iNVS: Repurposing Diffusion Inpainters for Novel View Synthesis

YASH KANT, University of Toronto, Canada and Snap Research, Canada
ALIAKSANDR SIAROHIN, Snap Research, USA
MICHAEL VASILKOVSKY, Snap Research, USA
RIZA ALP GULER, Snap Research, UK
JIAN REN, Snap Research, USA
SERGEY TULYAKOV, Snap Research, USA
IGOR GILITSCHENSKI, University of Toronto, Canada

Fig. 1. **Novel view synthesis results for *unseen* objects.** Our system synthesizes novel view from a single image for unseen objects. We obtain detailed generations, while respecting the appearance of the region that is visible in the input image by maximizing reuse of source pixels.

We present a method for generating consistent novel views from a single source image. Our approach focuses on maximizing the reuse of visible pixels from the source image. To achieve this, we use a monocular depth estimator that transfers visible pixels from the source view to the target view. Starting from a pre-trained 2D inpainting diffusion model, we train our method on the large-scale *Objaverse* dataset to learn 3D object priors. While training we use a novel masking mechanism based on epipolar lines to further improve the quality of our approach. This allows our framework to perform zero-shot novel view synthesis on a variety of objects. We evaluate the zero-shot abilities of our framework on three challenging datasets: Google Scanned Objects, Ray Traced Multiview, and Common Objects in 3D. See our webpage for more details: https://yashkant.github.io/invs/

CCS Concepts: • **Computing methodologies** → **Computer vision**; *Computer graphics.*

Additional Key Words and Phrases: Novel View Synthesis, Diffusion Models, Inpainting

1 INTRODUCTION

Synthesizing novel views from a *single* image has broad applications in the computer vision fields, including object reconstruction for Augmented and Virtual Reality [Choi et al. 2019; Gao et al. 2021; Tucker and Snavely 2020], animating humans in the game and movie industry [Bhatnagar et al. 2019; Hu et al. 2021a], and environment understanding in robotics [Hu et al. 2021b; Wiles et al. 2020], *etc*. However, generating high-fidelity novel views from one image is a longstanding challenging problem, as the image generation algorithm needs to infer the correct geometry from a partial observation of an object. Existing literature still struggles to build the generic framework to reconstruct high-quality 3D objects for image rendering. For instance, some studies require a manual 3D prior, such as human bodies [Liao et al. 2023; Xiu et al. 2023] and faces [Jiang et al. 2018; Kemelmacher-Shlizerman and Basri 2010] for reconstruction, limiting the generation to specific categories. Other approaches suffer from inadequate reconstruction quality when dealing with single images [Chen and Williams 1993; Debevec et al. 1996; Shih et al. 2020] and often require multiple views as input [Martin-Brualla et al. 2021; Mildenhall et al. 2020], which may not always be available.

To tackle such a problem, we first step back to understand why we, as humans, can understand how objects should look in any view direction by just viewing the object from a single view. The reason might be that the 3D prior knowledge equipped by humans is obtained by observing a huge number of objects from various viewpoints in the real world. Thus, leveraging the knowledge learned

from different objects can be a solution for accurately reconstructing objects. Recently, the emerging efforts on large-scale text-to-image diffusion models [Ramesh et al. 2022, 2021; Rombach et al. 2022; Saharia et al. 2022] prove the capability of learning a generic object prior by training on large-scale image datasets, *e.g.*, LAION [Schuhmann et al. 2022]. However, these models operate in the 2D domain and lack precise control over camera view directions, limiting their effectiveness in view synthesis tasks.

This work empowers the pre-trained large-scale text-to-image diffusion model with the ability for camera viewpoint control to generate novel views. We make the following contributions: first, we attempt to reuse pixels from the input view when camera views are not significantly far away. This is achieved by back-projecting such pixels into the 3D space using monocular depth and reprojecting them back onto the novel view. Second, we apply inpainting to recover the missing regions by leveraging Inpainting Stable Diffusion (*ISD*) [Rombach et al. 2022]. However, naïvely applying *ISD* fails to generalize well to the masks that arise from the reprojection procedure since the *ISD* model is trained with masks that randomly cover a part of the image. Therefore, we propose to train *ISD* on a dataset in which we can compute such masks easily. One prominent choice is a dataset of 3D assets, *e.g.*, *Objaverse* [Deitke et al. 2023], which can be rendered from multiple views and for which such masks can be computed. After training, our method can predict missing pixels in the novel view image, while at the same time preserving pixels that are initially visible. We abbreviate our method as *iNVS* which stands for inpainting-driven Novel View Synthesis.

We conduct experiments on synthetic and real datasets, and find that our method can achieve strong novel view synthesis (NVS) results from single images as shown in Figure 1. We conduct ablative and failure mode analyses which demonstrates that a good monocular depth estimator is important to preserve structure and allow maximal reuse of source pixels.

2 RELATED WORKS

Novel View Synthesis in Space. Novel view synthesis is a long-standing problem in computer vision and graphics. Early methods rely on the images from multiple viewpoints and attempt to incorporate the knowledge from epipolar geometry to perform smooth interpolation between the different views [Chen and Williams 1993; Debevec et al. 1996]. One of the important milestones in novel view synthesis is the introduction of Neural Radiance Fields (NeRFs) [Mildenhall et al. 2020]. NeRFs can synthesize smooth interpolations between different views with the help of volumetric rendering. Since then, numerous improvements have been introduced to improve the original design [Barron et al. 2021; Chen et al. 2021; Kuang et al. 2022; Wang et al. 2021; Zhang et al. 2020]. However, most of them share the same limitation of relying on multiple views for learning 3D representation.

Newer works have demonstrated that using deep networks is a promising approach for synthesizing novel views from few images owing to their generalization capabilities [Chan et al. 2023; Deng et al. 2023; Mirzaei et al. 2023; Sajjadi et al. 2022; Zhou and Tulsiani 2023]. In its limit, this approach allows for generating novel views given exactly a *single* image [Gu et al. 2023; Shen et al. 2023; Shih et al.

2020; Tang et al. 2023; Wiles et al. 2020] and we adapt this setting in our work. Zero123 [Liu et al. 2023] proposes to fine-tune Stable Diffusion for NVS task. They condition diffusion both on the source image and on the CLIP embedding of the source image. However, this method largely ignores the inability of U-Net [Ronneberger et al. 2015] networks to generate output that is not aligned with the source [Siarohin et al. 2019a,b]. In contrast, our method relies on geometry clues to align the source and target views. This helps us to preserve the content from the source image well.

Novel View Synthesis in Time. Video generation conditional on one or more input image(s) can be seen as a novel view synthesis task with generated images unrolling both in space and time. Prior works [Denton and Fergus 2018; Finn et al. 2016; Hsieh et al. 2018; Villegas et al. 2017; Vondrick and Torralba 2017; Wang et al. 2017] in this domain proposed the use of spatiotemporal conditioning to handle dynamic scenes. More recent works have trained on large indoor scene and video datasets further improving quality of generations task [Koh et al. 2021; Lee et al. 2021; Ye et al. 2019; Yu et al. 2022].

Our work is particularly inspired by InfiniteNature and Infinite-Nature Zero works [Li et al. 2022a; Liu et al. 2021] which utilize softmax-splatting [Niklaus and Liu 2020] for synthesizing infinite videos of nature with a fly-through camera. Recent works in this space have tackled generation of novel views with full 360-degree camera control [Chai et al. 2023], as well as learning domain-specific dynamics of abstract scenes [Mahapatra and Kulkarni 2022], and very recently general dynamic prior from largescale videos [Li et al. 2023].

3D Generative Models. The recent surge in the quality and diversity of generations orchestrated by 2D image diffusion models poses the question of whether the prior knowledge learned by these models can be used for generating 3D objects and scenes. Indeed, diffusion models have some textual control over the viewpoint. For example, DreamBooth [Ruiz et al. 2023] shows that the diffusion model can properly react to the words "front", "back", and "side" words the prompt. The seminal work that exploits 2D diffusion for 3D generation, DreamFusion [Poole et al. 2023], proposes to optimize NeRF representation by judging the novel view generations with the large-scale pre-trained text-to-image diffusion model [Saharia et al. 2022]. Several follow-up works [Chen et al. 2023a; Lin et al. 2023] improve the resolution and quality of the resulting 3D assets. On the other hand, Dreambooth3D [Raj et al. 2023] introduces additional image control. Although these works can generate reasonable novel views, they require a lengthy optimization process.

Additionally, several works [Chen et al. 2023b; Richardson et al. 2023] have proposed to utilize stable diffusion for mesh texturing. TEXTure [Richardson et al. 2023] uses 2D diffusion models to sequentially in-paint novel regions over the existing mesh, projecting results via a differentiable renderer. Text2Tex [Chen et al. 2023b] extends this strategy with an automatic viewpoint-finding approach for optimized re-projection. In both of these works, Stable Diffusion utilizes text as conditioning, which provides only limited control over the generation.
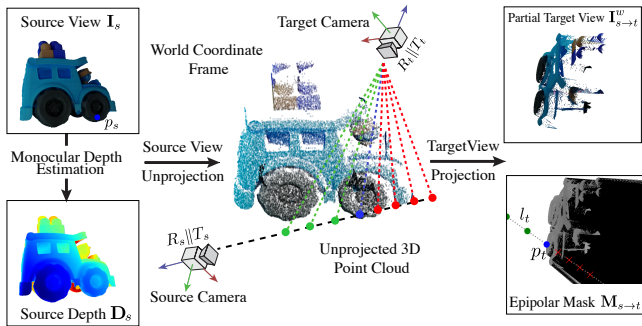
Fig. 2. **Epipolar mask and partial view generation strategy.** Starting with a source view, we use a pre-existing monocular depth estimator to calculate the distance of each point in the image from the camera, creating a depth map. We then use this depth map to "unproject" the 2D image into 3D space, generating a partial point cloud. Next, we take the partial point cloud and "reproject" it onto the target view, essentially projecting the 3D points back onto a 2D image from a different angle. As we do this, we also generate a "visibility mask" which identifies any new areas that become visible in the target view that were not visible in the source view.

3 METHOD

In this section, we introduce the overall task setup (Sec. 3.1), the strategy used for generating inputs to the inpainting network (Sec. 3.2), three different losses used throughout training (Sec. 3.3), and our inference technique (Sec. 3.4).

3.1 Overview

Novel View Synthesis Task. Given a single RGB image of a 3D asset (source view) $I_s \in \mathbb{R}^{h \times w \times 3}$, and the corresponding camera pose $C_s \in \mathbb{R}^{3 \times 4}$, we aim to generate a target view $I_t \in \mathbb{R}^{h \times w \times 3}$ of this asset from a novel viewpoint, say $C_t \in \mathbb{R}^{3 \times 4}$.

Inpainter Inputs. We start by preparing inputs for the inpainting model, which takes in a partial view of the scene as well as a binary mask that denotes the region to be inpainted. We then obtain source view depth $D_s \in \mathbb{R}^{h \times w}$, which is available for our synthetic training set, and can be calculated using an off-the-shelf monocular depth estimator [Bhat et al. 2023] during inference. Using this depth map, we warp the pixels from the source to the target viewpoint by creating a partial target view $I_{s \to t} \in \mathbb{R}^{h \times w \times 3}$. We train our *ISD* model to inpaint the missing regions of this partial view. Additionally, we provide the inpainter network with a mask $M_{s \to t} \in \mathbb{R}^{h \times w}$ that indicates parts of the image which require inpainting.

Training and Inference. We train our *iNVS* model intialized from the Stable Diffusion inpainting checkpoint[1] on source-target pairs sampled at random from 20M rendered views of the largescale synthetic *Objaverse* [Deitke et al. 2023] dataset. The finetuning process is outlined in Sec. 3.3. Finally, we also modify the DDIM inference which helps to significantly reduce artifacts in the NVS generations, described in Sec. 3.4.
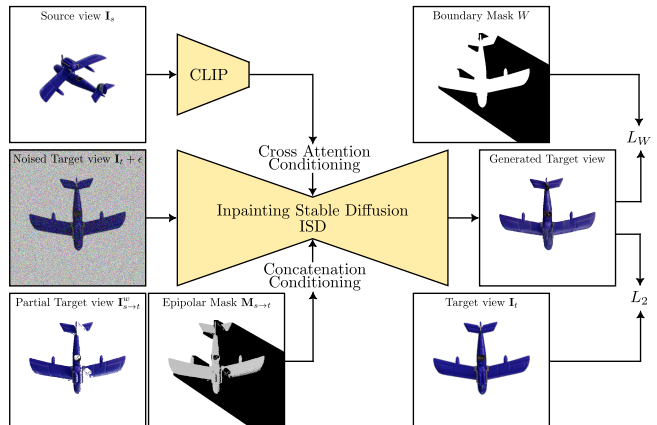
[1]https://huggingface.co/runwayml/stable-diffusion-inpainting

Fig. 3. **Inpainter model training using denoising, and boundary losses.** Inpaining Stable Diffusion (*ISD*) accepts a noised target view as well as a partial target view and epipolar mask. All three of these inputs are concatenated before they are fed into the diffusion. Instead of the text condition we use CLIP [Radford et al. 2021] embedding of the source view, that is provided to the *ISD* through cross attention layers. The final generated image is compared against ground truth with L_2 loss, moreover to enforce object shape discovery we introduce additional boundary mask loss.

3.2 Generating Partial View and Epipolar Mask

Warping source view using depth. Next, we describe how to unproject the pixels from source view I_s into 3D space, and then reproject them into target view I_t (*a.k.a.* warping). Let any source pixel from I_s be $p_s = [x, y, 1]$ defined in homogenous coordinates. We can unproject it into 3D world space by:

$$p_w = R_s \cdot d_s \cdot K_s^{-1} p_s + T_s, \qquad (1)$$

where $K_s \in \mathbb{R}^{3 \times 3}$ is the source view camera intrinsic, $C_s = [R_s|T_s] \in \mathbb{R}^{3 \times 4}$ is the source camera, and $d_s \in \mathbb{R}$ is the scalar depth value for the point p_s. Finally, world space point p_w can be reprojected in the target view with camera as:

$$p_t = K_t \cdot d_t^{-1} \cdot R_t^{-1} \cdot (p_w - T_t), \qquad (2)$$

where $C_t = [R_t|T_t]$ is the target camera, $d_t \in \mathbb{R}$ is scalar target depth, and p_t is target pixel in homogenous coordinates. Applying the above transform for all foreground pixels in the source view can obtain the partial target view $I_{s \to t}$. Additionally, when reprojecting points to the target view, we use forward softmax-splatting [Niklaus and Liu 2020] similar to [Li et al. 2022b] to handle overlapping points using z-values. In Figures 2 and 3, we visualize the warped target outputs.

How to create inpainting mask? Reprojecting source pixels to target view only gives us the information about the visible pixels of the object. However, it does not tell the inpainting network anything regarding which regions in the image are newly discovered and which already exist. The simplest method to construct an inpainting mask would be to use all pixels that are not part of the object in the partial target view. However, we find that using such strategy creates a very large inpainting mask, and subsequently *ISD* struggles to generalize or maintain consistency with the source view. We show results with this in Sec. 4.

Inpainting mask with Epipolar Geometry. When a light ray falls onto a particular pixel $\mathbf{p}_s$ in the source view, it corresponds to a line l_t in the target view. This line is known as an epipolar line, as illustrated in Fig. 2. It is worth noting that only a portion of this line in the target view is obstructed, while the rest remains visible. The point $\mathbf{p}_w$ precisely determines which part is obstructed. Anything preceding $\mathbf{p}_w$ is visible, whereas anything following it is obstructed. To create the inpainting or visibility mask, we generate rays from each pixel in the source view to the target view until they intersect with the reprojected point. This process yields $\mathbf{M}_{s \rightarrow t}$, which we refer to as the Epipolar Mask. Fig. 2 provides a visual depiction of this procedure along with the resulting mask.

Using smooth inpainting mask using ray angles. When creating the epipolar inpainting mask, instead of using a binary value of 0/255 (black/white) at each pixel, we use a smooth value (linearly interpolated) between the source and target camera ray angle at the corresponding 3D world point (projected onto this pixel). Providing this inpainting mask indicates to the ISD *how much camera angle variation has happened at each point* (180 degrees is black, 0 degrees is white). This information is used while training and helps the inpainter ignore/overwrite flipped pixels. Figure 5 shows a visual example of this.

3.3 Training *iNVS*: Inpainter for Novel View Synthesis

Denoising Novel Views. Equipped with the partial target view and the epipolar mask highlighting regions to be inpainted, we can now train our inpainting model. Concretely, our *ISD* takes as conditioning the epipolar mask $\mathbf{M}_{s \rightarrow t}$, partial target view $\mathbf{I}_{s \rightarrow t}$, as well as the CLIP embedding of the source view X_s, and it is trained to denoise a noisy target view $\mathbf{I}_t + \epsilon$. Following previous works [Dhariwal and Nichol 2021; Rombach et al. 2022], we utilize epsilon parameterization of diffusion and optimize our network with the following loss:

$$L_2 = \| \epsilon - ISD\left(\mathbf{I}_t + \epsilon, \mathbf{M}_{s \rightarrow t}, \mathbf{I}_{s \rightarrow t}, X_s\right) \|^2 . \tag{3}$$

Encoding Source Views with CLIP. We replace the text-encoder of CLIP [Radford et al. 2021] used in the Stable Diffusion model with its image-encoder, to condition generation on source view $\mathbf{I}_s$. Since the source view does not align well with the target view in RGB image space, we choose to formulate the conditioning via cross-attention instead of using concatenation, unlike previous work [Watson et al. 2022].

Stricter Boundary Loss. Stable Diffusion inpainting model is trained on real images with diverse backgrounds, and we find that it struggles to generate uniform solid color (white or black) backgrounds. It shows an affinity towards inpainting backgrounds with patterns, or enlarges the object boundaries to cover entirety of inpainting mask. To tackle this issue, we propose a loss re-weighting that puts more emphasis on target regions where the model has to discover object boundary. Concretely, we introduce re-weighting coefficient $W[\mathbf{p}_t] = 1$ if $\mathbf{p}_t$ is a pixel that falls within the boundary of known regions, and $W[\mathbf{p}_t] = 2$ if $\mathbf{p}_t$ is a pixel where the boundary is unknown (as shown in Fig. 3). Finally, we obtain the re-weighted loss:

$$L_W = \| W\left(\epsilon - ISD(\mathbf{I}_t + \epsilon, \mathbf{M}_{s \rightarrow t}, \mathbf{I}_{s \rightarrow t}, X_s)\right) \|^2 . \tag{4}$$

Training on early denoising steps. We observe that *ISD* mostly struggles to decode the shape of the object, while reasonable textures can be obtained even with non-finetuned *ISD* when using ground truth boundary masks [Chen et al. 2023b]. We find that during inference denoising, the shape boundary is discovered much earlier compared to texture, so we additionally fine-tune *ISD* by sampling noise levels in the first 10% of the denoiser schedule.

3.4 Inference

Rescale and Recenter. At inference time we wish to generate a novel view of the object from a single source view. Since we do not have ground truth source view depth we rely on monocular depth predictor ZoeDepth [Bhat et al. 2023]. However, depth estimators can predict object depth only up to unknown scale, hence, we recenter the projected world points to the origin and rescale them into a cube, which follows the setup used in rendering our dataset. **Guiding Inference using Partial Target View.** We observe that instead of starting the backward denoising process from pure noise, we can significantly boost the quality of the generated views by starting with a noisy version of the image $\mathbf{I}_{s \rightarrow t}$ (see Sec. 4.5).

4 EXPERIMENTS

In this section, we provide details on our training and evaluation datasets (Sec. 4.1, Sec. 4.2), compare our method to three NVS models (Sec. 4.3), and provide an ablation study (Sec. 4.5) of each component.

4.1 Training Setup.

Objaverse and Rendering Setup. To train *iNVS* we require paired data consisting of source and target views. To generate this data, we utilize the extensive *Objaverse* dataset, which contains nearly 800,000 3D assets. We employ Blender as our rendering software and begin by recentering all scenes at the origin. Additionally, we rescale the bounding box of each scene to fit within a $[-1, 1]^3$ cube. For each object in the dataset, we randomly generate 24 camera viewpoints within predefined boundaries. The radius of the viewpoint is sampled from a range of 3 to 4, and the field of view (FoV) is set to 50 degrees. Using these viewpoints, we render both the images and corresponding depth maps. We utilize the Cycles engine in Blender and employ 128 samples per ray for rendering. All images are rendered at a resolution of 512×512 pixels.

Selecting good camera poses. To add diversity to the lighting conditions, we randomly sample lighting from a collection of 100 environmental maps. These maps provide a range of indoor and outdoor lighting conditions with varying intensity levels. It's worth noting that in the Objaverse dataset, most assets are oriented with the Z-axis pointing upwards. Consequently, synthesizing the object from extreme bottom or top view angles can be challenging. For instance, when objects are placed on a platform, viewing them from below makes it almost impossible to accurately determine the opposite view without additional information. To address this issue, we empirically determine that sampling the polar angle θ from a uniform distribution between -65 and 75 degrees provides reasonably accurate views on average. We sample the azimuth angle ϕ randomly between 0 and 360 degrees.

By utilizing all 800,000 assets in the Objaverse dataset, we render a total of 19 million images to train our *iNVS* model.

Table 1. Comparison with baselines on **Google Scanned Objects** dataset.

# Method	PSNR ↑		SSIM ↑		LPIPS ↓
	mask	unmask	mask	unmask	
1 *Point-E*	8.90	12.04	0.18	<u>0.82</u>	<u>0.25</u>
2 *Shap-E*	10.39	12.18	<u>0.30</u>	<u>0.82</u>	0.29
3 *Zero-1-to-3*	14.74	<u>14.70</u>	**0.34**	**0.84**	<u>0.25</u>
4 *Original* ISD	<u>15.03</u>	13.25	0.09	0.49	0.38
5 *iNVS* (ours)	**18.95**	**19.83**	<u>0.30</u>	0.80	**0.24**

Table 2. Comparison with baselines on **Ray-traced Multiview** data.

# Method	PSNR ↑		SSIM ↑		LPIPS ↓
	mask	unmask	mask	unmask	
1 *Point-E*	7.40	10.44	0.14	**0.67**	**0.41**
2 *Shap-E*	8.35	9.74	**0.17**	<u>0.65</u>	<u>0.48</u>
3 *Zero-1-to-3*	9.09	8.29	<u>0.16</u>	0.58	0.50
4 *Original* ISD	<u>14.61</u>	<u>11.25</u>	0.09	0.27	0.65
5 *iNVS* (ours)	**16.83**	**17.82**	0.09	0.5	0.49

Table 3. Comparison with baselines on **Common Objects in 3D** dataset.

# Method	PSNR ↑		SSIM ↑		LPIPS ↓
	mask	unmask	mask	unmask	
1 *Point-E*	9.37	10.10	0.22	<u>0.72</u>	<u>0.38</u>
2 *Shap-E*	10.67	10.01	**0.33**	**0.73**	0.42
3 *Zero-1-to-3*	12.32	9.91	**0.33**	0.69	0.42
4 *Original* ISD	<u>16.43</u>	<u>13.56</u>	<u>0.24</u>	0.46	0.44
5 *iNVS* (ours)	**17.58**	**17.39**	**0.33**	0.65	**0.36**

Model and Training details. We performed fine-tuning on the pretrained Inpainting Stable Diffusion (*ISD*) v1.5 checkpoint to adapt it for our task. This model is capable of generating high-resolution images with dimensions of 512×512, utilizing a latent space of dimensions $64 \times 64 \times 4$. During the fine-tuning process, we employed a sequential training approach, consisting of three stages with separate losses previously introduced: a) denoising, b) boundary loss, and c) early steps training. Our final model was trained on 96 A100 GPUs with each stage training over 7 days.

4.2 Evaluation

Datasets. We evaluate how well *iNVS* generalizes at generating novel target views across three different datasets. Specifically, we use two synthetics datasets, *Google Scanned Objects* (*GSO*) [Downs et al. 2022; ?], and *Ray-Traced Multi-View* (*RTMV*) [Tremblay et al. 2022]. The *GSO* dataset contains nearly one thousand photorealistic 3D models, which we render using Blender following the same setup used for generating training data (described in 4.1). The *RTMV* dataset contains high quality renderings of nearly 2000 scenes from

4 different sources, and we filter out the scenes that contain *GSO* objects. Finally, we also evaluate on real videos from the *Common Objects in 3D CO3D* [Reizenstein et al. 2021] dataset, which is a dataset of 19,000 videos of common objects spanning 50 categories. **Metrics.** Following prior work [Liu et al. 2023], we compare *iNVS* and baselines with three different metrics covering different aspects of image similarity: PSNR, SSIM [Wang et al. 2004] and LPIPS [Zhang et al. 2018]. During evaluation, for every object (or scene) we sample two random views I_s and I_t along with their camera poses C_s and C_t. Next, starting from I_s we can compute relative camera transformation and generate a target view.

Masked Metrics. For PSNR and SSIM metrics we find that filtering out background (using the ground truth mask) before comparison helps to avoid spurious gains, hence we report masked metrics.

4.3 Baseline Comparisons

***Zero-1-to-3* [Liu et al. 2023]** is an image and camera pose conditioned diffusion model which leveraged a pretrained Stable Diffusion called Image Variations [Labs 2023] and finetuned it on Objaverse renders. Unlike our method, Zero-1-to-3 is trained to generate novel views from scratch, and is prone to cause inconsistency between source and target views (see results for more details). We use the official codebase[2] and checkpoints with our datasets for evaluation. ***Point-E* [Nichol et al. 2022]** is an image-conditioned diffusion model which operates over 3D point clouds to generate objects. We use the official codebase and checkpoint[3], and use the settings mentioned in the paper to generate point clouds with 4,000 points. We render the point cloud from the target viewpoints for novel views.

***Shap-E* [Jun and Nichol 2023]** is a conditional generative model for 3D, which directly outputs the parameters of implicit functions that can be rendered directly as neural radiance fields. It is a two-stage model that involves generating a latent code for each 3D asset and then uses a diffusion model to denoise this latent code. We use the official codebase and checkpoint[4], and render outputs as neural radiance field from the target viewpoint.

Our method achieves good PSNR and comparable LPIPS. We find that our method achieves the higher PSNR and comparable LPIPS scores compared to all other baselines on the Google Scanned Objects (*GSO*) and Common Objects in 3D (*CO3D*) benchmarks (Table 1 and 2). This indicates that our method performs well in terms of noise reduction and perceptual similarity in both synthetic and real data scenarios. The *CO3D* dataset consists of real-world views captured from free-form videos, while the *GSO* dataset contains virtual scans of 3D photorealistic assets. On the Ray-traced Multiview (*RTMV*) dataset, we find that our method is able to outperform all baselines in PSNR, but falls short on the SSIM and LPIPS metrics (shown in Table 2). We attribute this low performance to out-of-distribution variations in lighting across our rendering setup (described in Section 4.1 compared to *RTMV*).

Structural Similarity is compromised in generated views. It is worth mentioning that our method consistently underperforms

---

[2]https://github.com/cvlab-columbia/zero123
[3]https://github.com/openai/point-e
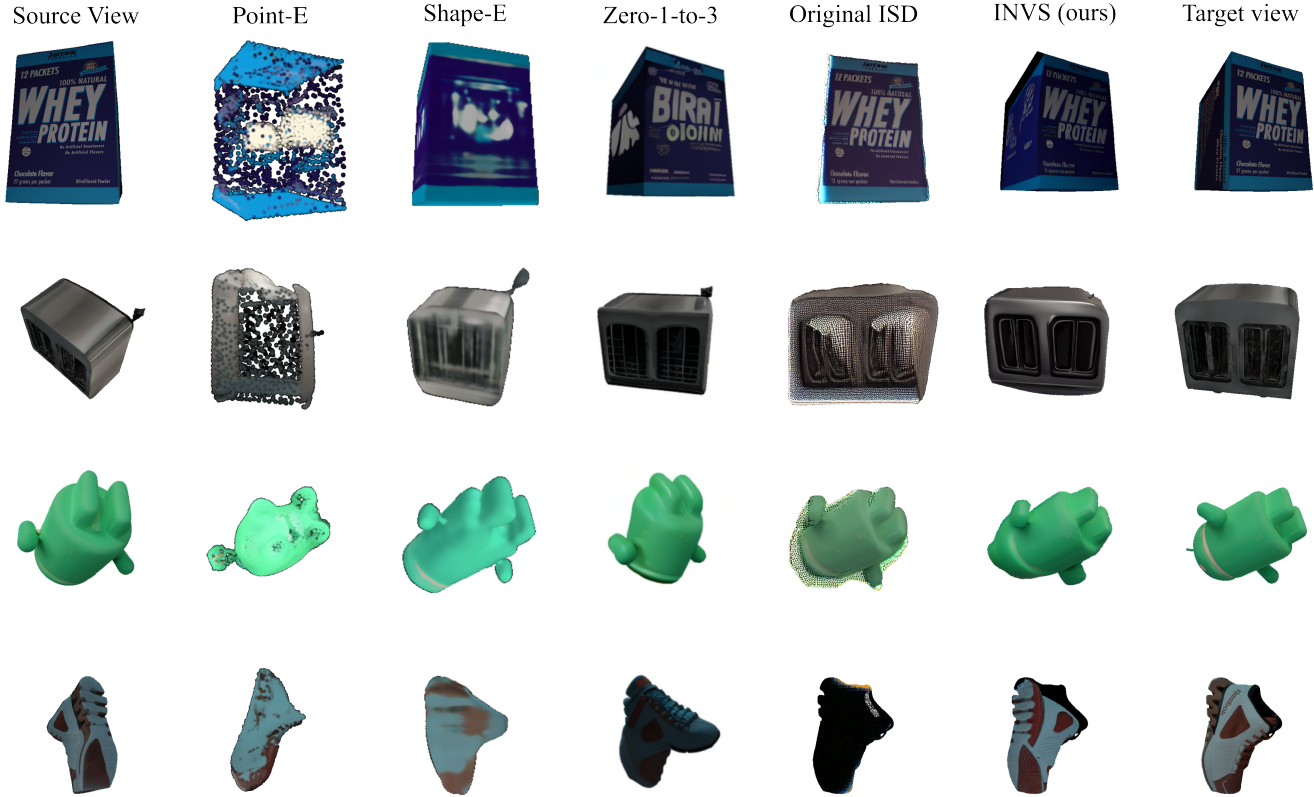[4]https://github.com/openai/shap-e

Fig. 4. **Comparison with SoTA methods on NVS task (GSO dataset).** The first column is input, columns two to four are baselines: *Point-E* [Nichol et al. 2022], *Shap-E* [Jun and Nichol 2023] and *Zero-1-to-3* [Liu et al. 2023]. Fifth column is untrained *ISD*, and last two columns is *iNVS* and ground truth.
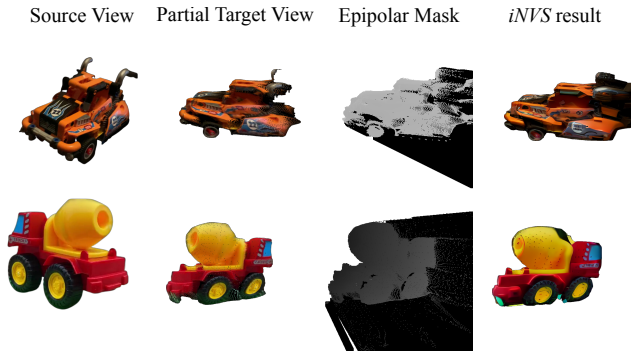

Fig. 5. **Partial target view and epipolar masks on CO3D dataset.** The first column shows input. Second column show the warped source view. The third column demonstrates the smooth inpainting epipolar mask. Notice that the inpainting mask for cement dump truck (second row) is much darker compared to the race car (first row) due to larger angle variation (details in Sec. 3.2). Last column shows generated result.

on the SSIM metric across all datasets. We find that this occurs primarily due to misalignment in monocular depth estimator. We observe that under significant viewpoint variations, the monocular depth estimator fails to generate consistent depth across different parts of the objects. This inconsistency leads to distortions in the generated images and lower SSIM scores (see Section 4.6).

Masked metrics help disambiguate performance across baselines. We notice that Shap-E and Point-E often produce tiny objects, and their white background pixels (matching the target) lead to majority of their unmasked gains, thus outperforming Zero123 quantitatively. However, using masked metrics we notice these trends change (see PSNR and SSIM in Tables 2 and 3).

4.4 Qualitative Results

We present the visualization of the results obtained from our method and the baselines in Figure 4 and draw the following conclusions:

Preservation of Text and Fine Details. We employ monocular depth to unproject pixels into 3D space and warp them to the target viewpoint. This technique enables us to preserve a significant amount of textual information and finer details across multiple viewpoints. This is clearly demonstrated in Row 1, where our method successfully retains the text "Whey Protein" between the source and target views, unlike the *Zero-1-to-3* baseline. Additionally, our method shows less distortion in the structural details of objects, as seen in Row 2, where *Zero-1-to-3* alters the toaster oven into grills, while our method accurately preserves this detail. The benefits of

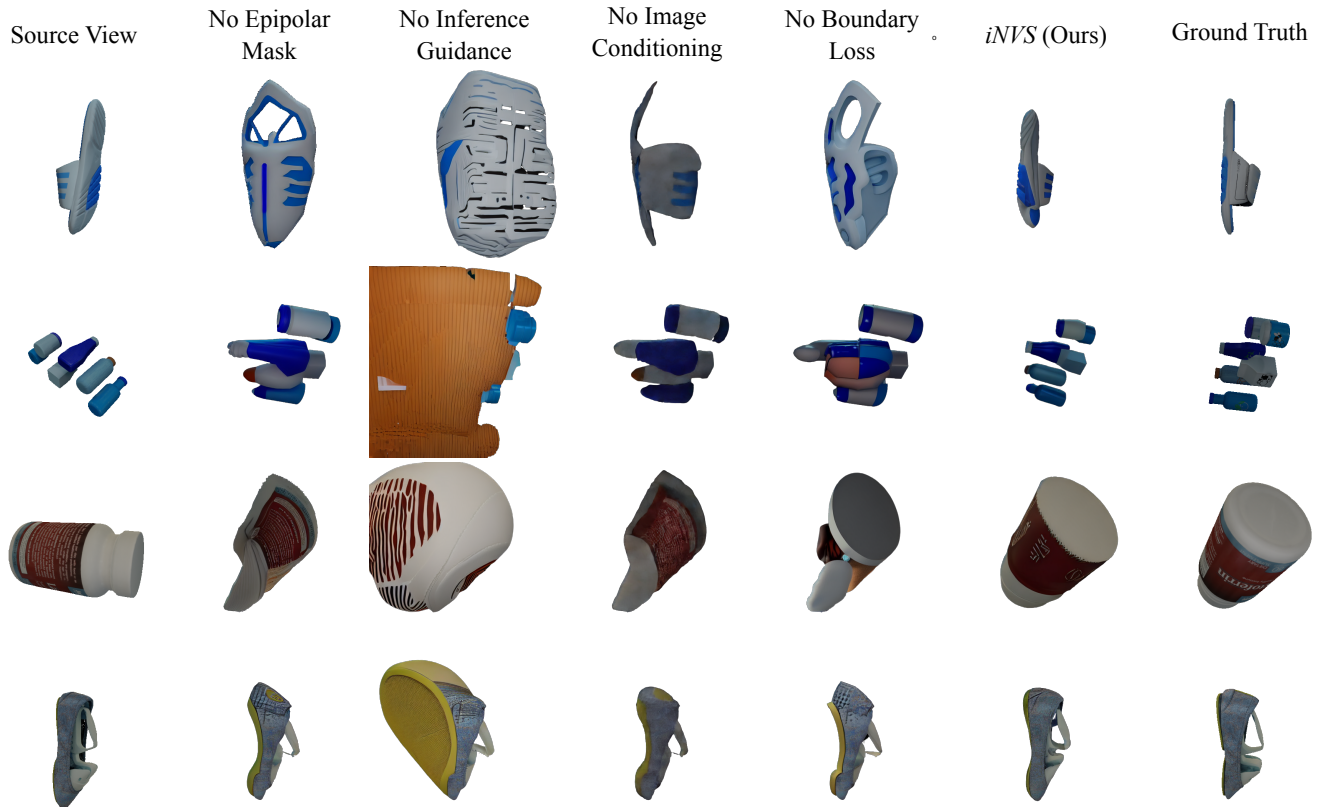| Source View | No Epipolar Mask | No Inference Guidance | No Image Conditioning | No Boundary Loss | *iNVS* (Ours) | Ground Truth |

Fig. 6. **Ablation Study of *iNVS* on GSO dataset.** We describe each column from left to right. First, we show input image. Second, we show generation without the epipolar mask. Third, we show generation without inference guidance (Section 3.4). Fourth, we skip source-view conditioning via CLIP during inference. Fifth, we show a variant of our method without boundary loss L_W. The last two columns show result from *iNVS* and ground truth.

this technique is further highlighted in Fig. 5, where we show that for many viewpoints most of the partial target view can be directly reused to significantly simplify the job of inpainting network.

Consistent synthesis for single object across multiple views. In Figures 8 and 9, we show NVS results across many different objects from a fixed source view, and randomly selecting six target views. We find our generations to be largely coherent across viewpoints, and more stochastic under large viewpoint variations.

Faithful Compliance with Viewpoint Variations: Another challenge in novel view synthesis is maintaining control over the generated views, particularly when dealing with significant viewpoint variations. Rows 3 and 4 exemplify this issue, showing that *Zero-1-to-3* struggles to exercise accurate control over the generated images, resulting in the android and the shoe being created from incorrect viewpoints.

Outperforming 3D Diffusion Models: Our method surpasses the performance of the *Point-E* and *Shap-E* baselines, both of which are variants of 3D diffusion models. We believe that utilization of a large, pretrained inpainting model in our method contributes to the generation of visually superior results.

4.5 Ablation Study

We conduct a series of ablation experiments and analysis to evaluate the effectiveness of our proposed changes and additional training. The quantitative performance results for all metrics are reported in Table 4, and visuals are presented in Figure 6.

Epipolar Mask helps to constrain NVS generation. The epipolar mask allows us to control the extent of the inpainting necessary in the warped image, as depicted in Figures 2 and 3. When we omit the use of this mask, the inpainting model faces challenges in understanding the relative orientation between the source view and the target view. As a result, the performance of the model is compromised, leading to distorted and exaggerated generated images.

Guiding denoising inference with partial target view yields significant benefits. We find that during the 500 step DDIM inference *iNVS* generates a rough outline of the target view within the first 100 steps. Given that we have a partial target view available after warping the source pixels to the target viewpoint, we utilize it as guidance. Specifically, we replace the output of the first 10 DDIM steps with a noised version of the partial target view, similar to Repaint [Lugmayr et al. 2022]. This approach proves to be immensely

Table 4. **Ablation Study.** First row is our method *iNVS*. The second row shows our method where the epipolar mask is replaced with a full mask that covers all non-splatted pixels. The third row is our method without using a partial target view as diffusion input in the first steps as guidance. The fourth row is our method without conditioning on the source image. The fifth row demonstrates our method without L_W. Finally we highlight our model with the original training schedule. Also for the reference we show the performance of the original unaltered *ISD*.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
iNVS (ours)	**19.83**	**0.8**	**0.24**
- epipolar mask	$17.39_{-2.44}$	$0.65_{-0.15}$	$0.36_{+0.12}$
- inference guidance	$17.48_{-2.35}$	$0.70_{-0.1}$	$0.31_{+0.07}$
- image conditioning	$16.57_{-3.26}$	$0.70_{-0.01}$	$0.30_{+0.06}$
- boundary loss	$19.10_{-0.73}$	$0.77_{-0.03}$	$0.27_{+0.03}$
- early steps training	$19.70_{-0.13}$	$0.78_{-0.03}$	$0.26_{+0.02}$
Original ISD	$13.25_{-6.58}$	$0.49_{-0.31}$	$0.38_{+0.14}$

helpful in preventing the inpainting model from generating arbitrary boundaries, shown in Figure 6, third column.

Without image conditioning our model is unable exploit source view information well. When there is a significant variation in viewpoint between the target and source images, the warped image becomes less informative for the inpainting process. In such scenarios, it becomes essential for the inpainting model to heavily rely on the source view and the learned 3D priors for accurate autocompletion. By removing the conditioning on the source view, the model's ability to generate high-quality and consistent views is compromised, as demonstrated in Figure 6, fourth column.

Boundary loss enhances generations. Furthermore, we find that enhancing the boundary loss improves the generation quality leading to consistent visuals, evident from Figure 6, fifth column.

4.6 Current Limitations

Imprecise monocular depth can lead to structure and texture problems. We notice that ZoeDepth [Bhat et al. 2023] can generate depth maps that distorts flat surfaces, which leads to unrealistically deformed surfaces or incorrect texture predictions. We put visuals of these in failure modes #1 and #2 in Figure 7.

Inference guidance can occasionally lead to incomplete or flipped output images. Recall from Section 3.4 that we use the noisy partial target view for the first 10 DDIM steps. However, since this view is incomplete and may contain flipped pixels (under high camera variations), this trick occasionally creates incomplete outputs or lead to reversed-view generated images. We put visuals of these in failure modes #3 and #4 in Figure 7.

5 CONCLUSION

In this work we propose an approach for novel view synthesis that provides a significant advancement in terms of quality of the results and the coverage of the different object categories. Our approach combines recent advancement of diffusion models with epipolar geometry. By training on a large scale *Objaverse* dataset we were able to re-purpose Inpainting Stable Diffusion for the novel view

synthesis task. Surprisingly we found that after finetuning our method gained an understanding of the underlying shape, even for the objects that were not seen during the training. Our approach demonstrates sizable improvement over state-of-the-art novel view synthesis methods, especially when considering texture preservation from the input image. One limitation of our approach is inability to generate consistent textures in the regions that are not visible in the original image. This however opens the exciting opportunity for future research, which can explore auto-regressive schemes of novel view generation.

ACKNOWLEDGMENTS

| Input View | Unprojected View | Generated | Ground Truth | Input View | Unprojected View | Generated | Ground Truth |

Failure Mode #1: Deformed partial view.

Failure Mode #3: Inference guidance hinders completion.

Failure Mode #2: Holes in partial view blend into texture.

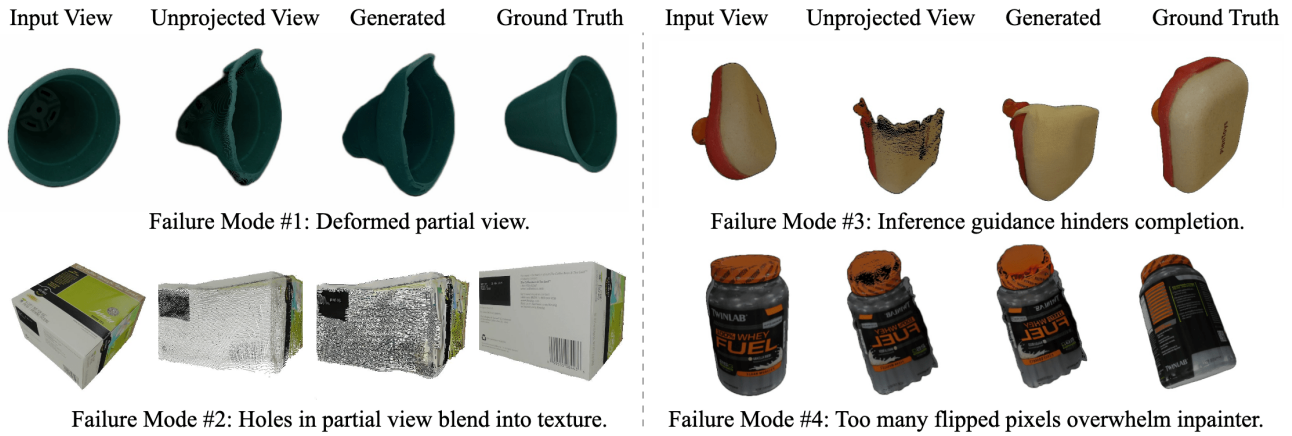Failure Mode #4: Too many flipped pixels overwhelm inpainter.

Fig. 7. **Failure Modes. Left:** Imperfect depth maps cause issues in structure and texture. **Right:** Inference time tricks can occasionally hinder generations.

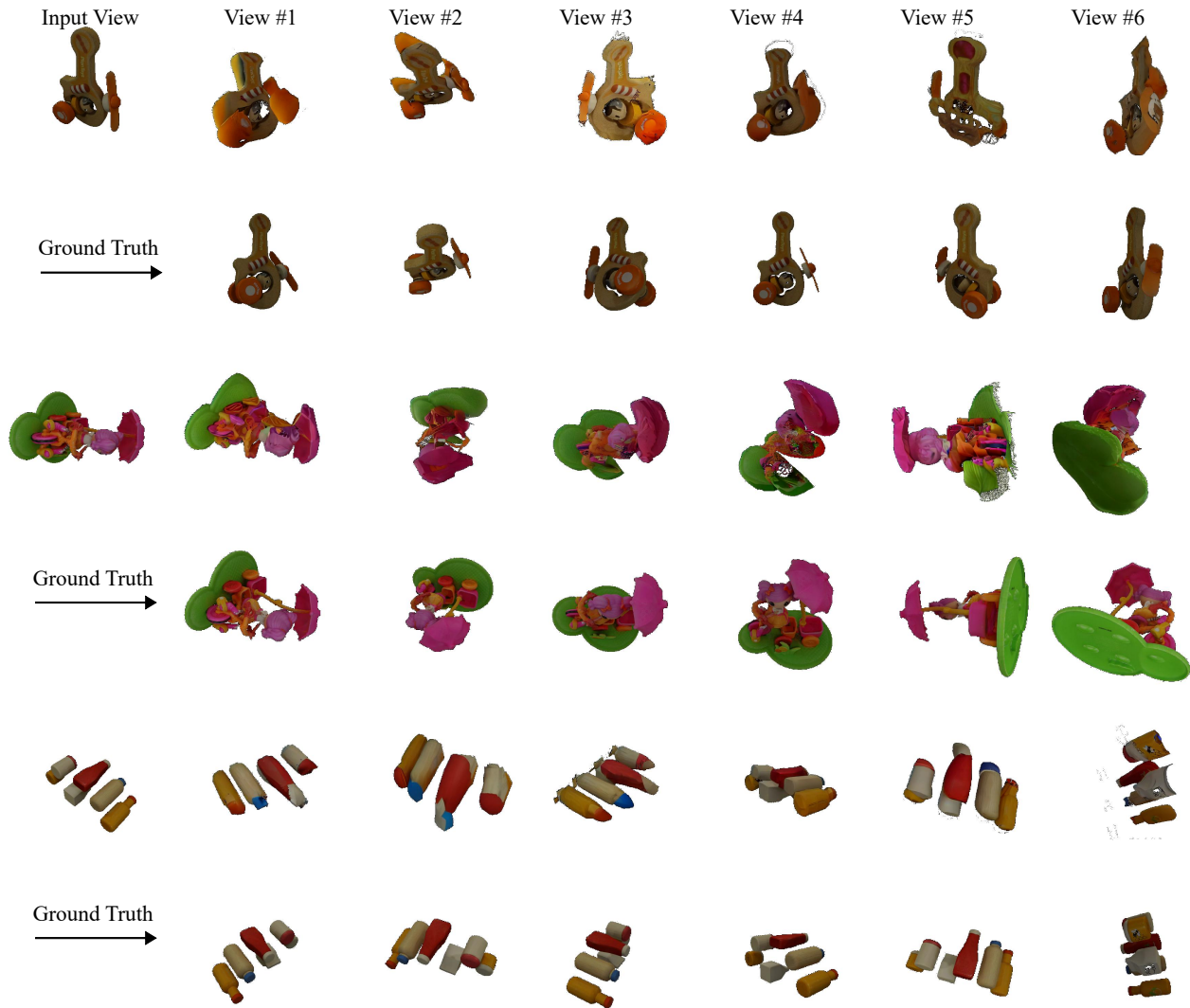| Input View | View #1 | View #2 | View #3 | View #4 | View #5 | View #6 |

Fig. 8. **Multiple novel views from single image.** We show six randomly sampled camera views given an input image, and corresponding ground truth.
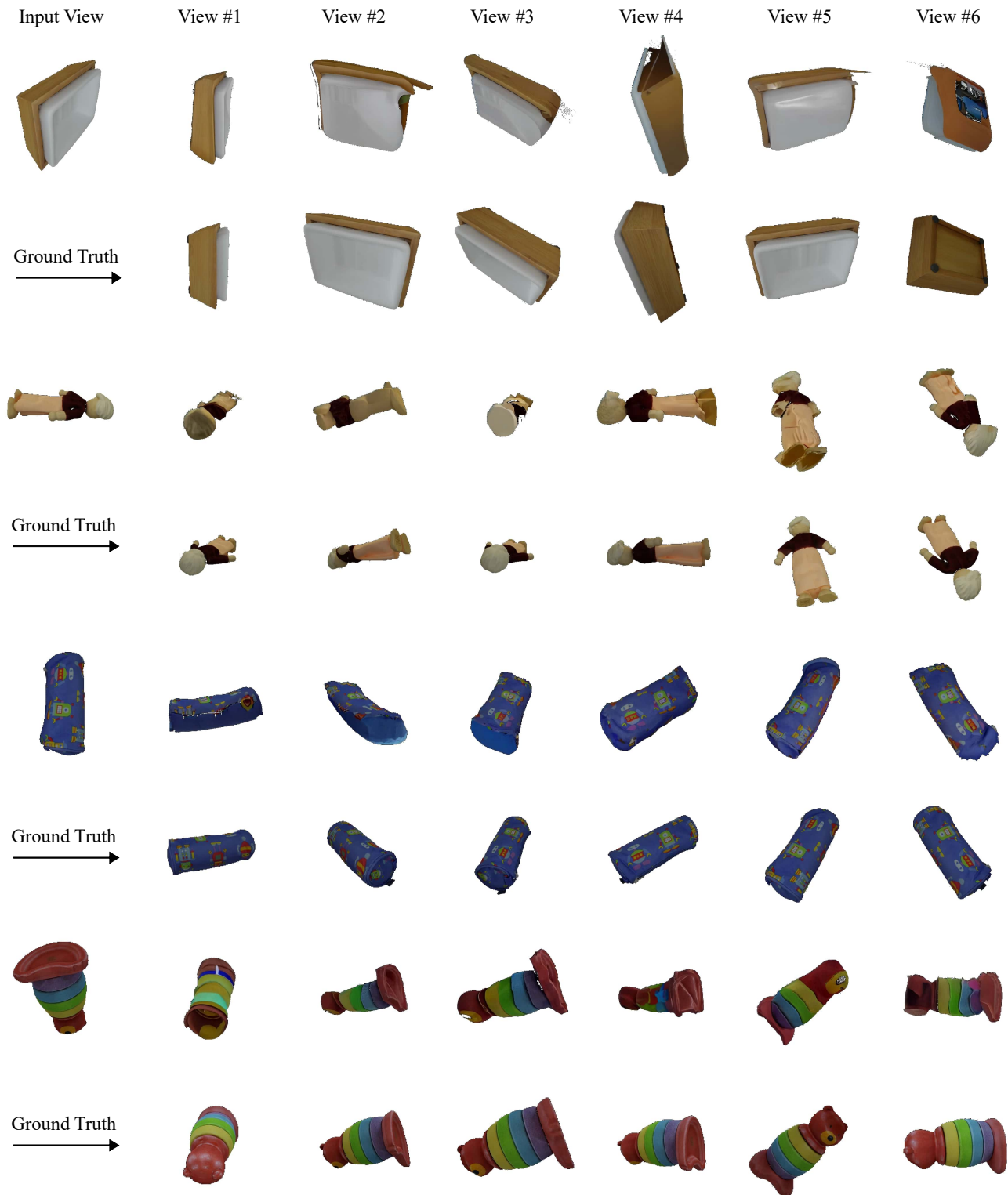
Fig. 9. **Multiple novel views from single image.** We show six randomly sampled camera views given an input image, and corresponding ground truth.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE International Conference on Computer Vision*.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023).

Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE International Conference on Computer Vision*.

Lucy Chai, Richard Tucker, Zhengqi Li, Phillip Isola, and Noah Snavely. 2023. Persistent Nature: A Generative Model of Unbounded 3D Worlds. In *CVPR*.

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. 2023. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision*.

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. 2023b. Text2Tex: Text-driven Texture Synthesis via Diffusion Models. In *Proceedings of the IEEE International Conference on Computer Vision*.

Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. 2021. Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems* (2021).

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *Proceedings of the IEEE International Conference on Computer Vision*.

Shenchang Eric Chen and Lance Williams. 1993. View Interpolation for Image Synthesis. In *Special Interest Group on Computer Graphics and Interactive Techniques*.

Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. 2019. Extreme view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*.

Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. 1996. Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach. In *Special Interest Group on Computer Graphics and Interactive Techniques*.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kangle Deng, Gengshan Yang, Deva Ramanan, and Jun-Yan Zhu. 2023. 3d-aware conditional image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Emily Denton and Rob Fergus. 2018. Stochastic video generation with a learned prior. In *Proceedings of the International Conference on Machine Learning*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* (2021).

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *Proceedings of the IEEE International Conference on Robotics and Automation*.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. 2016. Unsupervised learning for physical interaction through video prediction. *Advances in Neural Information Processing Systems* (2016).

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. 2021. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE International Conference on Computer Vision*.

Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. 2023. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *Proceedings of the International Conference on Machine Learning*.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. *Advances in Neural Information Processing Systems* (2018).

Pengpeng Hu, Edmond Shu-Lim Ho, and Adrian Munteanu. 2021a. 3DBodyNet: fast reconstruction of 3D animatable human body shape from a single commodity depth camera. *IEEE Transactions on Multimedia* (2021).

Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. 2021b. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*.

Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 2018. 3D face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing* (2018).

Heewoo Jun and Alex Nichol. 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463* (2023).

Ira Kemelmacher-Shlizerman and Ronen Basri. 2010. 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2010).

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. 2021. Pathdreamer: A world model for indoor navigation. *ICCV*.

Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. 2022. NeROIC: Neural Rendering of Objects from Online Image Collections. In *Special Interest Group on Computer Graphics and Interactive Techniques*.

Lambda Labs. 2023. Stable Diffusion Image Variations. https://huggingface.co/spaces/lambdalabs/stable-diffusion-image-variations. Accessed on 2023-05-22.

Wonkwang Lee, Whie Jung, Han Zhang, Ting Chen, Jing Yu Koh, Thomas Huang, Hyungsuk Yoon, Honglak Lee, and Seunghoon Hong. 2021. Revisiting hierarchical approach for persistent long-term video prediction. In *Proceedings of the International Conference on Learning Representations*.

Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. 2023. Generative Image Dynamics. *arXiv preprint arXiv:2309.07906* (2023).

Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. 2022a. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *Proceedings of the European Conference on Computer Vision*.

Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. 2022b. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *Proceedings of the European Conference on Computer Vision*.

Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. 2023. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. 2021. Infinite Nature: Perpetual View Generation of Natural Scenes from a Single Image. In *Proceedings of the IEEE International Conference on Computer Vision*.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. In *Proceedings of the IEEE International Conference on Computer Vision*.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Aniruddha Mahapatra and Kuldeep Kulkarni. 2022. Controllable Animation of Fluid Elements in Still Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3667–3676.

Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G Derpanis, and Igor Gilitschenski. 2023. Reference-guided Controllable Inpainting of Neural Radiance Fields. In *Proceedings of the IEEE International Conference on Computer Vision*.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751* (2022).

Simon Niklaus and Feng Liu. 2020. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *Proceedings of the International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. 2023. DreamBooth3D: Subject-Driven Text-to-3D Generation. *arXiv preprint arXiv:2303.13508* (2023).

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. 2021. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision.*

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. TEXTure: Text-Guided Texturing of 3D Shapes. In *Special Interest Group on Computer Graphics and Interactive Techniques.*

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention.*

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* (2022).

Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. 2022. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* (2022).

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

Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 2020. 3D Photography using Context-aware Layered Depth Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019a. First order motion model for image animation. *Advances in Neural Information Processing Systems* (2019).

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019b. Animating Arbitrary Objects via Deep Motion Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. 2023. LDM3D: Latent Diffusion Model for 3D. *arXiv preprint arXiv:2305.10853* (2023).

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *Proceedings of the IEEE International Conference on Computer Vision.*

Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, and Stan Birchfield. 2022. RTMV: A Ray-Traced Multi-View Synthetic Dataset for Novel View Synthesis. *Proceedings of the European Conference on Computer Vision Workshop* (2022).

Richard Tucker and Noah Snavely. 2020. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. 2017. Decomposing motion and content for natural video sequence prediction. In *Proceedings of the International Conference on Learning Representations.*

Carl Vondrick and Antonio Torralba. 2017. Generating the Future With Adversarial Transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. 2017. PredRNN: Recurrent Neural Networks for Predictive Learning Using Spatiotemporal LSTMs. *Advances in Neural Information Processing Systems* (2017).

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* (2004).

Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. 2021. NeRF−−: Neural Radiance Fields Without Known Camera Parameters. *arXiv preprint arXiv:2102.07064* (2021).

Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. 2022. Novel View Synthesis with Diffusion Models. In *Proceedings of the International Conference on Learning Representations.*

Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. 2020. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. 2019. Compositional video prediction. In *Proceedings of the IEEE International Conference on Computer Vision.*

Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. 2022. Generating videos with dynamics-aware implicit generative adversarial networks. In *Proceedings of the International Conference on Learning Representations.*

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020).

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Zhizhuo Zhou and Shubham Tulsiani. 2023. SparseFusion: Distilling View-conditioned Diffusion for 3D Reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

A   TRAINING HYPERPARAMETERS

#	Hyperparameters	Value
1	Base Learning rate	1e-5
2	Learning rate decay	N/A
3	Loss Type	L1
4	Source-view / Inpaint Mask dropout	0.05
5	Classifier-free guidance	9.0
6	Effective batch size	1152
7	Effective batch size	1152
8	DDIM Steps	300/500
9	Partial View Guidance Steps	10
10	Boundary Loss Weight	2.0
11	CLIP Frozen	False
12	Renders background color	Black
13	Image Resolution	512
14	Learning rate linear warmup	100 steps

Table 5. Hyperparameter choices for *iNVS*.

B   EXPERIMENT WITH DIFFUSION-BASED MONOCULAR DEPTH NETWORK

Idea and Setup. Since incorrect predictions from monocular depth network lead to failure cases, we trained a separate diffusion-based monocular depth prediction network on the Objaverse dataset. For this, we simply inflate the first convolution layer of Stable Diffusion [Rombach et al. 2022] to take input RGB image as additional conditioning via concatenation. The model is trained with L1 loss to generate monocular depths, and it is initialized with pretrained text-to-image Stable Diffusion weights.

Result. We found that the compression in latent-space of diffusion model due to the VAE caused issues in the predicted depth maps. This in turn led to *jittery outputs and reprojections* leading to subpar performance. Since we trained only on Objaverse dataset, we did not require explicit rescaling and recentering when using predicted depth. We believe improvement in this setup could be achieved by retraining the Autoencoders to reconstruct depth maps, similar to [Stan et al. 2023]. *We put visuals of this experiment in supplementary video.*