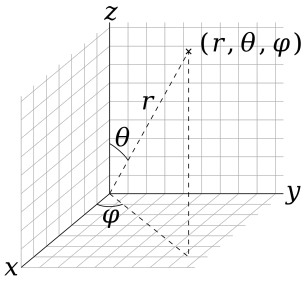

Figure 11: Spherical Coordinate System [57].

We use a spherical coordinate system to represent camera locations and their relative transformations. As shown in Figure 11, assuming the center of the object is the origin of the coordinate system, we can use θ, ϕ, and r to represent the polar angle, azimuth angle, and radius (distance away from the center) respectively. For the creation of the dataset, we normalize all assets to be contained inside the XYZ unit cube $[-0.5, 0.5]^3$. Then, we sample camera viewpoints such that $\theta \in [0, \pi]$, $\phi \in [0, 2\pi]$ uniformly cover the unit sphere, and r is sampled uniformly in the interval $[1.5, 2.2]$. During training, when two images from different viewpoints are sampled, let their camera locations be (θ_1, ϕ_1, r_1) and (θ_2, ϕ_2, r_2). We denote their *relative* camera transformation as $(\theta_2 - \theta_1, \phi_2 - \phi_1, r_2 - r_1)$. Since the camera is always pointed at the center of the coordinate system, the extrinsics matrices are uniquely defined by the location of the camera in a spherical coordinate system. We assume the horizontal field of view of the camera to be $49.1°$, and follow a pinhole camera model.

Due to the incontinuity of the azimuth angle, we encode it with $\phi \mapsto [\sin(\phi), \cos(\phi)]$. Subsequently, at both training and inference time, four values representing the relative camera viewpoint change, $[\theta, \sin(\phi), \cos(\phi), r]$ are fed to the model, along with an input image, in order to generate the novel view.

B. Dataset Creation

We use Blender [2] to render training images of the finetuning dataset. The specific rendering code is inherited from a publicly released repository[1] by authors of Objaverse [8]. For each object in Objaverse, we randomly sample 12 views and use the Cycles engine in Blender with 128 samples per ray along with a denoising step to render each image. We render all images in 512×512 resolution and pad transparent backgrounds with white color. We also apply randomized area lighting. In total, we rendered a dataset of around 10M images for finetuning.

[1] https://github.com/allenai/objaverse-rendering

C. Finetuning Stable Diffusion

We use the rendered dataset to finetune a pretrained Stable Diffusion model for performing novel view synthesis. Since the original Stable Diffusion network is not conditioned on multimodal text embeddings, the original Stable Diffusion architecture needs to be tweaked and finetuned to be able to take conditional information from an image. This is done in [1], and we use their released checkpoints. To further adapt the model to accept conditional information from an image along with a relative camera pose, we concatenate the image CLIP embedding (dimension 768) and the pose vector (dimension 4) and initialize another fully-connected layer ($772 \mapsto 768$) to ensure compatibility with the diffusion model architecture. The learning rate of this layer is scaled up to be $10\times$ larger than the other layers. The rest of the network architecture is kept the same as the original Stable Diffusion.

C.1. Training Details

We use AdamW [28] with a learning rate of 10^{-4} for training. First, we attempted a batch size of 192 while maintaining the original resolution (image dimension 512×512, latent dimension 64×64) for training. However, we discovered that this led to a slower convergence rate and higher variance across batches. Because the original Stable Diffusion training procedure used a batch size of 3072, we subsequently reduce the image size to 256×256 (and thus the corresponding latent dimension to 32×32), in order to be able to increase the batch size to 1536. This increase in batch size has led to better training stability and a significantly improved convergence rate. We finetuned our model on an 8×A100-80GB machine for 7 days.

C.2. Inference Details

To generate a novel view, Zero-1-to-3 takes only 2 seconds on an RTX A6000 GPU. Note that in prior works, typically a NeRF is trained in order to render novel views, which takes significantly longer. In comparison, our approach inverts the order of 3D reconstruction and novel view synthesis, causing the novel view synthesis process to be fast and contain diversity under uncertainty. Since this paper addresses the problem of a single image to a 3D object, when an in-the-wild image is used during inference, we apply an off-the-shelf background removal tool [35] to every image before using it as input to Zero-1-to-3.

D. 3D Reconstruction

Different from the original Score Jacobian Chaining (SJC) implementation, we removed the "emptiness loss" and "center loss". To regularize the VoxelRF representation,

we differentiably render a depth map, and apply a smoothness loss to the depth map. This is based on the prior knowledge that the geometry of an object typically contains less high-frequency information than its texture. It is particularly helpful in removing holes in the object representation. We also apply a near-view consistency loss to regularize the difference between an image rendered from one view and another image rendered from a nearby randomly sampled view. We found this to be very helpful in improving the cross-view consistency of an object's texture. All implementation details can be found in the code that is submitted as part of the appendix. Running a full 3D reconstruction on an image takes around 30 minutes on an RTX A6000 GPU.

Mesh extraction. We extract the 3D mesh from the VoxelRF representation as follows. We first query the density grids at resolution 200^3. Then we smooth the density grids using a mean filter of size $(7, 7, 7)$, followed by an erosion operator of size $(5, 5, 5)$. Finally, we run marching cubes on the resulting density grids. Let $\bar{d}$ denote the average value of the density grids. For the GSO dataset, we use a density threshold of $8\bar{d}$. For the RTMV dataset, we use a density threshold of $4\bar{d}$.

Evaluation. The ground truth 3D shape and the predicted 3D shape are first normalized within the unit cube. To compute the chamfer distance (CD), we randomly sample 2000 points. For Point-E and MCC, we sample from their predicted point clouds directly. For our method and SJC-I, we sample points from the reconstructed 3D mesh. We compute the volumetric IoU at resolution 64^3. For our method, Point-E and SJC-I, we vocalize the reconstructed 3D surface meshes using marching cubes. For MCC, we directly voxelize the predicted dense point clouds by occupancy.

E. Baselines

To be consistent with the scope of our method, we compare only to methods that (1) operate in a zero-shot setting, (2) use single-view RGB images as input, and (3) have official reference implementations available online that can be adapted in a reasonable timeframe. In the following sections, we describe the implementation details of our baselines.

E.1. DietNeRF

We use the official implementation located on GitHub[2], which, at the time of writing, has code for low-view NeRF optimization from scratch with a joint MSE and consistency loss, though provides no functionality related to finetuning PixelNeRF. For fairness, we use the same hyperparameters as the experiments performed with the NeRF synthetic dataset in [23]. For the evaluation of novel view synthesis, we render the resulting NeRF from the designated camera poses in the test set.

E.2. Point-E

We use the official implementation and pretrained models located on GitHub[3]. We keep all the hyperparameters and follow their demo example to do 3D reconstruction from single input image. The prediction is already normalized, so we do not need to perform any rescaling to match the ground truth. For surface mesh extraction, we use their default method with a grid size of 128.

E.3. MCC

We use the official implementation located on GitHub[4]. Since this approach requires a colorized point cloud as input rather than an RGB image, we first apply an online off-the-self foreground segmentation method [35] as well as a state-of-the-art depth estimation method [42, 41] for preprocessing. For fairness, we keep all hyperparameters the same as the zero-shot, in-the-wild experiments described in [59]. For the evaluation of 3D reconstruction, we normalize the prediction, rotate it according to camera extrinsics, and compare it with the 3D ground truth.

[2] https://github.com/ajayjain/DietNeRF

[3] https://github.com/openai/point-e

[4] https://github.com/facebookresearch/MCC