The More You See in 2D, the More You Perceive in 3D

Xinyang Han*
UC Berkeley
hanxinyang66@gmail.com

Zelin Gao*
Zhejiang University
jamesgzl@zju.edu.cn

Angjoo Kanazawa
UC Berkeley
kanazawa@berkeley.edu

Shubham Goel†
Avataar
shubhamgoel@avataar.ai

Yossi Gandelsman†
UC Berkeley
yossi_gandelsman@berkeley.com

Abstract

Humans can infer 3D structure from 2D images of an object based on past experience and improve their 3D understanding as they see more images. Inspired by this behavior, we introduce SAP3D, a system for 3D reconstruction and novel view synthesis from an arbitrary number of unposed images. Given a few unposed images of an object, we adapt a pre-trained view-conditioned diffusion model together with the camera poses of the images via test-time fine-tuning. The adapted diffusion model and the obtained camera poses are then utilized as instance-specific priors for 3D reconstruction and novel view synthesis. We show that as the number of input images increases, the performance of our approach improves, bridging the gap between optimization-based prior-less 3D reconstruction methods and single-image-to-3D diffusion-based methods. We demonstrate our system on real images as well as standard synthetic benchmarks. Our ablation studies confirm that this adaption behavior is key for more accurate 3D understanding.[1]

1. Introduction

Imagine you are shopping online and see a picture of the bunny in Figure 1 you want to buy. You can build a rough mental 3D model of the object based on the first image alone following your understanding of what bunnies look like, but you can only guess the geometry and appearance of the unseen parts. When you see more images, you implicitly estimate the camera viewpoints and consolidate information from all the views to build a better 3D understanding of the object. To accomplish this task you combine 2D priors coming from your vast previous visual experience with the actual observations you see, to understand the 3D object. As you see more examples, your 3D understanding improves.

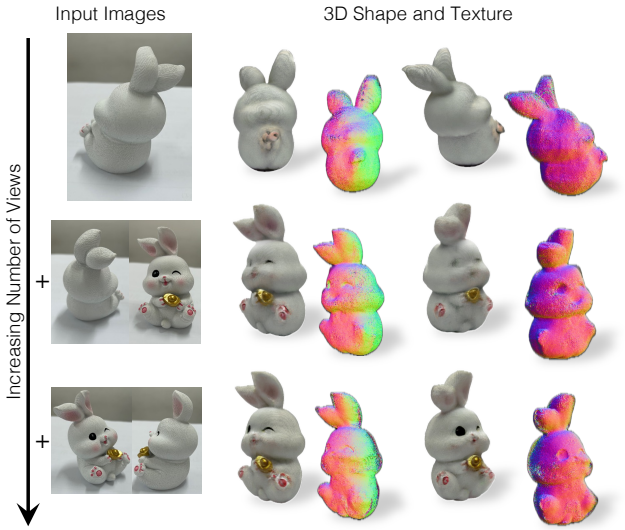

Figure 1. **3D from one or more unposed views.** Our system reconstructs the 3D shape and texture of an object with a variable number of real input images. The first, second, and third rows show reconstructions from 1, 3, and 5 input images. The quality of 3D shape and texture improves with more views.

In this paper, we present a system that follows the same principles - it handles a variable number of images, incorporates learned priors from large-scale 2D image data, and improves as it sees more captures of the object. Specifically, starting from an arbitrary number of input images *without camera poses*, we use a pre-trained generative image model, adapt it to the object of interest, and consolidate the information, enabling consistent 3D reconstruction and novel view synthesis (NVS). Therefore, we call our system **S**ee-**A**dapt-**P**erceive, or SAP3D in short.

To reconstruct real-world 3D objects from only a few images (e.g. 3 photos), our system relies on priors from models that are pre-trained on large-scale datasets. We incorporate a generative model that was trained on a large-scale image

* Equal contribution. † Equal contribution.
[1]Project page: https://sap3d.github.io/

Methods	# Input Views	Handles Unposed Images	3D Recon.	Leverages 2D Data
MV-Dream [45]	0 (text)			✓
Zero123 [23]	1		✓	✓
SyncDreamer [24]	1		✓	✓
PixelNeRF [55]	≥ 1		✓	
SparseFusion [60]	≥ 1		✓	✓
RelPose++ [20]	≥ 1	✓	✓*	
Ours (SAP3D)	≥ 1	✓	✓	✓

Table 1. **Related approaches:** A comparison of different methods, highlighting the key difference in inputs needed, outputs recovered, and type of data used. Only our system (SAP3D) can reconstruct 3D from a variable number of unposed input images. *Camera outputs from RelPose++ have been shown useful for 3D reconstruction, but only on a limited number of examples with 7 or more images.

dataset (LAION [44]) and distill it into a 3D model, akin to Zero-1-to-3 [23]. We also train a camera pose estimator [20] on a large 3D object dataset (Objaverse [3]), to enable its generalization to a diverse set of objects. The rough camera poses predicted by these models provide us with enough signal for exploiting the generative model prior for 3D reconstruction. We believe that future improvements in any of the pre-trained models will directly improve our system.

While other methods that rely on 2D priors use a fixed number of input images, SAP3D can incorporate information from a varying number of views. To achieve this, we adapt the 2D generative model at test-time by fine-tuning it on the input views and simultaneously refining their estimated relative camera poses. The resulting instance-specific generative model can be used for sampling novel views of the instance and for distillation into a 3D model (e.g. via NeRF [27]). Thus, our system bridges the gap between single-view and multi-view reconstruction. It can recover 3D geometry and appearance from an *arbitrary number* of un-posed images, as shown in Figure 1.

As the number of input images grows, the quality of the 3D reconstruction and the estimated camera poses improve. Similarly, the consistency of the novel views sampled from the obtained instance-specific generative model improves with more views. We illustrate this qualitatively for real images and quantitatively for 3D models from the GSO dataset [4].

Finally, we present an extensive ablation study of different components of our system. We show that test-time adaption, as well as training the camera pose estimator on a large-scale 3D dataset, improves 3D reconstruction, camera pose predictions, and novel view synthesis quality.

2. Related Work

Instance-specific 3D Reconstruction. There is extensive literature on 3D reconstructing objects and scenes from input images. COLMAP [43] is the culmination of a long line of classical work on SfM [12, 31] and MVS [6, 7]. NeRF [26], VolSDF [54], SRN [47] are modern neural counterparts. However, these approaches do not rely on priors that can be learned from large-scale datasets. Therefore, they require many input views, where camera poses are either known or can be found using SfM. Some approaches [10, 13, 29, 48, 56] attempt to reconstruct from a limited number of views but they need camera poses and cannot reconstruct unseen parts because of the lack of data-driven priors.

Single-view 3D Reconstruction. Several works learn priors that enable full 3D reconstruction, including unseen areas. They are often formulated as single-view reconstruction of volumetric occupancy [1, 25, 42, 59], meshes [8, 11, 51] and category-specific shape deformation models [5, 9, 14, 15, 19, 53], all from a single image.

An alternative two-stage approach for single-view 3D reconstruction relies on implicit 3D priors that 2D generative models learn. First, a 2D diffusion model [35, 39, 41] is pre-trained on a large-scale dataset of images. Second, the pre-trained diffusion model is used as a score function to supervise the optimization of an instance-specific 3D model (e.g. NeRF [27]) via a score distillation sampling (SDS) loss [32] or a similar variant [49]. Zero-1-to-3 [23] trains a 2D diffusion model conditioned on a relative camera viewpoint and an input image, and distills the diffusion model into a NeRF via SDS loss by conditioning the diffusion model on a single input image. However, the sampled generated novel views are not multiview-consistent. Other methods [24, 45] train a diffusion model to generate multiview-consistent images from a single-view image/text, before the distillation. As shown in Table 1, all these approaches utilize a single input at test-time, and can not improve the 3D reconstruction by utilizing multiple unposed images of the captured object.

Few-view 3D Reconstruction Single-view and multi-view reconstruction are only two ends of the input spectrum. We would like systems that can reconstruct an approximate shape from a single input image but progressively improve with more views. LSM [16] was one of the first to show such behavior in a learning framework, by merging features from a variable number of input images to predict a 3D voxel grid. Recent works [36, 55] also achieved this with an implicit volumetric representation. However, these approaches require accurate camera poses, and even then they are susceptible to generating blurry outputs due to their deterministic nature.

A recent line of work [20, 46, 50, 57] learns to predict relative camera poses from a few input images with small overlap. Even though they are motivated by the goal of sparse view 3D reconstruction, they only show minimal proof-of-concept experiments for 3D reconstruction using their output poses. Our system, SAP3D, reconstructs 3D

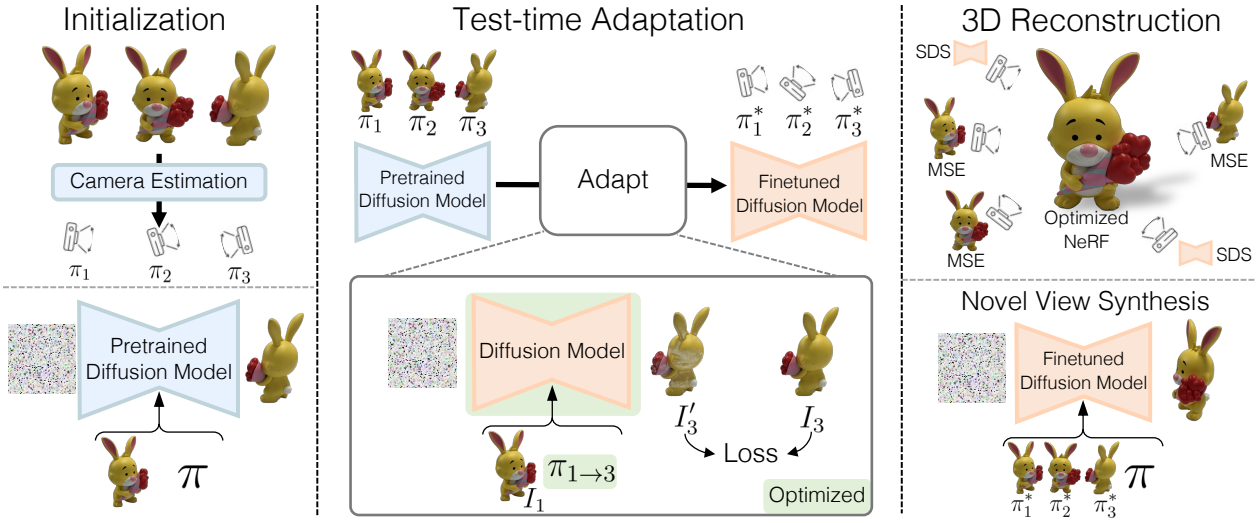

Figure 2. **Overview of SAP3D.** We first compute coarse relative camera poses using an off-the-shelf model. We fine-tune a view-conditioned 2D diffusion model on the input images and simultaneously refine the camera poses via optimization. The resulting instance-specific diffusion model and camera poses enable 3D reconstruction and novel view synthesis from *an arbitrary number of input images*.

shape from only a few images without the need for input camera poses.

Test-Time Adaptation. Test-time adaptation aims to fine-tune a pre-trained model on test example(s) before making a prediction. In the context of generative models, and in our case, test-time adaptation is often used for personalizing the model and improving the generation quality on the test distribution. Pivotal Tuning [37] and MyStyle [30] allow for real image editing by adapting a generative adversarial network (GAN) to reconstruct face image(s) of a single person. DreamBooth [40] and CustomDiffusion [18] obtain a personalized text-conditioned model that can synthesize novel images of a subject, by fine-tuning a pre-trained text-to-image diffusion model on a few images of this subject. Dreambooth3D [34] distills a personalized text-conditioned model into a NeRF, to enable personalized text-to-3D generation. Similarly, we use test-time adaptation to obtain an *instance-specific* diffusion model, by fine-tuning a view-conditioned diffusion model on a few images of it. Differently from all these methods, our goal is 3D reconstruction and novel view synthesis, which requires preserving the 3D structure of the captured instance.

3. SAP3D

We start by describing our three-stage system for 3D object reconstruction and novel view synthesis from a few unposed images, as shown in 2. Section 3.1 presents the initialization stage of relative camera poses between the input images and a view-conditioned 2D diffusion model that provides implicit 3D priors for the 3D reconstruction. Section 3.2 describes the refinement stage of the relative camera poses together with the diffusion model to obtain instance-specific 3D priors via test-time optimization. Finally, we show how the refined camera poses and the instance-specific diffusion model are utilized for 3D reconstruction (Section 3.4) and for NVS (Section 3.3).

3.1. Initialization

Initial camera poses. Given only a small set of k images of an object $S = \{I_1, ..., I_k\}$, $I_i \in \mathbb{R}^{H \times W \times 3}$, we estimate their corresponding camera poses π_i following [20]. We first compute crude estimates of the distribution of relative poses $\pi_{i \to j}$ between each pair of images I_i, I_j as a distribution on relative camera rotations $R_{i \to j} \in \mathbb{R}^{3 \times 3}$ and translations $T_{i \to j} \in \mathbb{R}^3$. We then find a consistent set of poses π_i that maximize the likelihood under the predicted distribution. We further refine these camera poses in Section 3.2.

Initial view-conditioned 2D diffusion model. As 3D reconstruction from a small number of images is an ill-posed problem, we rely on 3D priors that are learned from large-scale image datasets. We use a pre-trained view-conditioned diffusion model F that takes an input object image I and a relative camera pose π as conditioning signals, and is trained to generate a new image I' of the object under this camera transformation. F is first pre-trained on an internet-scale dataset of text-image pairs, and then adapted via fine-tuning for view-conditioning, as described in Zero-1-to-3 [23]. F consists of a camera pose conditioning network c_ϕ and a learned denoiser ϵ_θ. It is based on stable diffusion [38] and operates in the latent space of a pre-trained VAE with a fixed encoder $\mathcal{E}$ and decoder $\mathcal{D}$. The diffusion process adds noise

to the encoded latent $z = \mathcal{E}(I')$ to produce a noisy latent z_t with an increasing noise level over timesteps t. The denoiser ϵ_θ learns to predict the noise added to the noisy latent z_t given I and encoded camera pose $c_\phi(\pi)$. Formally, the optimization objective is:

$$\min_{\theta,\phi} \mathbb{E}_{I,z,\pi,t,\epsilon\sim\mathcal{N}(0,1)} \left\|\epsilon - \epsilon_\theta\left(z_t, t, c_\phi(\pi), I\right)\right\|_2^2 \quad (1)$$

where ϵ is the noise added to z to create z_t.

3.2. Test-time optimization

To incorporate information from a varying number of views into our system, we apply test-time optimization on the initial diffusion model and camera poses. Our goals are to (1) obtain an instance-specific view-conditioned diffusion model F_S and (2) improve the camera pose estimates for the images for downstream novel view synthesis and 3D reconstruction.

Finetuning the diffusion model. We adapt the diffusion model F to our images S during test-time. We fine-tune F on the images in S using the estimated relative camera poses from between every image pair from the initialization step (resulting in $k(k-1)$ training examples). This stage can be applied with any k, as each optimization step requires pairs of images.

Optimizing camera poses. During the finetuning process, we also refine the initial estimates for π_i by directly backpropagating to these parameters. The view-conditioned 2D diffusion model F, which can synthesize novel views of an object, models the probability distribution $P(I'|I,\pi)$ of novel views I' of an image I conditioned on the relative camera pose π. Therefore, by Bayes Rule, it also implicitly models the distribution of camera poses given images I and I'. We exploit F to further optimize our camera poses. While fine-tuning F, we simultaneously optimize our camera pose estimates, by backpropagating gradients into their parameters. Formally, the test-time adaptation objective is:

$$\min_{\theta,\phi,\pi_i} \mathbb{E}_{i,j,t,\epsilon} \left\|\epsilon - \epsilon_\theta\left(z_t^j, t, c_\phi, (\pi_{i\to j}), I_i\right)\right\|_2^2 \quad (2)$$

where z_t^j is the noisy latent encoding of image I_j.

3D prior preservation loss. The test-time adaptation of the diffusion model on a few pairs of images can lead to catastrophic forgetting of the learned 3D priors. To address this problem, we incorporate a *3D prior preservation loss* as a regularization term, following DreamBooth [40]. We sample pairs of object images and relative camera poses from large-scale object dataset, and include these samples during the test-time optimization. Specifically, we sample images that are similar to the test object from F's training data. We use CLIP [33] as a similarity metric and retrieve the nearest neighbors of the average image representation of the images of the object. The loss in test time is:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{denoise}} + \mathcal{L}_{\text{prior}} \quad (3)$$

where $\mathcal{L}_{\text{denoise}}$ is the diffusion loss, $\mathcal{L}_{\text{prior}}$ is the prior preservation loss and $\alpha \in \mathbb{R}$ is the loss coefficient.

3.3. Novel View Synthesis

The instance-specific diffusion model obtained via test-time optimization allows us to sample novel views of the given object directly. We use stochastic conditioning [52] to incorporate *all of the input images* and the refined camera poses during the sampling process from the diffusion model. The original sampling process uses the same input image and relative camera pose in all of the denoising steps to generate a novel view. Differently, we make use of all the images during the sampling process via stochastic conditioning: We first compute the relative camera poses from each of the input images I_i given the refined camera poses π_i. Then, at each denoising step, we sample a different input image and corresponding camera pose and conditioning ϵ_θ on it.

3.4. 3D Reconstruction

Given the input images $\{I_i\}$ of the object, refined estimated camera poses $\{\pi_i\}$, and the instance-specific diffusion model F that generates novel views of the object, we reconstruct the object in 3D as a neural radiance field [26] with parameters ψ. We adapt an existing single-image 3D reconstruction pipeline [23] that uses view-conditioned diffusion models, to multiple images.

Losses. The loss comprises a data term on the reference images, a 2D diffusion prior term for novel views, and 3D shape regularization terms. The data term is a photometric loss between the input images I_i and the rendered image from the corresponding viewpoint $\mathcal{R}_\psi(\pi_i)$:

$$\mathcal{L}_{\text{data}} = \mathbb{E}_i \left\|\mathcal{R}_\psi^{\text{RGB}}(\pi_i) - I_i^{\text{RGB}}\right\|^2 + \mathbb{E}_i \left\|\mathcal{R}_\psi^{\text{Mask}}(\pi_i) - I_i^{\text{Mask}}\right\|^2$$

The 2D diffusion prior is a score distillation sampling loss [32] adapted to the view-conditioned diffusion model. Intuitively, it guides the renderings from novel viewpoints π to be similar to what the diffusion model would predict when conditioned on a randomly selected input view I_i:

$$\nabla_\psi \mathcal{L}_{\text{SDS}} = \mathbb{E}_{\pi,i,t,\epsilon}\left[\epsilon_\theta\left(z_t, t, c(\pi/\pi_i), I_i\right) - \epsilon\right] \frac{\partial \mathcal{R}_\psi^{\text{RGB}}(\pi)}{\partial \psi}$$

Here, z_t is the noisy latent encoded from the rendered image $\mathcal{R}_\psi(\pi)$ and π/π_i is the relative camera pose between π and π_i. Lastly, we regularize the 3D NeRF being optimized by

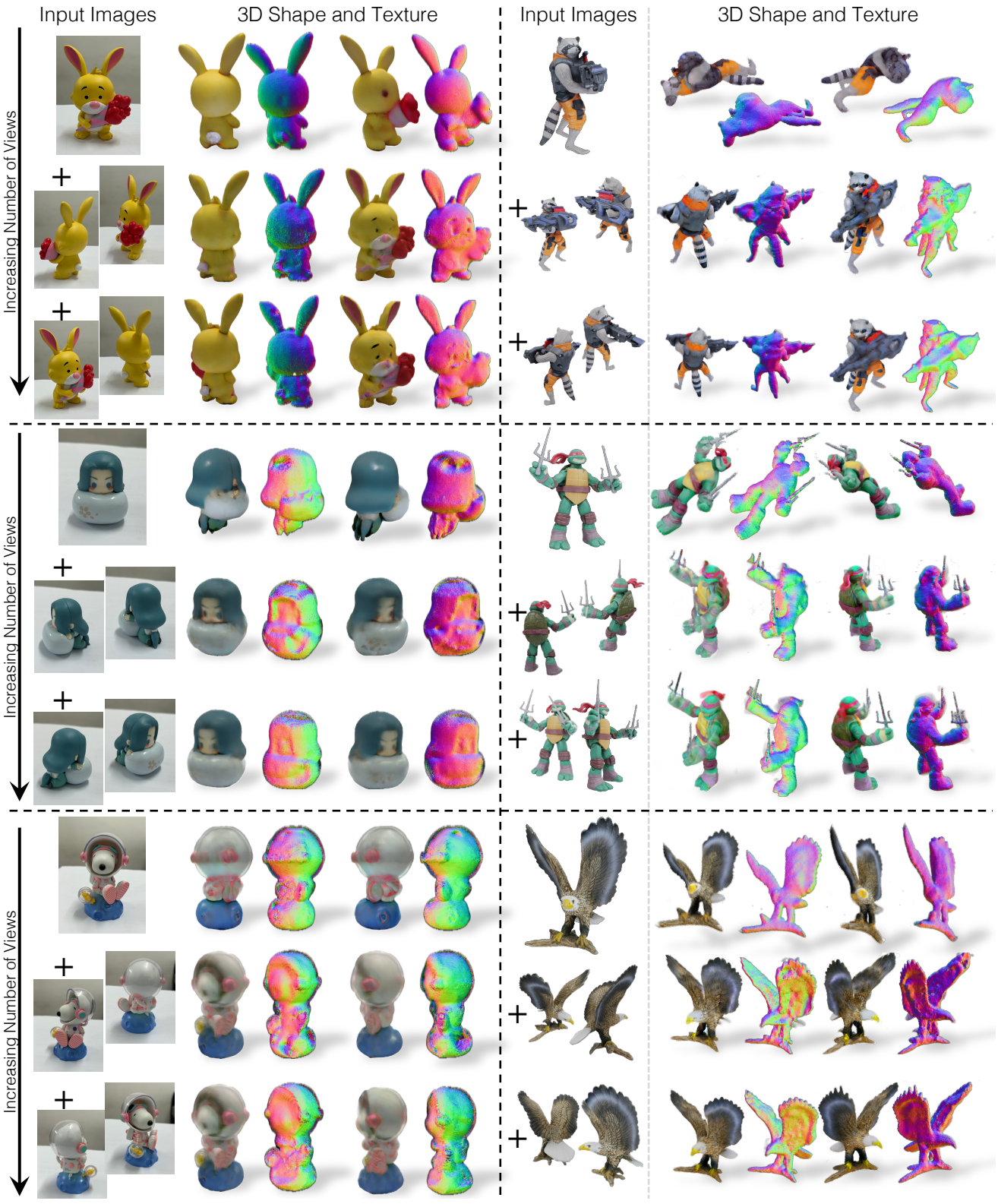

Figure 3. **3D reconstructions with one or more images.** Qualitative visualizations with 1, 3, and 5 views for SAP3D on real images (left column) and instances from the synthetic GSO dataset (right column). Observe how the wings of the eagle, the spiky weapon of the green turtle, and the yellow bunny's bouquet of flowers, all become more detailed and accurate with more views.

#Images	LPIPS ↓	PSNR ↑	SSIM ↑	CD ↓	VolumeIoU ↑
1 (Zero1-to-3)	0.23	14.1	0.82	0.168	0.25
2	0.16	18.0	0.83	0.041	0.51
3	0.14	18.5	0.83	0.024	0.57
4	0.11	19.6	0.85	0.023	0.67
5	0.11	19.8	0.86	0.019	0.68

Table 2. **3D reconstruction benchmark.** We compare geometry and appearance accuracy. As more images are provided, the 3D reconstruction quality improves.

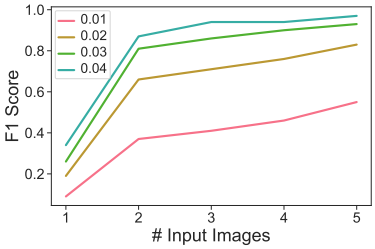

Figure 4. **F1-Score for 3D reconstruction per input set size.**

#Images	LPIPS ↓	PSNR ↑	SSIM ↑
1 (Zero-1-to-3)	0.23	15.2	0.79
2	0.15	17.6	0.83
3	0.13	18.2	0.83
4	0.10	19.4	0.85
5	0.10	19.5	0.85

Table 3. **2D benchmark**. We evaluate geometry-free novel-view-synthesis on 20 objects from the GSO dataset. As the number of input images increases the results improve.

guiding the normals to be smooth and the rendered masks to be sparse and low-entropy, using a regularization loss $\mathcal{L}_{\text{reg}}$ (see the Appendix for more details).

Our final loss is a weighted sum of these losses:

$$\mathcal{L} = \mathcal{L}_{\text{SDS}} + \lambda_{\text{data}}\mathcal{L}_{\text{data}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{4}$$

4. Experiments

In this section, we experimentally verify our system SAP3D. Section 4.1 provides more details on the implementation, Section 4.2 evaluates how SAP3D effectively ingests a variable number of images , and Section 4.3 verifies our system design choices and their impact on the quality of 3D reconstruction and 2D novel view synthesis.

4.1. Implementation details

Initial camera poses. We initialize our camera poses using RelPose++ [20]. We find that the original model, trained on the Co3D [36] (containing ∼19,000 objects), does not generalize well to in-the-wild objects with few views. Therefore, we re-train the model on a larger dataset - Objaverse [3] (∼800K objects) to obtain more accurate camera poses. In our ablations, we refer to this model as RelPose++*.

View-condition 2D diffusion model. We use Zero-1-to-3 [23] as the view-conditioned diffusion model. This model is pre-trained on Objaverse [3]. The input camera pose to this model is parameterized by azimuth angle, elevation angle, and scale. We therefore convert the initial camera pose estimations to these parameters, and estimate the absolute elevation of the first camera following [22].

Test-time optimization. We apply test-time optimization for 1000 steps, with a batch size of 1 and a learning rate of 0.1. For the prior preservation loss, we retrieve 50 nearest neighbors by computing the CLIP-ViT-L similarity to each example in Objaverse training data, and use $\alpha = 1.5$. We use a learning rate of 100 for optimizing the camera poses, and further restrict the range of the radius to make optimization more stable. Please see the Appendix for more details.

3D Reconstruction. We use InstantNGP [28] to represent our NeRF and train it for 3000 steps with the losses described in Section 3.4, $\lambda_{\text{data}} = 500$ and and learning rate of 0.01. Every minibatch renders one image for the data loss and one for the SDS loss. For the novel views in the SDS loss, we randomly sample azimuth from $\mathcal{U}(0, 360)$ and elevation from $\mathcal{U}(-60, 60)$.

4.2. More Sight, More Insight

SAP3D adapts diffusion-based SoTA single-image reconstruction to incorporate information from multiple views. Here, we demonstrate its effectiveness as the number of views increases, *i.e.* more sight gives more insight into the shape and appearance of the object. We evaluate the output 3D shape and appearance qualitatively and quantitatively.

Datasets. We benchmark quantitatively on Google's Scanned Objects (GSO) [4]. We randomly selected 20 objects and rendered k views of each object as inputs to our method, $k \in \{1, 2, 3, 4, 5\}$. We set the change in the azimuth and the elevation between the sampled random camera poses to be at least 30 degrees. Additionally, we provide qualitative results for real objects that we captured for $k \in \{1, 3, 5\}$.

Evaluation Metrics. To evaluate the appearance of the 3D reconstructions and the synthesized novel views, we use PSNR, SSIM, and LPIPS [58]. For evaluating the geometry of the 3D reconstructions we extract a mesh from the optimized NeRF and benchmark Chamfer Distance, F1 score at different thresholds, and VolumeIoU. For evaluating cameras, we compute the error in relative rotations (in degrees) between all pairs of input cameras, following [20].

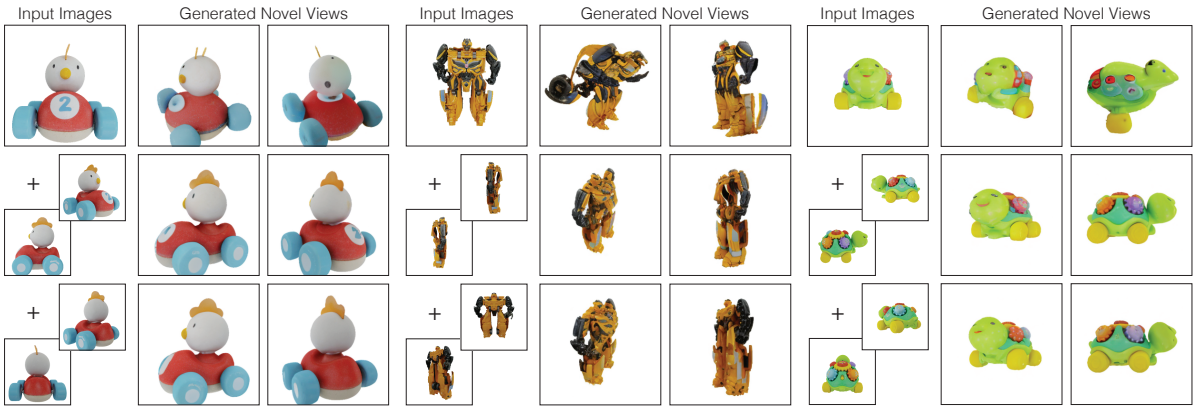

Figure 5. **SAP3D novel view qualitative results.** We present results for 1, 3, and 5 input images. With more input images, SAP3D improves fidelity of generated 3D details.

3D reconstruction improves with more images. We evaluate the 3D reconstruction quality of SAP3D. When only one view is provided, no test-time adaptation is performed as we do not have labels to fine-tune the diffusion model. As shown in Table 2 and Figure 4, both the geometry and the appearance improve with more input images. The largest improvement is between one view and two views, where test-time optimization is added to the process. The improvements diminish with more views and saturate with 4 input views. Figure 3 shows the same trend for 3D reconstruction from real images. When the number of input images is raised from 3 to 5, more fine-grained details are reconstructed correctly (the appearance of the bunny's eyes, the geometry of Snoopy's nose, and the eagle's wings).

2D NVS improves with more images. We evaluate our method for novel view synthesis. When $k = 1$, our method is similar to sampling from Zero-1-to-3 [23]. Table 3 presents the average performance on GSO. Similarly to the 3D reconstruction, the accuracy of the generated novel views improves with more input images. The relative improvement by adding additional images is larger than for 3D reconstruction, as the instance-specific diffusion model is used directly to sample novel views, instead of via distillation into InstantNGP (that incorporates additional priors in the form of regularization terms). Moreover, the improvements do not saturate with 5 images. Qualitative results are presented in Figure 5. As shown, with more images provided, the diffusion model hallucinates fewer incorrect details (e.g. the sign on the back of the toy car disappears), and generates fine details more accurately (e.g. the parts of the robot).

Camera poses improve with more images. We evaluate the accuracy of estimated relative camera poses. As shown in Figure 6, the relative camera poses from SAP3D improve significantly with more views.

4.3. System Verification

In this section, we ablate different components in our pipeline to verify their effects on the downstream 3D reconstruction (Table 4) and novel view synthesis performance (Table 5). Additionally, we ablate the effect of scaled RelPose++ and the fine-tuning stage on camera poses (Figure 6).

All of our ablations are done on 20 randomly chosen objects from GSO, and use 3 input images. We incorporate all the 3 images by using stochastic conditioning when sampling from the ablated models.

Initial camera poses ablations. We compare the downstream 3D reconstruction and novel view synthesis performance with different camera pose initializations. We compare the provided RelPose++ model [20] to our version, trained on Objaverse (RelPose++*), before introducing the adaptation stage to the model. As shown quantitatively in Table 4 (first two rows) and qualitatively in Figure 7, both the geometry and appearance of the downstream 3D reconstruction improve when RelPose++ is replaced with RelPose++*. Similar results hold for NVS downstream evaluation, as shown in Table 5. We conclude that large-scale pre-training of the camera pose estimator results in better initialization of the camera poses that improves downstream performance.

Test-time adaptation ablation. We compare our model with and without test-time training (appears in the tables as "SAP3D w/o adaptation"). Note that without test-time optimization, our system uses the original Zero-1-to-3 model as a prior. As shown in Table 5, test-time optimization improves the novel view synthesis in all metrics. As shown in Table 4, test-time optimization improves 3D reconstruction as well, although the improvement in appearance quality is smaller compared to NVS. Nevertheless, there is a significant improvement in geometry metrics. Finally, the test-time adaptation stage results in more accurate camera rotation predictions, as shown in Figure 6.

	LPIPS ↓	PSNR ↑	SSIM ↑	CD ↓	F1@0.04 ↑	VolumeIoU ↑
RelPose++	0.22	13.9	0.82	0.229	0.44	0.41
SAP3D w/o adapt.	0.17	17.1	0.85	0.029	0.87	0.60
SAP3D w/o L_{SDS}	0.51	9.8	0.57	0.281	0.33	0.22
SAP3D w/o L_{data}	0.20	15.3	0.84	0.030	0.76	0.52
SAP3D	**0.16**	**18.1**	**0.86**	**0.015**	**0.94**	**0.63**

Table 4. **SAP3D ablations for 3D reconstruction.** We ablate the SDS loss, photometric loss, and test-time adaptation. We evaluate on 13 objects as InstantNGP failed to converge when camera poses were initialized with RP++ for 7 objects.

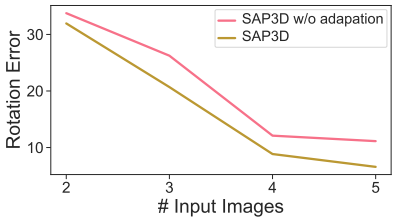

Figure 6. **Camera Pose Evaluation.** The rotation error (in degrees) with and without test-time adaptation for varying numbers of input images.

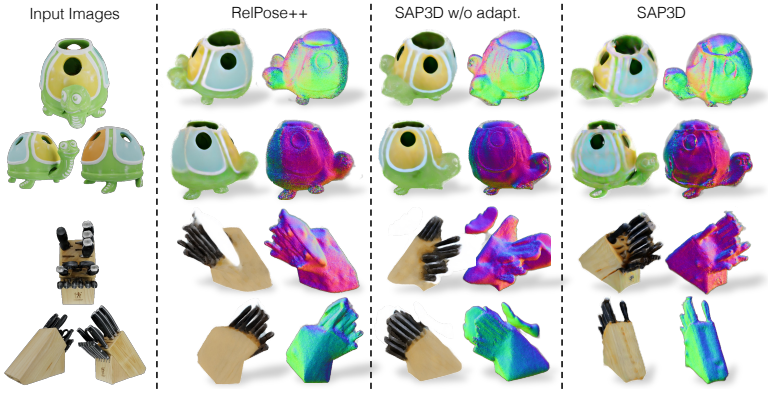

Figure 7. **Qualitative ablation results for 3D reconstruction.** Training RelPose++ on large-scale data and adapting the view-conditioned diffusion model at test-time improve results.

Figure 8. **Qualitative ablation results for NVS.**

3D reconstruction losses ablations. We ablate the contributions of the data and SDS losses to the 3D reconstruction quality. Without SDS, the 3D model optimization reduces to InstantNGP, and the 2D priors are not used for the reconstruction. As shown in Table 4, the reconstruction quality is the lowest in this case. The reconstruction quality after applying SDS and MSE is higher than applying each of the losses separately. This suggests that the reconstruction can benefit from applying 2D priors as well as a photometric loss that relies on the predicted relative camera poses of SAP3D.

5. Discussion and Limitations

We presented a system that enables 3D reconstruction and generation of novel views from an arbitrary number of images, and improves with more images. We discuss here our two main limitations and conclude with future work.

	LPIPS ↓	PSNR ↑	SSIM ↑
RelPose++	0.30	13.3	0.77
SAP3D w/o adaptation	0.18	16.3	0.78
SAP3D	**0.13**	**18.2**	**0.83**

Table 5. **SAP3D ablations for novel view synthesis**.

Camera pose parametrization. Our system inherits the camera pose parameterization of the pre-trained diffusion model, restricted to 3 degrees of freedom in the case of [23]. We believe that replacing the diffusion model with a generative model that is conditioned on a parametrization with more degrees of freedom will result in more control in novel view synthesis.

Optimization-based system. When the number of input images is larger than 1, our system requires an optimization stage of a large-scale diffusion model and therefore can not be applied in real-time. Specifically, the diffusion fine-tuning stage takes around 15 minutes per object on one A100 GPU.

Future work. While our current system contains a few independent components, we believe that an end-to-end approach can lead to better performance. Additionally, we hypothesize that model performance can improve at test-time with more inputs, without the need for fine-tuning, by directly training the model to use input examples in context, as done in large language models. We plan to explore this in future work.

References

[1] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[2] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21126–21136, 2022. 12

[3] 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. 2, 6

[4] 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, 2022. 2, 6

[5] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. *CVPR*, 2022. 2

[6] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2

[7] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2

[8] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019. 2

[9] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 88–104. Springer, 2020. 2

[10] Shubham Goel, Georgia Gkioxari, and Jitendra Malik. Differentiable stereopsis: Meshes from multiple views using differentiable rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8635–8644, 2022. 2

[11] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[12] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. 2

[13] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, 2021. 2

[14] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2

[15] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1966–1974, 2015. 2

[16] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 364–375, Red Hook, NY, USA, 2017. Curran Associates Inc. 2

[17] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples. *ACM Transactions on Graphics (TOG)*, 36:1 – 13, 2017. 12

[18] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023. 3

[19] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, 2020. 2

[20] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 2, 3, 6, 7

[21] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 12, 13

[22] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 6

[23] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 2, 3, 4, 6, 7, 8, 12, 13

[24] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[25] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[26] 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, 4

[27] 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 *The European Conference on Computer Vision (ECCV)*, 2020. 2

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 6

[29] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[30] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *arXiv preprint arXiv:2203.17272*, 2022. 3

[31] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59:207–232, 2004. 2

[32] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2, 4

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 4

[34] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 3

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[36] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2, 6

[37] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *CoRR*, abs/2106.05744, 2021. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

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

[40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 3, 4

[41] 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. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 2

[42] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 2

[43] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[44] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 2

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

[46] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. SparsePose: Sparse-view camera pose regression and refinement. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[47] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 2

[48] Nagabhushan Somraj and Rajiv Soundararajan. Vip-nerf: Visibility prior for sparse input neural radiance fields. In *ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2023. 2

[49] 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

[50] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2

[51] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2

[52] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 4

[53] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, pages 1–12, 2023. 2

[54] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2

[55] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[56] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Conference on Neural Information Processing Systems*, 2021. 2

[57] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018. 6

[59] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Joshua B Tenenbaum, William T Freeman, and Jiajun Wu. Learning to Reconstruct Shapes From Unseen Classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[60] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2