MVD-Fusion: Single-view 3D via Depth-consistent Multi-view Generation

Hanzhe Hu[1]*   Zhizhuo Zhou[2]*   Varun Jampani[3]   Shubham Tulsiani[1]

[1] Carnegie Mellon University   [2] Stanford University   [3] Stability AI

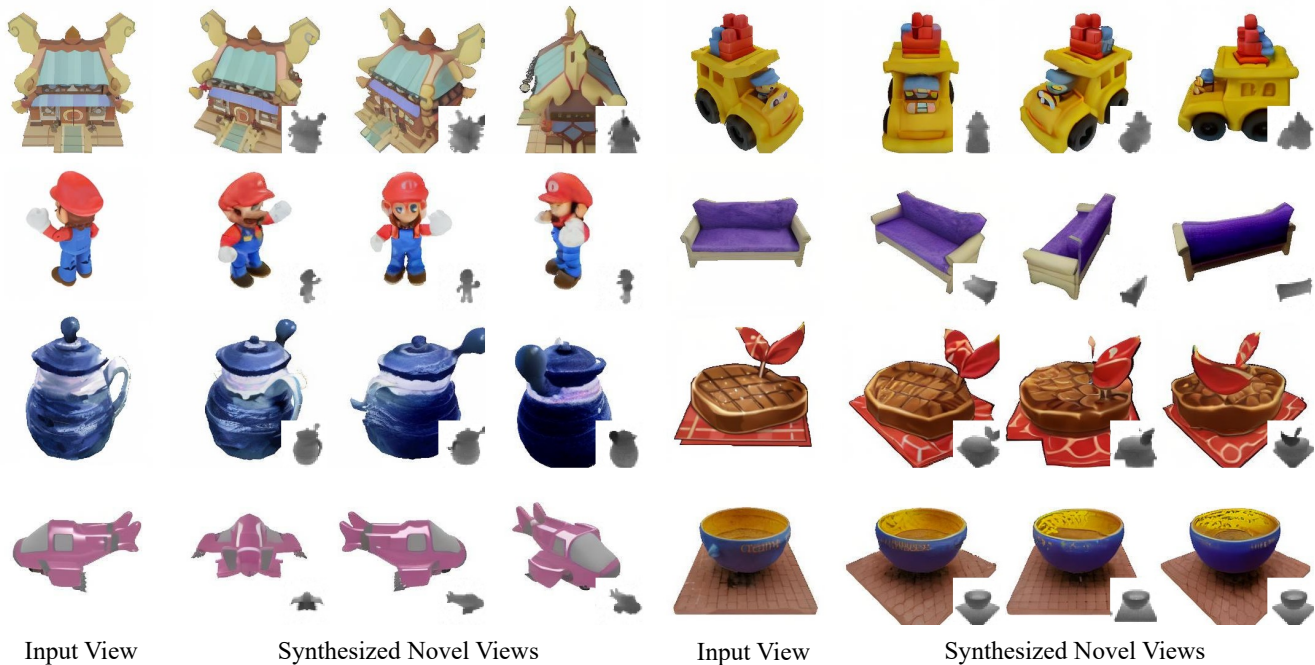https://mvd-fusion.github.io/

Figure 1. **Single-view 3D Inference**. Given an input RGB image, MVD-Fusion allows synthesizing multi-view RGB-D images using a depth-guided attention mechanism for enforcing multi-view consistency. We visualize the input RGB image (left) and three synthesized novel views (with generated depth in inset).

Abstract

We present MVD-Fusion: a method for single-view 3D inference via generative modeling of multi-view-consistent RGB-D images. While recent methods pursuing 3D inference advocate learning novel-view generative models, these generations are not 3D-consistent and require a distillation process to generate a 3D output. We instead cast the task of 3D inference as directly generating mutually-consistent multiple views and build on the insight that additionally inferring depth can provide a mechanism for enforcing this consistency. Specifically, we train a denoising diffusion model to generate multi-view RGB-D images given a sin-gle RGB input image and leverage the (intermediate noisy) depth estimates to obtain reprojection-based conditioning to maintain multi-view consistency. We train our model using large-scale synthetic dataset Obajverse as well as the real-world CO3D dataset comprising of generic camera viewpoints. We demonstrate that our approach can yield more accurate synthesis compared to recent state-of-the-art, including distillation-based 3D inference and prior multi-view generation methods. We also evaluate the geometry induced by our multi-view depth prediction and find that it yields a more accurate representation than other direct 3D inference approaches.

---

*Equal contribution.

1. Introduction

The task of recovering 3D from a single 2D image has witnessed a recent wave of generative-modeling based approaches. In particular, while initial 3D prediction methods pursued inference of volumetric [1, 6], point clouds [5, 45], or meshes [7, 12] representations, a class of recent approaches [19, 22, 52] instead formulate the task as learning (conditional) generation of novel views. By adapting large-scale pre-trained generative models, these methods can learn generalizable view synthesis that performs remarkably well even for generic objects in-the-wild. However, the synthesized novel views are not mutually consistent and these 2D generative methods rely on a (costly) 'score distillation' [27] based optimization to then recover a consistent 3D. While this process can yield impressive results, these come at the cost of a reduction in both the efficiency of inference and the diversity of the generations.

In this work, we seek to overcome these limitations, and pursue an approach that allows directly generating diverse outputs. We do so by re-formulating the task of 3D inference as that of generating a set of (mutually consistent) multiple views, and learn a (conditional) generative prior to model this joint distribution. While some recent (and concurrent) methods do similarly 'co-generate' multiple views given a single input image [33, 34], these are typically not geometrically consistent. Instead, our approach is inspired by the recent work from Liu *et al*. [20] which incorporates a 3D bottleneck with unprojection and reprojection as an inductive bias for ensuring geometric consistency across views. In this work, we explore an alternate mechanism for enforcing such consistency. In particular, we formulate a depth-generation-guided approach that: a) allows improved generation via depth-based reprojection, and b) enables directly producing an estimate of the 3D geometry via the inferred multi-view 2.5D representation.

Our approach for enforcing multi-view consistency stems from a simple question: *what does it mean for images to be 3D-consistent?* Drawing inspiration from classical 3D reconstruction methods, one answer to this question is that if a pixel in one image corresponds to a point that is also visible in another, then the local appearance should match. However, how can we know where the 3D point corresponding to a pixel in one image may project in the other? This inspires our solution for generating multi-view consistent images, where we not only generate the RGB images but also reason about the corresponding depth for these generations (and thus allow such reprojection-based multi-view consistency). More specifically, we adopt existing 2D diffusion models to generate RGB-D images while adding multi-view projection based on (noisy) depth estimates to enforce 3D consistency.

We train our system using a large-scale synthetic dataset, and empirically demonstrate the efficacy of our approach on held out objects as well as scanned real world data. We show that our approach allows more accurate generation compared to prior state of the art while also (directly) generating plausible geometry via the synthesized depth images. Finally, we also highlight the ability of our method to sample diverse outputs and its ability to generalize zero-shot to in-the-wild generic objects.

2. Related Work

Single-view 3D Prediction. The task of inferring 3D from 2D images is a long-standing one in computer vision, and the pre-dominant learning-based approach has been to frame it as a *prediction* task where a data-driven predictor is trained to output a 3D representation given image input. In particular, several deep learning based methods pursued this task by inferring a plethora of 3D representations such as volumetric 3D [1, 6, 38, 48], meshes [7, 12, 14, 15, 41, 47], point clouds [5, 25, 45], or neural implicit fields [16, 23, 39]. While these approaches have shown promising results, they often struggle to generate detailed outputs for complex objects owing to the ambiguity in unobserved regions. Indeed, any such regression-based approach fundamentally cannot model the inherent uncertainty in single-view reconstruction. In contrast, our generative modeling-based approach can synthesize and generate high-fidelity outputs and also yield multiple modes.

3D Inference via 2D Diffusion. Instead of directly predicting 3D shapes in a feed-forward way, this line of work utilizes 2D diffusion prior to facilitating 3D inference. In particular, DreameFusion [27] and SJC [40] formulated a 'score distillation' objective that enabled incorporating pre-trained diffusion as a prior for optimization, and leveraged it to distill a 2D stable diffusion model for the text-to-3D inference. Inspired by this, several works [3, 8, 22, 28, 31, 35, 46] adopt this pipeline to optimize a neural radiance field (NeRF [24]) for the single-view reconstruction task. For instance, RealFusion [22] extracts from a single image of an object a 360° 3D reconstruction by leveraging a 2D diffusion model with a single-image variant of textual inversion whereas NeuralLift-360 [46] utilizes a depth-aware NeRF and learns to craft the scene by denoising diffusion models. However, as these methods only use pre-trained image diffusion models for 3D inference, they can suffer from implausible 3D outputs *e.g*. *janus effect* and do not always preserve the details in the observed image.

To circumvent this, SparseFusion [52] proposed to learn a novel-view diffusion model using epipolar feature transformer to build a view-conditioned features model for sparse-view reconstruction and distilled it to obtain more accurate 3D reconstructions. Moreover, Diffusion with Forward Models [37] applied 2D diffusion networks to denoise a consistent 3D scene. Our approach builds on Zero-1-to-
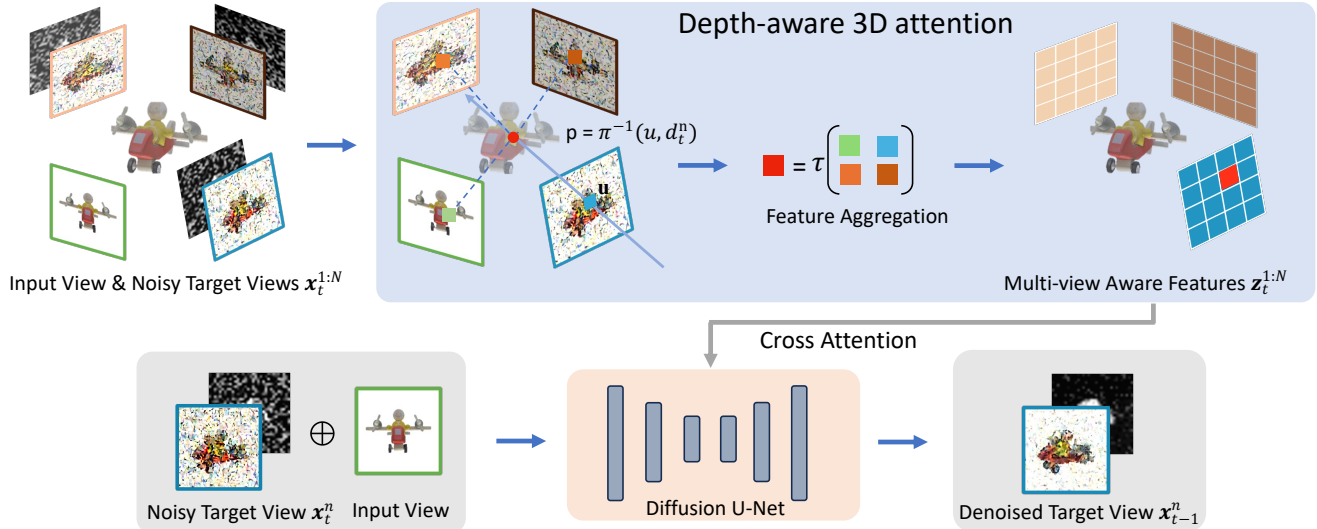
Figure 2. **Approach Overview.** MVD-Fusion learns a denoising diffusion model for generating multi-view RGB-D images given an input RGB image. At each diffusion timestep t, MVD-Fusion uses the current (noisy) depth estimates to compute depth-projection-based multi-view aware features (top). A novel-view diffusion based U-Net is modified to leverage these multi-view aware features as additional conditioning while producing denoised estimates of both, RGB and depth (bottom).

3 [19], which demonstrated that a large pre-trained image diffusion model can be finetuned for novel-view generation using a large-scale 3D dataset to achieve better generalization ability. While these methods are able to produce high-quality predictions, the reliance on score distillation sampling restricts them from obtaining diverse results with single-view 3D prediction as an under-constrained task.

Multi-view Image Generation. Unlike novel-view generation models which model the distribution over a single view given a reference image, many recent works have investigated generating multi-view images at the same time by using diffusion models, including text-based methods [33, 36] and image-based methods [20, 34]. Given a text conditioning, MVDiffusion [36] simultaneously generates all images with a global transformer to facilitate cross-view interactions. Similarly, MVDream [33] produces multi-view images via multi-view diffusion and leverages a self-attention layer to learn cross-view dependency and encourage multi-view consistency. While these methods rely on text input, Viewset Diffusion [34] adopts a similar approach for generating a multi-view image set given an input image and subsequently infers a radiance field to ensure consistent geometry. While these methods, similar to our goal, can model the distribution over novel views, they do leverage any geometric mechanism to enforce multi-view consistency. Perhaps most closely related to our work, Sync-Dreamer [20] proposes to use a 3D-aware feature attention mechanism that correlates the corresponding features across views to enforce multi-view consistency. Different from [20], we utilize depth information to learn consistency across views instead of a 3D bottleneck that contains redun-

dant information. Finally, there have been several promising concurrent works which also pursue multi-view inference [10, 13, 18, 21, 32, 44] but we believe that our method of depth-guided multi-view diffusion represents a complementary advance to the techniques proposed in these.

3. Method

Given a single RGB image, our method generates a set of multi-view consistent RGB-D predictions. In addition to allowing the synthesis of the object from any desired set of views, the generated multi-view depth maps also conveniently yield a (coarse) point cloud representation of the geometry. To ensure multi-view consistency among the generated images, we model the joint distribution over a set of posed images by adding depth-guided 3D cross-attention layers on top of pre-trained latent diffusion backbone from Stable Diffusion [30] and Zero-1-to-3 [19]. We first formalize this task of multi-view generation via denoising diffusion (Section 3.1), and then detail our specific approach to enforcing multi-view consistency (Section 3.2) and 2.5D image generation (Section 3.3). A diagram of our method is in Figure 2.

3.1. Multi-view Denoising Diffusion

A conditional denoising diffusion model can capture the distribution over a variable of interest $\mathbf{x}$ given some conditioning $\mathbf{c}$. In particular, by learning a function $\epsilon(\mathbf{x}_t, \mathbf{c}, t)$ that learns to denoise an input with time-dependent corruption, diffusion models can allow sampling from the distribution $p(\mathbf{x}|\mathbf{c})$. Towards our goal of multi-view generation, we are interested in an instantiation of this framework where
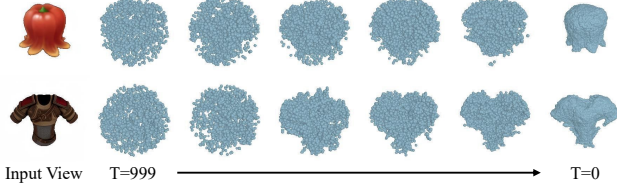
Figure 3. We visualize the unprojected point cloud obtained from a set of noisy RGB-D images at different timesteps during inference. We observe the gradual denoising of geometry from a random point cloud to a point cloud that matches the input object.

the conditioning corresponds to an observed RGB image $\mathbf{y}$ and a set of desired novel viewpoints $\{\pi^n\}$. Given these as input, we aim to generate a (mutually consistent) set of novel views $\{\mathbf{x}^n\}$ corresponding to the conditioning viewpoints and thus seek to learn a denoising diffusion model that captures $p(\{\mathbf{x}^n\}|\mathbf{y}, \{\pi^n\})$.

To learn such a diffusion model, we need to formulate an approach that can predict the noise added to a set of corrupted multi-view images:

$$\{\mathbf{x}_t^n\} = \{\sqrt{\bar{\alpha}_t}\mathbf{x}^n + \sqrt{1-\bar{\alpha}_t}\epsilon_n\}$$
$$\epsilon_{pred} = f(\{\mathbf{x}_t^n\}, \mathbf{y}, t)$$

Instead of learning such a prediction model from scratch, we propose to adapt a pre-trained novel-view generative model from Zero-1-to-3 [19]. Specifically, this model captures the distribution over a single novel view $p(\mathbf{x}|\mathbf{y}, \pi)$ given an RGB input by learning a denoising function $\epsilon_\phi(\mathbf{y}, \mathbf{x}_t, \pi, t)$. While we aim to leverage this pre-trained large-scale module for efficient learning and generalization, it only models the distribution over a single novel view whereas we aim to model the joint distribution over multiple views. To enable efficiently adapting this, we propose a first learn a separate module that computes view-aligned multi-view aware features $\{\mathbf{z}_t^n\}$. We then modify the pre-trained single-view diffusion model to additionally leverage this multi-view aware conditioning:

$$\mathbf{z}_t^n = f_\theta(\mathbf{y}, \{\mathbf{x}_t^n\}, \{\pi^n\}, t) \tag{1}$$
$$\epsilon_{pred} = \{\epsilon_{\phi'}(\mathbf{y}, \mathbf{x}_t^n, \pi^n, \mathbf{z}_t^n, t)\} \tag{2}$$

3.2. Depth-guided Multi-view Consistency

Generating a set of consistent images requires the network to attend across different images within the set. SyncDreamer [20] proposes a way of achieving multi-view attention by unprojecting features from the set of images $\{\mathbf{x}_t^0, \mathbf{x}_t^1, ...\mathbf{x}_t^N\}$ onto a 3D volume $\mathcal{V}$ and interpolating conditioning feature frustums $\{\mathbf{z}_t^0, \mathbf{z}_t^1, ...\mathbf{z}_t^N\}$. However, interpolating feature frustums linearly spaced across the whole 3D volume is an expensive operation that assumes no prior

knowledge of object surfaces. In contrast, we propose to explicitly reason about the surface by additionally generating depth and biasing sampling near the possible surface.

Given a target view $\mathbf{x}_t^i$, we obtain feature frustum $\mathbf{z}_t^i$ by shooting rays and sampling features at 3D locations along the rays. For each ray, we sample D depth values near the expected surface and aggregate projected features from target views $\{\mathbf{x}_t^n\}$ and input view $\mathbf{y}$. Let $(\mathbf{z}_t^i)_{md}$ be the feature for the m-th ray at d-th depth in $\mathbf{z}_t^i$. For a 3D point $\mathbf{p}_{md}^i$ corresponding to the feature $(\mathbf{z}_t^i)_{md}$, we sample $N+1$ features $\mathbf{c}_{mdn}$ from $\{\mathbf{x}_t^n\}$ and $\mathbf{y}$. We also include the plucker embedding of query ray $\mathbf{q}_m$ and reference rays $\mathbf{r}_{mdn}$ from $\mathbf{p}_{md}^i$ to $N+1$ camera centers along with the sampled features as input into the transformer f_θ that predicts view-aligned multi-view aware features:

$$(\mathbf{z}_t^i)_{md} = \sum_{n=0}^{N+1} w_\theta(\mathbf{v}_{mdn}, t) f_\theta(\mathbf{v}_{mdn}, t) \tag{3}$$
$$\text{where } \mathbf{v}_{mdn} = \{\mathbf{c}_{mdn}, \mathbf{r}_{mdn}, \mathbf{q}_m\}$$

Here, $w_\theta(\mathbf{v}_{mdn}, t)$ represents (normalized) weights predicted by the transformer, which are then used to aggregate the multi-view features to obtain the pointwise feature $(\mathbf{z}_t^i)_{md}$.

Naively, we can sample a large number of depth points along each ray linearly spaced throughout the scene bound; however, such exhaustive sampling quickly becomes a memory constraint while possibly making the learning task more difficult as the network may also observe features away from the surface. Thus, we sample $D = 3$ depths from a Gaussian distribution centered around an unbiased estimate of depth given the noisy depth d_t and a scaled version of the denoising diffusion model variance schedule in equation Eq. 4.

$$d \sim \mathcal{N}(\mathbb{E}[d_0], k\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}})$$
$$\text{where } d_t = \sqrt{\bar{\alpha}_t}d_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{4}$$
$$\text{and } \mathbb{E}[d_0] = \frac{d_t}{\sqrt{\bar{\alpha}_t}}$$

We then use these multi-view aware features $\{\mathbf{z}_t^n\}$ as conditioning input into our latent diffusion model in Section 3.3.

3.3. Learning Multi-view 2.5D Diffusion

Inspired by the success of finetuning pretrained Stable Diffusion models, we adapt Zero-1-to-3 [19] as our multi-view novel view synthesis backbone ϵ_ϕ. While Zero-1-to-3 is designed to only model single-view distributions and generate RGB output, we adapt it to predict an additional depth channel and cross-attend to the multi-view aware features.

First, we increase the input and output channels of the latent diffusion UNet backbone to predict normalized depth.
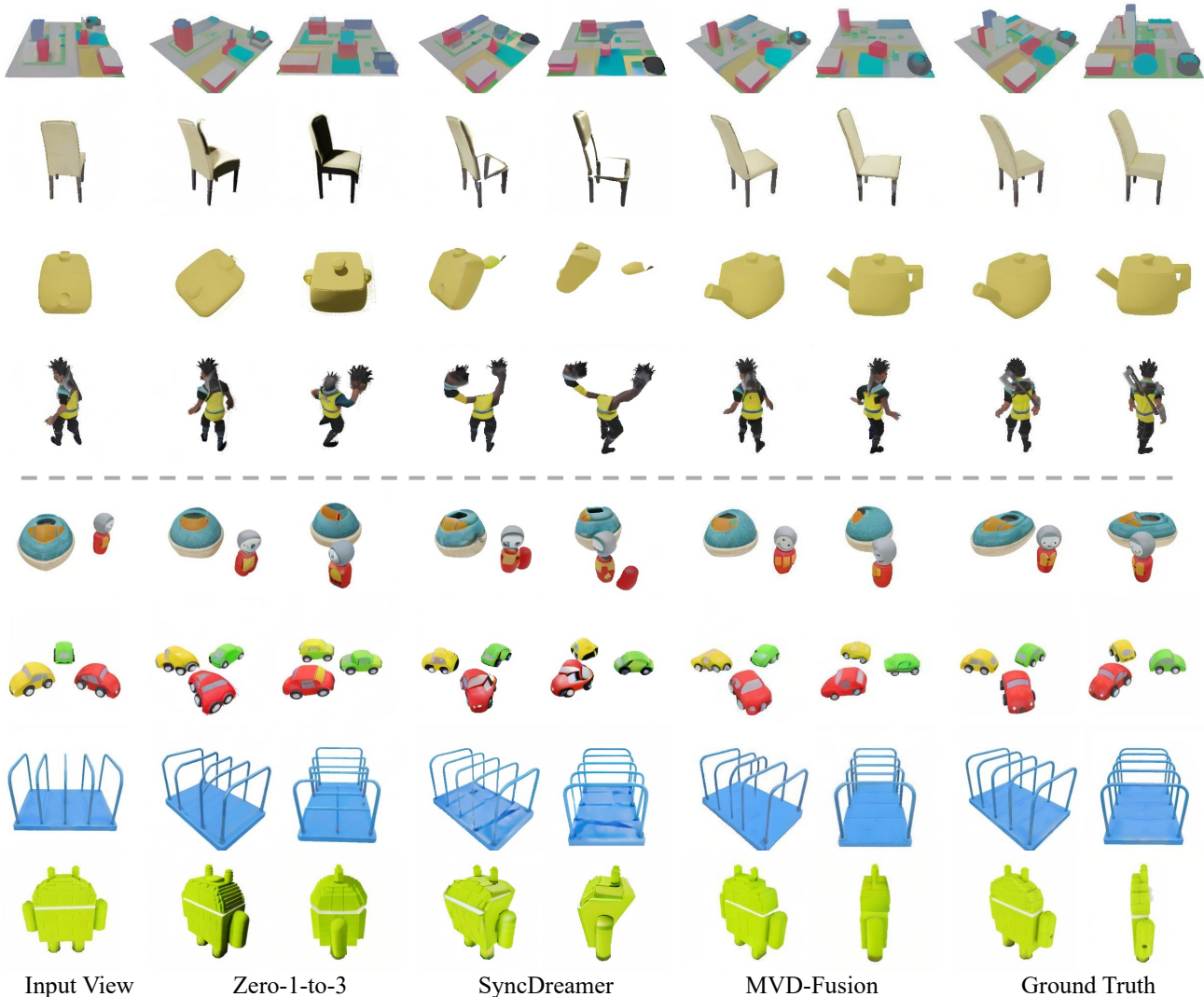
| Input View | Zero-1-to-3 | SyncDreamer | MVD-Fusion | Ground Truth |

Figure 4. **Qualitative results for novel view synthesis on instances from Objaverse (top) and Google Scanned objects (bottom)**. We compare our method with Zero-1-to-3 [19] and SyncDreamer [20]. We show the input image and two novel views generated by each method. Zero-1-to-3 independently generates novel views which are not consistent (*e.g.* the person in Objaverse). While both, SyncDreamer and MVD-Fusion yield consistent generations, we find that MVD-Fusion can generate more plausible output (*e.g.* the Android image) and is more faithful to details in the input (*e.g.* the three cars).

While the image latents can be decoded into high-resolution images, our predicted depth map remains at the lower resolution. This multi-resolution approach to predicting RGB-D lets us use the frozen Stable Diffusion VAE to decode high-resolution RGB images. Moreover, we add additional residual cross-attention layers at multiple levels of the UNet to attend to our multi-view aware features. Finally, we modify the camera parameterization used in Zero-1-to-3 from a 3-DoF azimuth, elevation, and radius parameterization to use the full perspective camera matrix. This makes our method capable of handling arbitrary camera poses in real datasets such as CO3D.

During training, we finetune all the parameters of our network and follow [9] to use a simplified variational lower bound objective in Eq. 5. During inference, we follow [20] and use a classifier free guidance of 2.0.

$$\mathcal{L}_{DM} = \mathbb{E}_{\mathbf{x}_0^n, \epsilon^n, t} \left[|| \epsilon^n - \{ \epsilon_{\phi'}(\mathbf{y}, \mathbf{x}_t^n, \pi^n, \mathbf{z}_t^n, t) \} || \right]$$
$$\text{where } \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon; \quad \epsilon \sim \mathcal{N}(0, 1) \quad (5)$$

4. Experiments

We train MVD-Fusion using a large-scale synthetic dataset (Section 4.1), and evaluate it on both, synthetic and real-
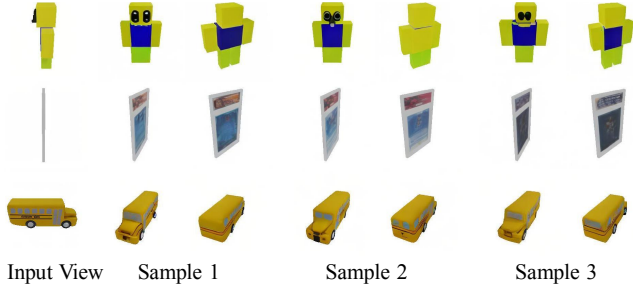
| Input View | Sample 1 | Sample 2 | Sample 3 |

Figure 5. **Sample Diversity.** MVD-Fusion is capable of generating diverse samples given the same input. We show the input image (left) followed by views synthesized in three randomly generated samples. We observe that there is meaningful variation in uncertain regions *e.g.* the eyes of the character and the colors on the screen vary across samples.

world objects for view synthesis (Section 4.2) and 3D reconstruction (Section 4.3). We show that our results achieves more accurate view synthesis compared to state-of-the-art as well as yields better 3D predictions compared to prior direct 3D inference methods. Finally, we present qualitative results on in-the-wild-objects (Section 4.4).

4.1. Experimental Setup

Datasets. We use the large-scale 3D dataset Objaverse [2] for training. Since Objaverse is large (800K instances) and contains several instances with poor texture, we filter the dataset with CLIP score to remove instances that match a set of hand-crafted negative text prompts. Our filtered set contains around 400K instances. For each instance, we render 16 views from an elevation of 30 degrees and azimuth linearly spaced across 360 degrees. Additionally, we hold out a subset of Objaverse instances following [19] for evaluation, which consists of about 4k instances.

Beyond evaluating on these held-out Objaverse instances, we also evaluate our method with the Google Scanned Object dataset (GSO) [4], which consists of high-quality scanned household items. For each object, we render 16 views evenly spaced in azimuth from an elevation of 30 degrees and choose one of them as the input image. For quantitative results, we randomly chose 30 objects to compute the metrics. Finally, to show the flexibility of our approach in modeling real-world datasets with general perspective cameras, as opposed to the common 3DoF cameras used in Objaverse and GSO, we finetune and evaluate our model on CO3D [29]. We follow [50] to train on 41 categories and evaluate on the held-out set of all 51 categories.

Baselines. For the novel view synthesis task, we adopt Zero-1-to-3 [19] and SyncDreamer [20] as our baseline methods. Given an input image, Zero-1-to-3 can synthesize

Table 1. **Results for novel view synthesis on the Objaverse dataset**. We compare our method with two other baselines on 100 instances from the test set of Objaverse dataset, same as in [19]. Our method outperforms existing baselines over three commonly used metrics: PSNR, SSIM and LPIPS.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Zero123 [19]	17.37	0.783	0.211
SyncDreamer [20]	19.22	0.817	0.176
MVD-Fusion	**21.19**	**0.835**	**0.146**

images from novel viewpoints. Built on Zero-1-to-3, SyncDreamer can simultaneously generate multiple images from different viewpoints with 3D consistency. For CO3D, we compare against PixelNeRF [49] as both Zero-1-to-3 and SyncDreamer are restricted to 3-DoF camera variation.

For 3D reconstruction, we compare our method with the aforementioned two methods together with RealFusion [22], Magic 123 [28], One-2-3-45 [17], Point-E [26] and Shape-E [11]. Note that the diffusion-based methods require neural field optimization using either rendering objectives or distillation objectives (*e.g.* Zero-1-to-3 requires a SDS distillation for extracting geometry whereas SyncDreamer relies on training Neus [42]), whereas our method allows 'directly' computing the geometry via un-projecting the predicted depth maps. To highlight this distinction, we categorize the reconstruction approaches as direct (One-2-3-45 [17], Point-E [26], Shape-E [11], and MVD-Fusion) or optimization based (RealFusion [22], Magic 123 [28], Zero-1-to-3 [19], and SyncDreamer [20]).

Metrics. For the novel view synthesis, we adopt commonly used metrics: PSNR, SSIM [43], and LPIPS [51]. For the 3D reconstruction task, we report Chamfer Distances between ground-truth and predicted point clouds.

Implementation Details. We train our model on a filtered version of the Objaverse dataset, which consists of about 400k instances. During training, for each instance, we randomly sample 5 views and choose the first one as the input view. We train the model for 400k iterations with 4 40G A100 GPUs using a total batch size of 16. We use Adam optimizer with a learning rate of 1e-5. Even though we only train with 5 views, we can sample an arbitrary set during inference as our depth-based projection followed by transformer-based aggregation trivially generalizes to incorporate more views. In our experiments, we render 16 views for each instance for evaluation.

4.2. Novel View Synthesis

Objaverse and Google Scanned Objects. We report quantitative results on the Objaverse dataset and GSO dataset in Table 1 and Table 2, respectively. For the Objaverse dataset, we use the held-out test set for evaluation

Table 2. **Results for novel view synthesis on the Google Scanned Objects (GSO) dataset.** We compare our method with two other baselines on 100 instances randomly chosen from the GSO dataset. Our method achieves consistent improvement over baseline methods on PSNR and LPIPS, while slightly worse than SyncDreamer on SSIM.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Zero123 [19]	17.42	0.756	0.207
SyncDreamer [20]	18.95	**0.796**	0.176
MVD-Fusion	**19.53**	0.790	**0.175**

Table 3. **Results for 51 category novel view synthesis on CO3D.** We significantly outperform PixelNeRF on perceptual quality.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
PixelNeRF [49]	**17.64**	0.484	0.378
MVD-Fusion	17.16	**0.701**	**0.220**

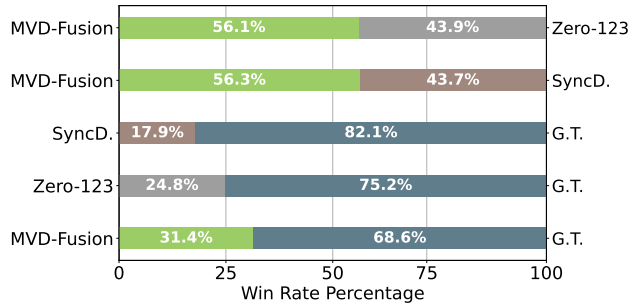

Figure 6. User preference percentage of MVD-Fusion against Zero-123, SyncDreamer (SyncD.), and ground truth (G.T.).

Input   PixelNeRF   MVD-Fusion

Figure 7. Qualitative results for novel view synthesis on instances from CO3D. MVD-Fusionis able to predict accurate and realistic novel views on real-world datasets with perspective camera poses.

baselines, we also highlight the ability of MVD-Fusion to generate multiple plausible outputs. In particular, since novel view synthesis from a single image is an under-constrained task, using the diffusion model can effectively generate more diverse samples given a single input image. As shown in Figure 5, MVD-Fusion is able to generate diverse plausible samples with different random seeds *e.g.* varying textures in the front of the bus.

User Study. We run a user study by randomly selecting 40 instances from Objaverse and GSO test set and asking 43 users to make 860 pairwise comparisons (users are shown an input image and two generated novel views per method). We show results in Figure 6. Our method tends to be chosen over Zero-123 and SyncDreamer, and is also more competitive with GT compared to these baselines.

Common Objects in 3D. Real-world data often have cameras that are not origin facing, making methods that model 3DoF origin facing cameras [19, 20] not suitable for real-world inference. We finetune our model on CO3D and show novel view synthesis results in Table 3. We also train a cross-category PixelNeRF [49] model as a baseline. While it is slightly better in PSNR (perhaps due to blurry mean predictions being optimal under uncertainty), our method vastly outperforms PixelNeRF in perceptual metrics SSIM and LPIPS (see Figure 7).

4.3. Single-view Reconstruction

Unlike previous methods such as Zero-1-to-3 and Sync-Dreamer, which have to fit a radiance field from generated multi-view images to obtain the 3D shape, MVD-Fusion can directly obtain the point cloud. With multi-view RGB-D generations, we can simply unproject the foreground pixels and obtain the object point cloud. We show quantitative results in Table 4, where we compare our method against previous methods on the GSO dataset using chamfer distance. We see that our method outperforms all of the methods that directly infer 3D shapes and most of the methods that require further optimization steps to get the 3D shapes.

whereas we use a subset of 30 random objects from GSO. We find that our method achieves consistent improvements over the baselines across metrics on both, the in-distribution Objaverse dataset and the out-of-distribution GSO dataset. We also provide qualitative comparisons on the Objaverse dataset and GSO dataset in Figure 4. Although Zero-1-to-3 [19] produces visually reasonable images, it suffers from multi-view inconsistency across generated viewpoints. In contrast SyncDreamer and MVD-Fusion, are able to obtain multi-view consistency among generated images. However, SyncDreamer struggles to obtain high alignment with the input image, sometimes leading to consistent multi-view images with unreasonable appearance. Our method, on the other hand, is able to generate multi-view consistent images with better alignment with the image input and more plausible completions in unobserved regions. Moreover, we also note that SyncDreamer is trained on the whole Obja-verse dataset and may have actually seen these instances whereas these represent held out instances for both, MVD-Fusion and Zero-1-to-3.
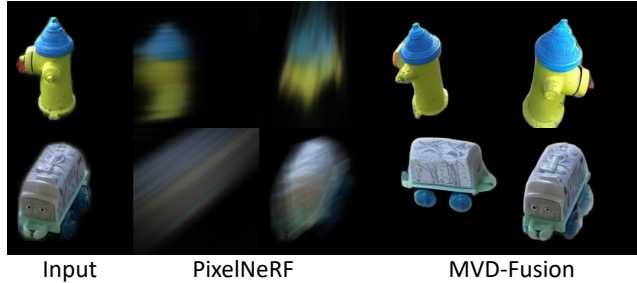
In addition to visualizing the comparative results with

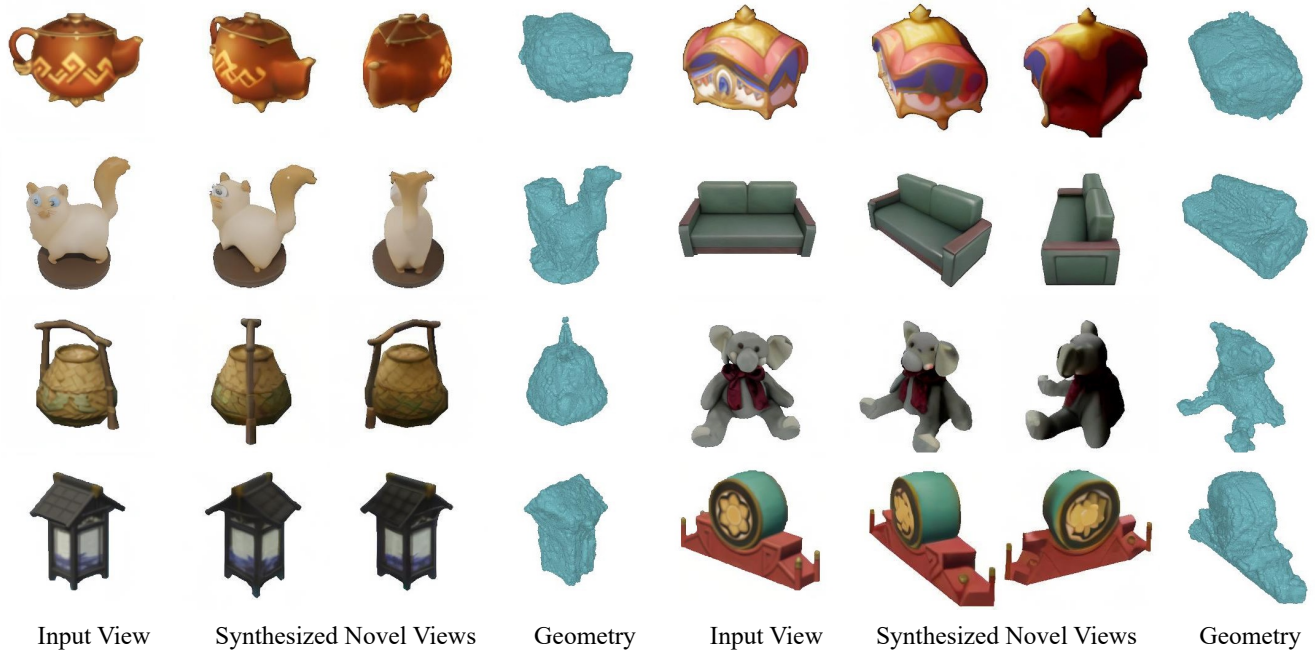| | Input View | Synthesized Novel Views | Geometry | Input View | Synthesized Novel Views | Geometry |

Figure 8. **In-the-wild Generalization.** We visualize the prediction from MVD-Fusion on in-the-wild internet images. We find that MVD-Fusion is able to preserve the rich texture in the input images and model the rough geometry without post-processing.

Table 4. **Results for 3D reconstruction on the Google Scanned Objects (GSO) dataset.** 'Optimization' denotes methods that require additional training such as fitting an occupancy field to obtain 3D shapes. 'Direct' denotes methods that can directly output 3D predictions. Following [20], we report Chamfer Distance on the same 30 instances from the GSO dataset. Our method demonstrates consistent improvement over 'direct' methods and outperforms most of the 'optimization' methods.

3D Extraction	Method	Chamfer Dist ↓
Optimization	RealFusion [22]	0.082
	Magic123 [28]	0.052
	Zero123 [19]	0.034
	SyncDreamer [20]	**0.026**
Direct	One-2-3-45 [17]	0.063
	Point-E [26]	0.043
	Shap-E [11]	0.044
	MVD-Fusion	0.031

4.4. In-the-wild Generalization

We also demonstrate the generalization ability of MVD-Fusion for reconstructing in-the-wild images from the internet. We show qualitative results depicting generated novel views and recovered point clouds in Figure 8. With challenging out-of-domain images as input, MVD-Fusion is still capable of generating consistent novel-view images and reasonable 3D shapes from single-view observation.

5. Discussion

In this work, we presented MVD-Fusion, which allowed co-generating multi-view images given a single input image. Our approach allowed adapting a pre-trained large-scale novel-view diffusion model for generating multi-view RGB-D images, and enforced consistency among these via depth-guided projection. While our results showed improvements over prior state-of-the-art across datasets, there are several challenges that still remain. First, the multi-view consistency is encouraged via inductive biases in the network design but is not guaranteed, and the network may generate (slightly) inconsistent multi-view predictions. Moreover, while our inferred multi-view depth maps can yield a point cloud representation that captures the coarse geometry, these coarse depth maps do not capture the fine details visible in the generated views and an optimization-based procedure may help extract these better. Finally, our approach has been trained on clean unoccluded instances and would not be directly applicable under cluttered scenes with partially visible objects, and it remains an open research question to build systems that can deal with such challenging scenarios.

Acknowledgements

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 *ECCV*, 2016. 2

[2] 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. In *CVPR*, 2023. 6

[3] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *CVPR*, 2023. 2

[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. In *ICRA*, 2022. 6

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

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

[7] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *ICCV*, 2019. 2

[8] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *ICML*, 2023. 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 5

[10] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, and Lu Sheng. Epidiff: Enhancing multiview synthesis via localized epipolar-constrained diffusion. In *CVPR*, 2024. 3

[11] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 6, 8

[12] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2

[13] Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. Spad : Spatially aware multiview diffusers. In *CVPR*, 2024. 3

[14] Nilesh Kulkarni, Justin Johnson, and David F Fouhey. Directed ray distance functions for 3d scene reconstruction. In *ECCV*, 2022. 2

[15] Nilesh Kulkarni, Linyi Jin, Justin Johnson, and David F Fouhey. Learning to predict scene-level implicit 3d from posed rgbd data. In *CVPR*, 2023. 2

[16] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. In *NeurIPS*, 2020. 2

[17] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023. 6, 8

[18] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *CVPR*, 2024. 3

[19] 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, 5, 6, 7, 8

[20] 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. In *ICLR*, 2024. 2, 3, 4, 5, 6, 7, 8

[21] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, 2024. 3

[22] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, 2023. 2, 6, 8

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

[24] 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

[25] K L Navaneet, Ansu Mathew, Shashank Kashyap, Wei-Chih Hung, Varun Jampani, and R Venkatesh Babu. From image collections to point clouds with self-supervised shape and pose networks. In *CVPR*, 2020. 2

[26] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 6, 8

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

[28] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *ICLR*, 2024. 2, 6, 8

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

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

[31] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023. 2

[32] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[33] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*, 2024. 2, 3

[34] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *ICCV*, 2023. 2, 3

[35] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *ICCV*, 2023. 2

[36] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023. 3

[37] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Frédo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *NeurIPS*, 2023. 2

[38] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 2

[39] Kalyan Alwala Vasudev, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *CVPR*, 2022. 2

[40] 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

[41] 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

[42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 6

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. In *TIP*, 2004. 6

[44] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 3

[45] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *CVPR*, 2023. 2

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

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

[48] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, 2021. 2

[49] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 6, 7

[50] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 6

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

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