One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization

Minghua Liu[1*] **Chao Xu**[2*] **Haian Jin**[3,4*] **Linghao Chen**[1,4*]
Mukund Varma T[1] **Zexiang Xu**[6]
Hao Su[1]

[1] UC San Diego [2] UCLA [3] Cornell University [4] Zhejiang University [5] Adobe Research

Project Website: `http://one-2-3-45.com`

Abstract

Single image 3D reconstruction is an important but challenging task that requires extensive knowledge of our natural world. Many existing methods solve this problem by optimizing a neural radiance field under the guidance of 2D diffusion models but suffer from lengthy optimization time, 3D inconsistency results, and poor geometry. In this work, we propose a novel method that takes a single image of any object as input and generates a full 360-degree 3D textured mesh in a single feed-forward pass. Given a single image, we first use a view-conditioned 2D diffusion model, Zero123, to generate multi-view images for the input view, and then aim to lift them up to 3D space. Since traditional reconstruction methods struggle with inconsistent multi-view predictions, we build our 3D reconstruction module upon an SDF-based generalizable neural surface reconstruction method and propose several critical training strategies to enable the reconstruction of 360-degree meshes. Without costly optimizations, our method reconstructs 3D shapes in significantly less time than existing methods. Moreover, our method favors better geometry, generates more 3D consistent results, and adheres more closely to the input image. We evaluate our approach on both synthetic data and in-the-wild images and demonstrate its superiority in terms of both mesh quality and runtime. In addition, our approach can seamlessly support the text-to-3D task by integrating with off-the-shelf text-to-image diffusion models.

1 Introduction

Single image 3D reconstruction, the task of reconstructing a 3D model of an object from a single 2D image, is a long-standing problem in the computer vision community and is crucial for a wide range of applications, such as robotic object manipulation and navigation, 3D content creation, as well as AR/VR [47; 9; 92]. The problem is challenging as it requires not only the reconstruction of visible parts but also the hallucination of invisible regions. Consequently, this problem is often ill-posed and corresponds to multiple plausible solutions because of insufficient evidence from a single image. On the other hand, humans can adeptly infer unseen 3D content based on our extensive knowledge of the 3D world. To endow intelligent agents with this ability, many existing methods [31; 19; 25; 11; 87; 91; 16; 83; 39; 10; 37] exploit class-specific priors by training 3D generative networks on 3D shape datasets [4]. However, these methods often fail to generalize to unseen categories, and their reconstruction quality is constrained by the limited size of public 3D datasets.

[*] Equal Contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Figure 1: One-2-3-45 reconstructs a full 360° mesh of any object in 45 seconds given a single image of it. In each example, we showcase the input image in the left column, alongside the generated textured and textureless meshes from three different views.

In this work, we pursue a generic solution to turn an image of any object, regardless of its category, into a high-quality 3D textured mesh. To achieve this, we propose a novel approach that can effectively utilize the strong priors learned by 2D diffusion models for 3D reconstruction. Compared to 3D data, 2D images are more readily available and scalable. Recent 2D generative models (*e.g.*, DALL-E [64; 63], Imagen [69], and Stable Diffusion [68]) and visual-language models (*e.g.*, CLIP [61]) have made significant strides by pre-training on Internet-scale image datasets. Since they learn a wide range of visual concepts and possess strong priors about our 3D world, it is natural to marry 3D tasks with them. Consequently, an emerging body of research [26; 23; 52; 60; 36], as exemplified by DreamField [26], DreamFusion [60], and Magic3D [36], employs 2D diffusion models or vision language models to assist 3D generative tasks. The common paradigm of them is to perform per-shape optimization with differentiable rendering and the guidance of the CLIP model or 2D diffusion models. While many other 3D representations have been explored, neural fields are the most commonly used representation during optimization.

Although these optimization-based methods have achieved impressive results on both text-to-3D [60; 26; 36] and image-to-3D tasks [48; 72], they face some common dilemmas: (a) **time-consuming**. Per-shape optimization typically involves tens of thousands of iterations of full-image volume rendering and prior model inferences, resulting in typically tens of minutes per shape. (b) **memory intensive**. Since the full image is required for the 2D prior model, the volume rendering can be memory-intensive when the image resolution goes up. (c) **3D inconsistent**. Since the 2D prior model only sees a single view at each iteration and tries to make every view look like the input, they often generate 3D inconsistent shapes (*e.g.*, with two faces, or the Janus problem [48; 60]). (d) **poor geometry**. Many methods utilize the density field as the representation in volume rendering. It is common that they produce good RGB renderings but extracting high-quality mesh tends to be difficult.

In this paper, instead of following the common optimization-based paradigm, we propose a novel approach to utilize 2D prior models for 3D modeling. At the heart of our approach is the combination of a 2D diffusion model with a cost-volume-based 3D reconstruction technique, enabling

the reconstruction of a high-quality 360° textured mesh from a single image in a feed-forward pass without per-scene optimization. Specifically, we leverage a recent 2D diffusion model, Zero123 [41], which is fine-tuned on Stable Diffusion [68] to predict novel views of the input image given the camera transformation. We utilize it to generate multi-view predictions of the input single image so that we can leverage multi-view 3D reconstruction techniques to obtain a 3D mesh. There are two challenges associated with reconstruction from synthesized multi-view predictions: (a) the inherent lack of perfect consistency within the multi-view predictions, which can lead to severe failures in optimization-based methods such as NeRF methods [53; 5]. (b) the camera pose of the input image is required but unknown. To tackle them, we build our reconstruction module upon a cost volume-based neural surface reconstruction approach, SparseNeuS [45], which is a variant of MVSNeRF [6]. Additionally, we introduce a series of essential training strategies that enable the reconstruction of 360-degree meshes from inherently inconsistent multi-view predictions. We also propose an elevation estimation module that estimates the elevation of the input shape in Zero123's canonical coordinate system, which is used to compute the camera poses required by the reconstruction module.

By integrating the three modules of multi-view synthesis, elevation estimation, and 3D reconstruction, our method can reconstruct 3D meshes of any object from a single image in a feed-forward manner. Without costly optimizations, our method reconstructs 3D shapes in significantly less time, *e.g.*, in just 45 seconds. Our method favors better geometry due to the use of SDF representations, and generates more consistent 3D meshes, thanks to the camera-conditioned multi-view predictions. Moreover, our reconstruction adheres more closely to the input image compared to existing methods. See Figure 1 for some of our example results. We evaluate our method on both synthetic data and real images and demonstrate that our method outperforms existing methods in terms of both quality and efficiency.

2 Related Work

2.1 3D Generation Guided by 2D Prior Models

Recently, 2D generative models (*e.g.*, DALL-E [64; 63], Imagen [69], and Stable Diffusion [68]) and vision-language models (*e.g.*, CLIP [61]) have learned a wide range of visual concepts by pre-training on Internet-scale image datasets. They possess powerful priors about our 3D world and have inspired a growing body of research to employ 2D prior models for assisting 3D understanding [38; 40] and generative tasks. Exemplified by DreamField [26], DreamFusion [60], and Magic3D [36], a line of works follows the paradigm of per-shape optimization. They typically optimize a 3D representation (*i.e.*, NeRF, mesh, SMPL human model) and utilize differentiable rendering to generate 2D images from various views. The images are then fed to the CLIP model [23; 26; 52; 35; 3; 32; 2; 28; 89; 43] or 2D diffusion model [60; 36; 72; 48; 13; 78; 88; 51; 99; 62; 75] for calculating the loss functions, which are used to guide the 3D shape optimization. In addition to optimization-based 3D shape generation, some works train a 3D generative model but leverage the embedding space of CLIP [8; 44; 71], and some works focus on generating textures or materials for input meshes using 2D models' prior [52; 82; 7; 51; 67].

2.2 Single Image to 3D

Before the emergence of CLIP and large-scale 2D diffusion models, people often learn 3D priors from 3D synthetic data [4] or real scans [65]. Unlike 2D images, 3D data can be represented in various formats and numerous representation-specific 3D generative models have been proposed. By combining 2D image encoder and 3D generators, they generate 3D data in various representations, including 3D voxels [19; 85; 11; 87; 86; 91], point clouds [16; 94; 20; 1; 49; 96], polygon meshes [31; 79; 83; 56], and parametric models [59; 100; 101]. Recently, there has been an increasing number of work on learning to generate a 3D implicit field from a single image [90; 50; 70; 25; 58; 18; 21; 27; 55; 84; 54].

As previously mentioned, several recent works leverage 2D diffusion models to perform per-shape optimization, allowing for the text-to-3D task [60; 36; 26] given that diffusion models are typically conditioned on text. To enable the generation of 3D models from a single image, some works [48; 13; 51] utilize textual inversion [17], to find the best-matching text embedding for the input image, which is then fed into a diffusion model. NeuralLift-360 [24] adds a CLIP loss to enforce similarity between the rendered image and the input image. 3DFuse [72] finetunes the Stable Diffusion model with LoRA layers [24] and a sparse depth injector to ensure greater 3D consistency. A recent work

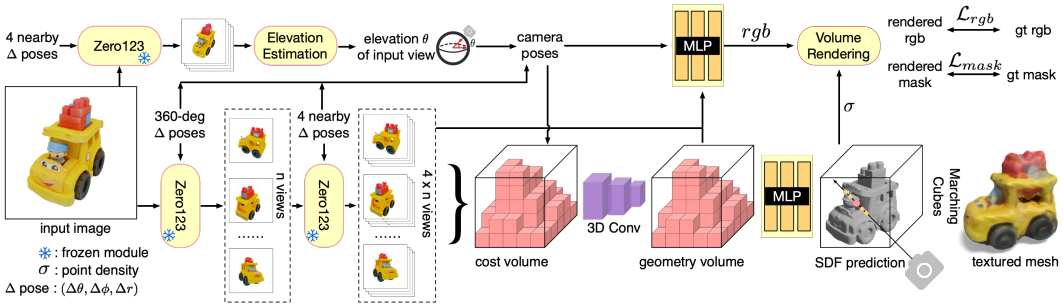

Figure 2: Our method consists of three primary components: (a) **Multi-view synthesis**: we use a view-conditioned 2D diffusion model, Zero123 [41], to generate multi-view images in a two-stage manner. The input of Zero123 includes a single image and a relative camera transformation, which is parameterized by the relative spherical coordinates $(\Delta\theta, \Delta\phi, \Delta r)$. (b) **Pose estimation**: we estimate the elevation angle θ of the input image based on four nearby views generated by Zero123. We then obtain the poses of the multi-view images by combining the specified relative poses with the estimated pose of the input view. (c) **3D reconstruction**: We feed the multi-view posed images to an SDF-based generalizable neural surface reconstruction module for 360° mesh reconstruction.

Zero123 [41; 73] finetunes the Stable Diffusion model [69] to generate a novel view of the input image based on relative camera pose. In addition to these methods, OpenAI trains a 3D native diffusion model Point-E [57], which uses several million internal 3D models to generate point clouds. Very recently, they published another model Shap-E [30] which is trained to generate parameters of implicit functions that can be used for producing textured meshes or neural radiance fields.

2.3 Generalizable Neural Reconstruction

Traditional NeRF-like methods [53; 80] use a neural network to represent a single scene and require per-scene optimization. However, some approaches aim to learn priors across scenes and generalize to novel scenes. These methods typically take a few source views as input and leverage 2D networks for extracting 2D features. The pixel features are then unprojected into 3D space, and a NeRF-based rendering pipeline is applied on top of them. In this way, they can generate a 3D implicit field given a few source views in a single feed-forward pass. Among the methods, some [81; 65; 22; 95; 93; 42; 34; 76; 77] directly aggregate 2D features with MLPs or transformers, while others explicitly construct the 3D feature/cost volume [6; 29; 98; 45], and utilize the voxel feature for decoding density and color. In addition to the density field representation, some methods such as SparseNeuS [45] and VolRecon [66] utilize SDF representations for geometry reconstruction.

3 Method

Our overall pipeline is illustrated in Figure 2. In Section 3.1, we introduce a view-conditioned 2D diffusion model, Zero123 [41], which is used to generate multi-view images. In Section 3.2, we show that traditional NeRF-based and SDF-based methods fail to reconstruct high-quality meshes from inconsistent multi-view predictions even given ground truth camera poses. Therefore, in Section 3.3, we propose a cost volume-based neural surface reconstruction module that can be trained to handle inconsistent multi-view predictions and reconstruct a 3D mesh in a single feed-forward pass. Specifically, we build upon the SparseNeuS [45] and introduce several critical training strategies to support 360° mesh reconstruction. Additionally, in Section 3.4, we demonstrate the necessity of estimating the pose of the input view in Zero123's canonical space for 3D reconstruction. While the azimuth and radius can be arbitrarily specified, we propose a novel module that utilizes four nearby views generated by Zero123 to estimate the elevation of the input view.

3.1 Zero123: View-Conditioned 2D Diffusion

Recent 2D diffusion models [64; 69; 68] have demonstrated the ability to learn a wide range of visual concepts and strong priors by training on internet-scale data. While the original diffusion

Figure 3: NeRF-based method [53] and SDF-based method [80] fail to reconstruct high-quality meshes given multi-view images predicted by Zero123. See Figure 1 for our reconstruction results.

models mainly focused on the task of text-to-image, recent work [97; 24] has shown that fine-tuning pretrained models allows us to add various conditional controls to the diffusion models and generate images based on specific conditions. Several conditions, such as canny edges, user scribbles, depth, and normal maps, have already proven effective [97].

The recent work Zero123 [41] shares a similar spirit and aims to add viewpoint condition control for the Stable Diffusion model [68]. Specifically, given a single RGB image of an object and a relative camera transformation, Zero123 aims to control the diffusion model to synthesize a new image under this transformed camera view. To achieve this, Zero123 fine-tunes the Stable Diffusion on paired images with their relative camera transformations, synthesized from a large-scale 3D dataset [12]. During the creation of the fine-tuning dataset, Zero123 assumes that the object is centered at the origin of the coordinate system and uses a spherical camera, *i.e.*, the camera is placed on the sphere's surface and always looks at the origin. For two camera poses (θ_1, ϕ_1, r_1) and (θ_2, ϕ_2, r_2), where θ_i, ϕ_i, and r_i denote the polar angle, azimuth angle, and radius, their relative camera transformation is parameterized as $(\theta_2 - \theta_1, \phi_2 - \phi_1, r_2 - r_1)$. They aim to learn a model f, such that $f(x_1, \theta_2 - \theta_1, \phi_2 - \phi_1, r_2 - r_1)$ is perceptually similar to x_2, where x_1 and x_2 are two images of an object captured from different views. Zero123 finds that such fine-tuning enables the Stable Diffusion model to learn a generic mechanism for controlling the camera viewpoints, which extrapolates outside of the objects seen in the fine-tuning dataset.

3.2 Can NeRF Optimization Lift Multi-View Predictions to 3D?

Given a single image of an object, we can utilize Zero123 [41] to generate multi-view images, but can we use traditional NeRF-based or SDF-based methods [5; 80] to reconstruct high-quality 3D meshes from these predictions? We conduct a small experiment to test this hypothesis. Given a single image, we first generate 32 multi-view images using Zero123, with camera poses uniformly sampled from the sphere surface. We then feed the predictions to a NeRF-based method (TensoRF [53]) and an SDF-based method (NeuS [80]), which optimize density and SDF fields, respectively. However, as shown in Figure 3, both methods fail to produce satisfactory results, generating numerous distortions and floaters. This is primarily due to the inconsistency of Zero123's predictions. In Figure 4, we compare Zero123's predictions with ground-truth renderings. We can see that the overall PSNR is not very high, particularly when the input relative pose is large or the target pose is at unusual locations (*e.g.*, from the bottom or the top). However, the mask IoU (most regions are greater than 0.95) and CLIP similarity are relatively good. This suggests that Zero123 tends to generate predictions that are perceptually similar to the ground truth and have similar contours or boundaries, but the pixel-level appearance may not be exactly the same. Nevertheless, such inconsistencies between the source views are already fatal to traditional optimization-based methods. Although the original Zero123 paper proposes another method for lifting its multi-view predictions, we will demonstrate in experiments that it also fails to yield perfect results and entails time-consuming optimization.

3.3 Neural Surface Reconstruction from Imperfect Multi-View Predictions

Instead of using optimization-based approaches, we base our reconstruction module on a generalizable SDF reconstruction method SparseNeuS [45], which is essentially a variant of the MVSNeRF [6] pipeline that combines multi-view stereo, neural scene representation, and volume rendering. As illustrated in Figure 2, our reconstruction module takes multiple source images with corresponding camera poses as input and generates a textured mesh in a single feed-forward pass. In this section, we will first briefly describe the network pipeline of the module and then explain how we train the module, select the source images, and generate textured meshes. Additionally, in Section 3.4, we will discuss how we generate the camera poses for the source images.

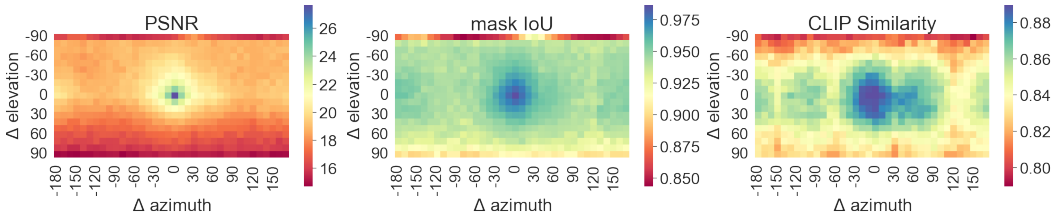

Figure 4: We analyze the prediction quality of Zero123 by comparing its predictions to ground truth renderings across various view transformations. For each view transformation, we report the average PSNR, mask IoU, and CLIP similarity of 100 shapes from the Objaverse [12] dataset. The prediction mask is calculated by considering foreground objects (*i.e.*, non-white regions). Zero123 provides more accurate predictions when the view transformation is small.

As shown in Figure 2, our reconstruction module takes m posed source images as input. The module begins by extracting m 2D feature maps using a 2D feature network. Next, the module builds a 3D cost volume whose contents are computed by first projecting each 3D voxel to m 2D feature planes and then fetching the variance of the features across the m projected 2D locations. The cost volume is then processed using a sparse 3D CNN to obtain a geometry volume that encodes the underlying geometry of the input shape. To predict the SDF at an arbitrary 3D point, an MLP network takes the 3D coordinate and its corresponding interpolated features from the geometry encoding volume as input. To predict the color of a 3D point, another MLP network takes as input the 2D features at the projected locations, interpolated features from the geometry volume, and the viewing direction of the query ray relative to the viewing direction of the source images. The network predicts the blending weights for each source view, and the color of the 3D point is predicted as the weighted sum of its projected colors. Finally, an SDF-based rendering technique is applied on top of the two MLP networks for RGB and mask rendering [80]. In each iteration, we randomly choose one view to build the cost volume and another view for rendering supervision.

2-Stage Source View Selection and Groundtruth-Prediction Mixed Training. Although the original SparseNeuS [45] paper only demonstrated frontal view reconstruction, we have extended it to reconstruct 360-degree meshes in a single feed-forward pass by selecting source views in a particular way. Specifically, our reconstruction model is trained on a 3D object dataset while freezing Zero123. We follow Zero123 to normalize the training shapes and use a spherical camera model. For each shape, we first render n ground-truth RGB images from n camera poses uniformly placed on the sphere. For each of the n views, we use Zero123 to predict four nearby views. During training, we feed all $4 \times n$ predictions with ground-truth poses into the reconstruction module and randomly choose one of the n ground-truth RGB images views as the target view. We call this view selection strategy as *2-stage source view selection*. We supervise the training with both the ground-truth RGB and mask values. In this way, the module can learn to handle the inconsistent predictions from Zero123 and reconstruct a consistent 360° mesh. We argue that our two-stage source view selection strategy is critical since uniformly choosing $n \times 4$ source views from the sphere surface would result in larger distances between the camera poses. However, cost volume-based methods [45; 29; 6] typically rely on very close source views to find local correspondences. Furthermore, as shown in Figure 4, when the relative pose is small (*e.g.*, 10 degrees apart), Zero123 can provide very accurate and consistent predictions and thus can be used to find local correspondences and infer the geometry.

During training, we utilize n ground-truth renderings in the initial stage. We find that employing n predicted images at this stage would suffer from notable inconsistencies across different views, complicating the network's ability to learn sharp details (see examples in ablation study). However, during inference, we can replace the n ground-truth renderings with Zero123 predictions, as shown in Figure 2, the network can automatically generalize to some extent. We will show in the experiments that this groundtruth-prediction mixed training strategy is also important. To export the textured mesh, we use marching cubes [46] to extract the mesh from the predicted SDF field and query the color of the mesh vertices as described in [80]. Although our reconstruction module is trained on a 3D dataset, we find that it mainly relies on local correspondences and can generalize to unseen shapes very well.

Figure 5: Qualitative examples of One-2-3-45 for both synthetic and real images. Each triplet showcases an input image, a textured mesh, and a textureless mesh.

3.4 Camera Pose Estimation

Our reconstruction module requires camera poses for the $4 \times n$ source view images. Note that we adopt Zero123 for image synthesis, which parameterizes cameras in a canonical spherical coordinate frame, (θ, ϕ, r), where θ, ϕ and r represent the elevation, azimuth, and radius. While we can arbitrarily adjust the azimuth angle ϕ and the radius r of all source view images simultaneously, resulting in the rotation and scaling of the reconstructed object accordingly, this parameterization requires knowing the absolute elevation angle θ of one camera to determine the relative poses of all cameras in a standard XYZ frame. More specifically, the relative poses between camera (θ_0, ϕ_0, r_0) and camera $(\theta_0 + \Delta\theta, \phi_0 + \Delta\phi, r_0)$ vary for different θ_0 even when $\Delta\theta$ and $\Delta\phi$ are the same. Because of this, changing the elevation angles of all source images together (*e.g.*, by 30 degrees up or 30 degrees down) will lead to the distortion of the reconstructed shape (see Figure 10 for examples).

Therefore, we propose an elevation estimation module to infer the elevation angle of the input image. First, we use Zero123 to predict four nearby views of the input image. Then we enumerate all possible elevation angles in a coarse-to-fine manner. For each elevation candidate angle, we compute the corresponding camera poses for the four images and calculate a reprojection error for this set of camera poses to measure the consistency between the images and the camera poses. The elevation angle with the smallest reprojection error is used to generate the camera poses for all $4 \times n$ source views by combining the pose of the input view and the relative poses. Please refer to the appendix for details on how we calculate the reprojection error for a set of posed images.

4 Experiments

4.1 Implementation Details

For each input image, we generate $n = 8$ images by choosing camera poses uniformly placed on the sphere surface and then generate 4 local images (10° apart) for each of the 8 views, resulting in 32 source-view images for reconstruction. During training, we freeze the Zero123 [41] model and train our reconstruction module on the Objaverse-LVIS [12] dataset, which contains 46K 3D models in 1,156 categories. We use BlenderProc [14] to render ground-truth RGB images. For images with background, we utilize an off-the-shelf segmentation network SAM [33] with bounding-box prompts for background removal. Please refer to the appendix for more details.

4.2 Single Image to 3D Mesh

We present qualitative examples of our method in Figures 1 and 5, illustrating its effectiveness in handling both synthetic images and real images. We also compare One-2-3-45 with existing zero-shot single image 3D reconstruction approaches, including Point-E [57], Shap-E [30], Zero123 (Stable

Figure 6: We compare One-2-3-45 with Point-E [57], Shap-E [30], Zero123 (Stable Dreamfusion version) [41], 3DFuse [72], and RealFusion [48]. In each example, we present both the textured and textureless meshes. As 3DFuse [72] and RealFusion [48] do not natively support the export of textured meshes, we showcase the results of volume rendering instead.

Table 1: Quantitative Comparison on GSO [15] and Objaverse [12] datasets.

	Prior Source	F-Score GSO	F-Score Obj.	F-Score avg.	CLIP Similarity GSO	CLIP Similarity Obj.	CLIP Similarity avg.	Time
Point-E [57]	internal 3D data	81.0	81.0	81.0	74.3	78.5	76.4	78s
Shap-E [30]		83.4	81.2	82.3	**79.6**	**82.1**	**80.9**	27s
Zero123+SD [41]	2D diffusion models	75.1	69.9	72.5	71.0	72.7	71.9	∼15min
RealFusion [48]		66.7	59.3	63.0	69.3	69.5	69.4	∼90min
3DFuse [72]		60.7	60.2	60.4	71.4	74.0	72.7	∼30min
Ours		**84.0**	**83.1**	**83.5**	76.4	79.7	78.1	45s

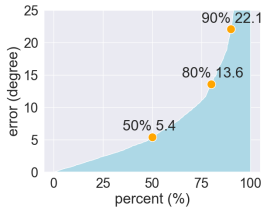

Figure 7: Error distribution of predicted elevations. The median and average are 5.4 and 9.7 degrees.

Dreamfusion version) [41], 3DFuse [72], and RealFusion [48]. Among them, Point-E and Shap-E are two 3D native diffusion models released by OpenAI, which are trained on several million internal 3D data, while others are optimization-based approaches leveraging priors from Stable Diffusion [68].

Figure 6 presents the qualitative comparison. While most methods can generate plausible 3D meshes from a single image, notable differences exist among them in terms of geometry quality, adherence to the input, and overall 3D consistency. In terms of geometry quality, approaches like RealFusion [48] and 3DFuse [72], which optimize a neural radiance field, face challenges in extracting high-quality meshes. Likewise, Point-E [57] produces a sparse point cloud as its output, resulting in numerous holes on the reconstructed meshes. In contrast, our approach utilizes an SDF presentation and favors better geometry. Regarding adherence to the input, we observe that most baseline methods struggle to preserve the similarity to the input image. Although Shap-E performs slightly better, it still produces lots of failure cases (see the backpack without shoulder straps, distorted shoe, and stool with three legs). In contrast, our approach leverages a powerful 2D diffusion model to directly produce high-quality multi-view images, rather than relying on 3D space hallucination. This strategy provides

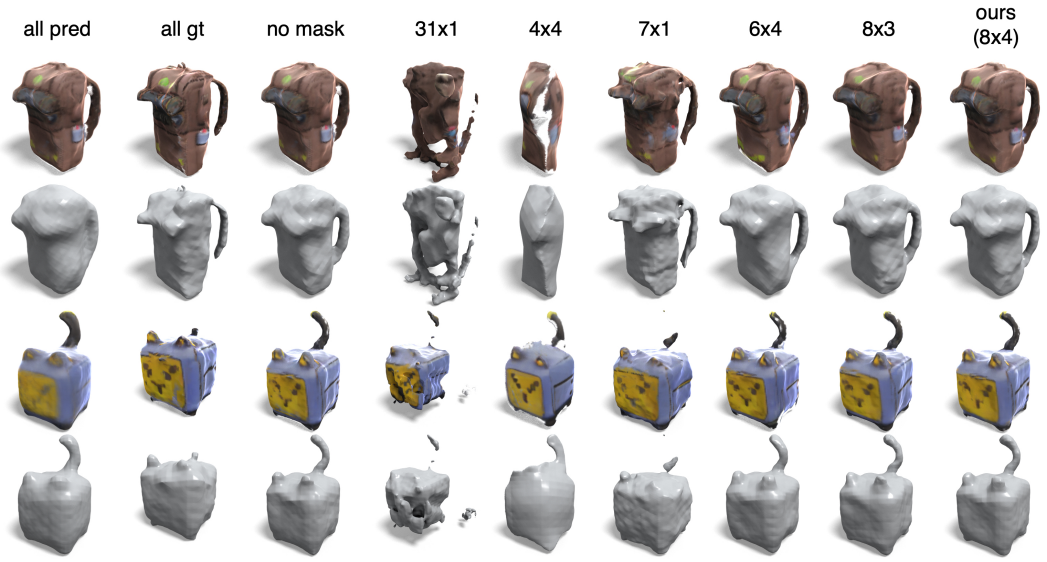

Figure 8: Ablations on training strategies of the reconstruction module and the number of views.

better adherence to the input views, alleviates the burden of the 3D reconstruction module, and yields results that are more finely attuned to the input. Furthermore, many approaches encounter challenges in achieving consistent 3D results (also known as the Janus problem [48; 60]), as highlighted in the right figure (two-handle mug, multi-face Mario, and two-face backpack). One of the contributing factors to this issue is that several methods optimize each view independently, striving to make each view resemble the input. In contrast, our method capitalizes on the view-conditioned 2D diffusion model, inherently enhancing 3D consistency.

We also quantitatively compare the approaches on Objaverse [12] and GoogleScannedObjects (GSO) [15] datasets. For each dataset, we randomly choose 20 shapes and render a single image per shape for evaluation. To align the predictions with the ground-truth mesh, we linearly search the scaling factor and the rotation angle, apply Iterative Closest Point (ICP) for sampled point clouds, and select the one with the most number of inliers. We follow RealFusion [48] to report F-score (with a threshold of 0.05) and CLIP similarity, and the runtime on an A100 GPU. As shown in Table 1, our method outperforms all baseline approaches in terms of F-Score. As for CLIP similarity, we surpass all methods except a concurrent work Shap-E [30]. We find that CLIP similarity is very sensitive to the color distribution and less discriminative in local geometry variations (*i.e.*, the number of legs of a stool, the number of handles of a mug). Regarding running time, our method demonstrates a notable advantage over optimization-based approaches and performs on par with 3D native diffusion models, such as Point-E [57] and Shap-E [30]. Specifically, our 3D reconstruction module reconstructs a 3D mesh in approximately 5 seconds, with the remaining time primarily spent on Zero123 predictions, which takes roughly 1 second per image on an A100 GPU.

4.3 Ablation Study

Training strategies. We ablate our training strategies in Figure 8. We found that without our 2-stage source view selection strategy, a network trained to consume 31 uniformly posed Zero123 predictions (fourth column) suffers from severe inconsistency among source views, causing the reconstruction module to fail completely. If we feed only 7 source views (sixth column) without the four nearby views, the reconstruction fails to capture local correspondence and cannot reconstruct fine-grained geometry. During training, we first render n ground-truth renderings and then use Zero123 to predict four nearby views for each of them. If we train directly on 8×4 ground-truth renderings without Zero123 prediction during training (second column), it fails to generalize well to Zero123 predictions

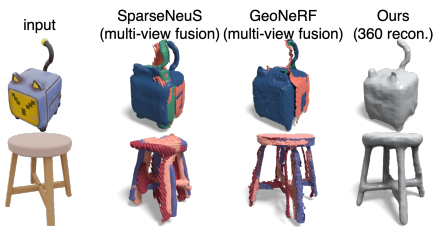

Figure 9: 360° reconstruction vs. multi-view fusion. Meshes from different views are in different colors.

Figure 10: Incorrect elevations lead to distorted reconstruction. Our elevation estimation module can predict an accurate elevation of the input view.

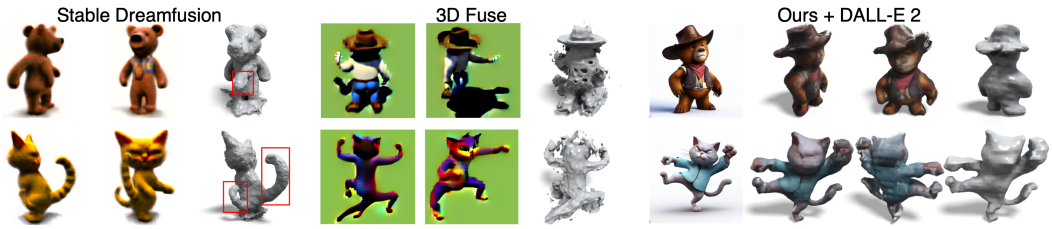

Figure 11: Text to 3D. First row: "a bear in cowboy suit." Second row: "a kungfu cat." We utilize DALL-E 2 [63] to generate an image conditioned on the text and then lift it to 3D. We compare our method with Stable Dreamfusion [60] and 3DFuse [72]. For baselines, volume renderings are shown.

during inference, with many missing regions. Instead, if we replace the n ground-truth renderings with n Zero123 predictions during training (first column), the network fail to generate sharp details (see the strips of the backpack).

Elevation estimation. Our reconstruction module relies on accurate elevation angles of the input view. In Figure 10, we demonstrate the impact of providing incorrect elevation angles (*e.g.*, altering the elevation angles of source views by $\pm 30°$), which results in distorted reconstruction results. Instead, utilizing our predicted elevation angles can perfectly match results with ground truth elevations. We also quantitatively test our elevation estimation module by rendering 1,700 images from random camera poses. As shown in Figure 7, our elevation estimation module predicts accurate elevations.

Number of source views. In Figure 8, we also investigate the impact of varying the number of source views on 3D reconstruction. We observe that our method is not very sensitive to the number of views as long as the reconstruction module is retrained with the corresponding setting.

360° **reconstruction vs. multi-view fusion.** While our method reconstructs a 360° mesh in a single pass, most existing generalizable neural reconstruction approaches [45; 29; 6] primarily focus on frontal view reconstruction. An alternative approach is to independently infer the geometry for each view and subsequently fuse them together. However, we have observed that this strategy often struggles with multi-view fusion due to inconsistent Zero123 predictions, as illustrated in Figure 9.

4.4 Text to 3D Mesh

As shown in Figure 11, by integrating with off-the-shelf text-to-image 2D diffusion models [68; 63], our method can be naturally extended to support text-to-image-3D tasks and generate high-quality textured meshes in a short time. See supplementary for more examples.

5 Conclusion

In this paper, we present a novel method for reconstructing a high-quality 360° mesh of any object from a single image of it. In comparison to existing zero-shot approaches, our results exhibit superior geometry, enhanced 3D consistency, and a remarkable adherence to the input image. Notably, our approach reconstructs meshes in a single forward pass without the need for time-consuming optimization, resulting in significantly reduced processing time. Furthermore, our method can be effortlessly extended to support the text-to-3D task.

Acknowledgments

This work is supported in part by gifts from Qualcomm. We would like to thank Ruoxi Shi, Xinyue Wei, Hansheng Chen, Jiayuan Gu, Fanbo Xiang, Xiaoshuai Zhang, and Yulin Liu for their helpful discussions and manuscript proofreading.

We would like to thank the following sketchfab users for the models used for the demo images in this paper: dimaponomar2019 (backpack), danielpeng (bag), pmlzbt233 (wooden barrel), felixyadomi (cactus), avianinda (burger), shedmon (robocat), ie-niels (stool), phucn (armchair), techCIR (mug), sabriny (fox). All models are CC-By licensed.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018.

[2] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Clipface: Text-guided editing of textured 3d morphable models. *arXiv preprint arXiv:2212.01406*, 2022.

[3] Zehranaz Canfes, M Furkan Atasoy, Alara Dirik, and Pinar Yanardag. Text and image guided 3d avatar generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4421–4431, 2023.

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

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022.

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[7] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023.

[8] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander Schwing, and Liangyan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. *arXiv preprint arXiv:2212.04493*, 2022.

[9] Han-Pang Chiu, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Automatic class-specific 3d reconstruction from a single image. *CSAIL*, pages 1–9, 2009.

[10] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. 2023.

[11] 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016.

[12] 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.

[13] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *arXiv preprint arXiv:2212.03267*, 2022.

[14] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023.

[15] 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 *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[16] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[18] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022.

[19] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 484–499. Springer, 2016.

[20] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018.

[21] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023.

[22] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4709, 2021.

[23] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022.

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[25] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Planes vs. chairs: Category-guided 3d shape learning without any 3d cues. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 727–744. Springer, 2022.

[26] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022.

[27] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021.

[28] Nikolay Jetchev. Clipmatrix: Text-controlled creation of 3d textured meshes. *arXiv preprint arXiv:2109.12922*, 2021.

[29] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022.

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

[31] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collectionsgirdhar2016learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018.

[32] Nasir Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Text to mesh without 3d supervision using limit subdivision. *arXiv preprint arXiv:2203.13333*, 2022.

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[34] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 198–216. Springer, 2022.

[35] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022.

[36] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

[37] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11596–11603, 2020.

[38] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *arXiv preprint arXiv:2305.10764*, 2023.

[39] Minghua Liu, Minhyuk Sung, Radomir Mech, and Hao Su. Deepmetahandles: Learning deformation meta-handles of 3d meshes with biharmonic coordinates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12–21, 2021.

[40] Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21736–21746, 2023.

[41] 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. *arXiv preprint arXiv:2303.11328*, 2023.

[42] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022.

[43] Zhengzhe Liu, Peng Dai, Ruihui Li, Xiaojuan Qi, and Chi-Wing Fu. Iss: Image as stetting stone for text-guided 3d shape generation. *arXiv preprint arXiv:2209.04145*, 2022.

[44] Zhengzhe Liu, Peng Dai, Ruihui Li, Xiaojuan Qi, and Chi-Wing Fu. Iss++: Image as stepping stone for text-guided 3d shape generation. *arXiv preprint arXiv:2303.15181*, 2023.

[45] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 210–227. Springer, 2022.

[46] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[47] Oier Mees, Maxim Tatarchenko, Thomas Brox, and Wolfram Burgard. Self-supervised 3d shape and viewpoint estimation from single images for robotics. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6083–6089. IEEE, 2019.

[48] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 {\deg} reconstruction of any object from a single image. *arXiv preprint arXiv:2302.10663*, 2023.

[49] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. $pc2$: Projection—conditioned point cloud diffusion for single-image 3d reconstruction. *arXiv preprint arXiv:2302.10668*, 2023.

[50] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

[51] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022.

[52] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022.

[53] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[54] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022.

[55] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2022.

[56] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International conference on machine learning*, pages 7220–

7229. PMLR, 2020.

[57] 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.

[58] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[59] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[60] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[62] 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.

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

[64] 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.

[65] 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 *International Conference on Computer Vision*, 2021.

[66] Yufan Ren, Fangjinhua Wang, Tong Zhang, Marc Pollefeys, and Sabine Süsstrunk. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. *arXiv preprint arXiv:2212.08067*, 2022.

[67] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[69] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[70] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019.

[71] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022.

[72] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023.

[73] 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.

[74] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[75] 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. *arXiv preprint arXiv:2303.14184*, 2023.

[76] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021.
[77] Mukund Varma, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that nerf needs? In *The Eleventh International Conference on Learning Representations*, 2022.
[78] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022.
[79] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018.
[80] 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. *arXiv preprint arXiv:2106.10689*, 2021.
[81] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.
[82] Jiacheng Wei, Hao Wang, Jiashi Feng, Guosheng Lin, and Kim-Hui Yap. Taps3d: Text-guided 3d textured shape generation from pseudo supervision, 2023.
[83] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1042–1051, 2019.
[84] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. *arXiv preprint arXiv:2301.08247*, 2023.
[85] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in neural information processing systems*, 30, 2017.
[86] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019.
[87] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12):2919–2935, 2020.
[88] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 {\deg} views. *arXiv preprint arXiv:2211.16431*, 2022.
[89] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704*, 2022.
[90] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019.
[91] Farid Yagubbayli, Yida Wang, Alessio Tonioni, and Federico Tombari. Legoformer: Transformers for block-by-block multi-view 3d reconstruction. *arXiv preprint arXiv:2106.12102*, 2021.
[92] Daniel Yang, Tarik Tosun, Benjamin Eisner, Volkan Isler, and Daniel Lee. Robotic grasping through combined image-based grasp proposal and 3d reconstruction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6350–6356. IEEE, 2021.
[93] Hao Yang, Lanqing Hong, Aoxue Li, Tianyang Hu, Zhenguo Li, Gim Hee Lee, and Liwei Wang. Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. *arXiv preprint arXiv:2303.11052*, 2023.
[94] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018.
[95] 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.
[96] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint*

arXiv:2210.06978, 2022.

[97] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[98] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5449–5458, 2022.

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

[100] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3955–3963, 2018.

[101] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6365–6373, 2017.

Appendix

We first show more qualitative comparison in Section A, which is followed by a demonstration of additional examples on real-world images and the text-to-3D task in Sections B and C respectively. Furthermore, we present the details of our elevation estimation module in Section D, training and evaluation details in Section E. We finally show the failure cases and discuss the limitations in Section F.

A More Qualitative Comparison

Figure 12: We compare One-2-3-45 with Point-E [57], Shap-E [30], Zero123 (Stable Dreamfusion version) [41], 3DFuse [72], and RealFusion [48]. In each example, we present both the textured and textureless meshes. As 3DFuse [72] and RealFusion [48] do not natively support the export of textured meshes, we showcase the results of volume rendering instead.

In Figure 12, we demonstrate more qualitative comparison on Objaverse [12] and GoogleScannedObjects (GSO) [15] datasets. Note that all test shapes are not seen during the training of our 3D reconstruction module.

B More Examples on Real-World Images

In Figure 13, we showcase more examples on real-world images and compare our method with the concurrent method Shap-E [30]. The input images are from `unsplash.com` or captured by ourselves. Note that our results exhibit a closer adherence to the input image.

C More Examples on Text-to-3D

In Figure 14, we present additional examples for the text-to-3D task. It is evident that existing approaches struggle to capture fine-grained details, such as a tree hollow, or achieve compositionality, as seen in examples like an orange stool with green legs, a pineapple-shaped Havana hat, or a rocking horse chair. In contrast, our method produces superior results that adhere more closely to the input text. We hypothesize that controlling such fine-grained attributes in the 3D space using existing optimization strategies is inherently challenging. However, by leveraging established 2D

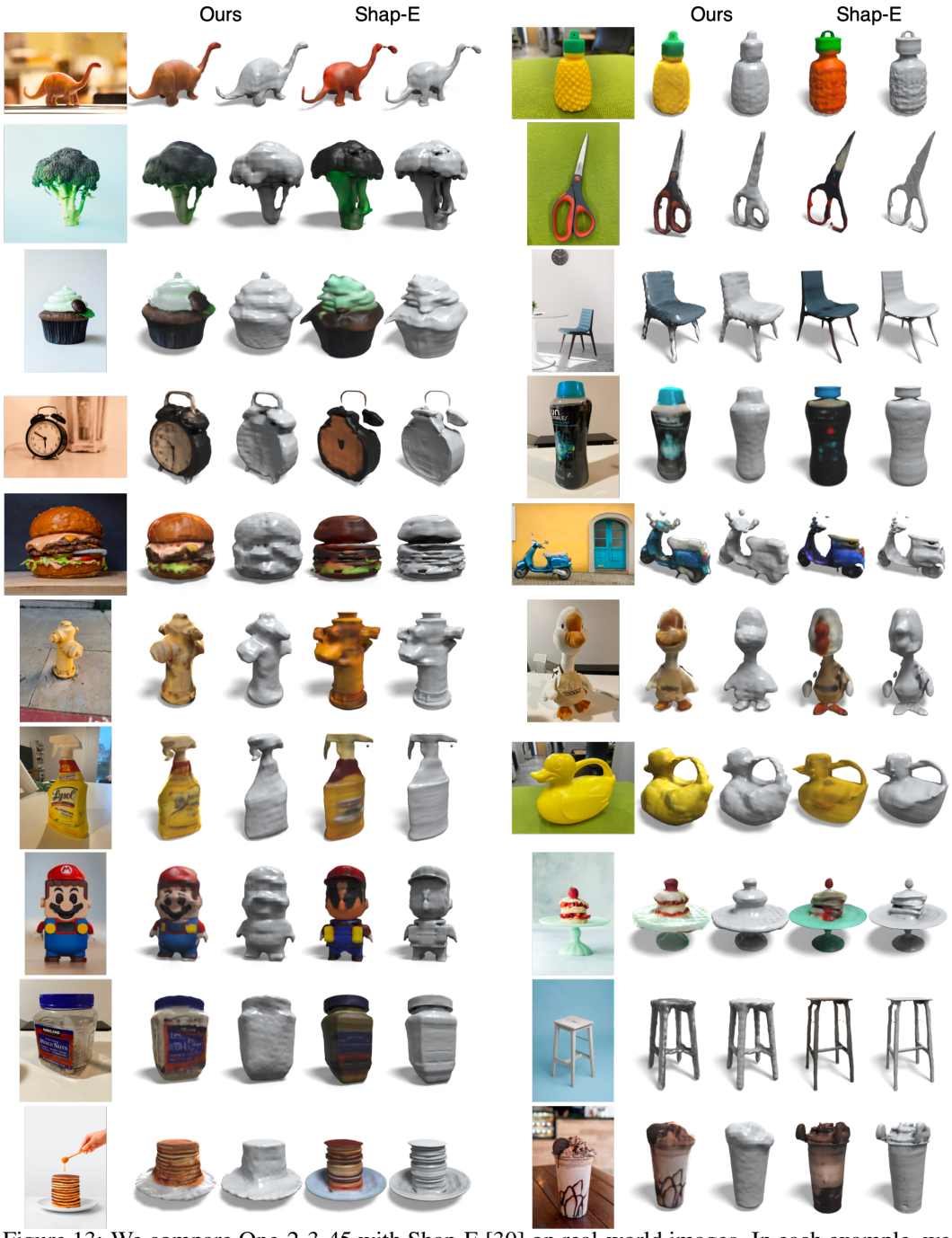

Figure 13: We compare One-2-3-45 with Shap-E [30] on real-world images. In each example, we present the input image, generated textured and textureless meshes.

text-to-image diffusion models, our method becomes more effective in lifting a single 2D image to a corresponding 3D textured mesh.

D Details of Elevation Estimation

To estimate the elevation angle θ of the input image, we first utilize Zero123 [41] to predict four nearby views (10 degrees apart) of the input view. With these predicted views, we proceed to enumerate all possible elevation angles and compute the re-projection error for each candidate angle.

Figure 14: Text-to-3D: We compare our method against two native text-to-3D approaches Stable DreamFusion [60] and 3DFuse [72]. To enable text-to-3D, our method first uses a pretrained text-to-image model DALL-E 2 [63] to generate an image from input text (prompted with "3d model, long shot"), and then uplifts the image to a 3D textured mesh.

The re-projection error assesses the consistency between camera poses and image observations, akin to the bundle adjustment module employed in the Structure-from-Motion (SfM) pipeline.

Specifically, we enumerate all candidate elevation angles in a coarse-to-fine manner. In the coarse stage, we enumerate elevation angles with a 10-degree interval. Once we have determined the elevation angle e^* associated with the smallest re-projection error, we proceed to the fine stage. In this stage, we enumerate elevation angle candidates ranging from $e^* - 10^\circ$ to $e^* + 10^\circ$ with a 1-degree

interval. This coarse-to-fine design facilitates rapid estimation, completing the elevation estimation module in under 1 second for each shape.

Given a set of four predicted nearby views, we perform feature matching to identify corresponding keypoints across each pair of images (a total of six pairs) using an off-the-shelf module LoFTR [74]. For each elevation angle candidate, we calculate the camera pose for the input image by employing the spherical coordinate system with a radius of 1.2 and an azimuth angle of 0. Note that the azimuth angle ϕ and the radius r can be arbitrarily adjusted, resulting in the rotation and scaling of the reconstructed object accordingly. Subsequently, we obtain the camera poses for the four predicted views by incorporating the specified delta poses.

Once we have the four posed images, we compute the re-projection error by enumerating triplet images. For each triplet of images (a, b, c) sharing a set of keypoints P, we consider each point $p \in P$. Utilizing images a and b, we perform triangulation to determine the 3D location of p. We then project the 3D point onto the third image c and calculate the reprojection error, which is defined as the $l1$ distance between the reprojected 2D pixel and the estimated keypoint in image c. By enumerating all image triplets and their corresponding shared keypoints, we obtain the mean projection error for each elevation angle candidate.

E Details of Training and Evaluation

Training We train the reconstruction module using the following loss function:

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda_1 \mathcal{L}_{eikonal} + \lambda_2 \mathcal{L}_{sparsity} \tag{1}$$

where $\mathcal{L}_{rgb}$ represents the $l1$ loss between the rendered and ground truth color, weighted by the sum of accumulated weights; $\mathcal{L}_{eikonal}$ and $\mathcal{L}_{sparsity}$ are the Eikonal and sparsity terms, respectively, following SparseNeuS [45]. We empirically set the weights as $\lambda_0 = 1$, $\lambda_1 = 0.1$, and $\lambda_2 = 0.02$. For λ_2, we adopt a linear warm-up strategy following SparseNeuS [45]. To train our reconstruction module, we utilize the LVIS subset of the Objaverse [12] dataset, which consists of 46k 3D models across 1,156 categories. The reconstruction module is trained for 300k iterations using two A10 GPUs, with the training process lasting approximately 6 days. It is important to note that our reconstruction module does not heavily rely on large-scale training data, as it primarily leverages local correspondence to infer the geometry, which is relatively easier to learn and generalize.

Evaluation We evaluate all baseline approaches using their official codebase. Since the approaches take only a single image as input, the predicted mesh may not have the same scale and transformation as the ground-truth mesh. To ensure a fair comparison, we employ the following process to align the predicted mesh with the ground-truth mesh. First, we align the up direction for the results generated by each approach. Next, for each generated mesh, we perform a linear search over scales and rotation angles along the up direction. After applying each pair of scale and z-rotation, we utilize the Iterative Closest Point (ICP) algorithm to align the transformed mesh to the ground-truth mesh. Finally, we select the mesh with the largest number of inliers as the final alignment. This alignment process helps us establish a consistent reference frame for evaluating the predicted meshes across different approaches. To calculate CLIP similarity, we render both ground-truth and generated meshes, capturing 24 views around the 3D shape from fixed viewpoints - 12 views at 30° elevation and 12 views at 0° elevation.

F Failure Cases and Limitations

Our method relies on Zero123 for generating multi-view images, which introduces challenges due to its occasional production of inconsistent results. In Figure 15, we present two typical cases that exemplify such inconsistencies. The first case involves an input view that lacks sufficient information, such as the back view of a fox. In this scenario, Zero123 struggles to generate consistent predictions for the invisible regions, such as the face of the fox. As a consequence, our method may encounter difficulties in accurately inferring the geometry for those regions. The second case involves an input view with ambiguous or complex structures, such as the pulp and peel of a banana. In such situations, Zero123's ability to accurately infer the underlying geometry becomes limited. As a result, our method may be affected by the inconsistent predictions generated by Zero123. It is important to acknowledge that these limitations arise from the occasional scenarios, and they can impact the

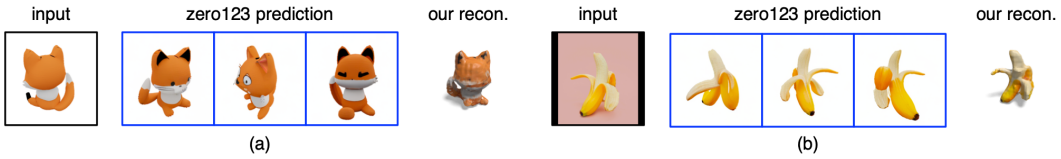

Figure 15: Failure cases. Our method relies on Zero123 to generate multi-view images, and we encounter challenges when Zero123 generates inconsistent results. (a) The input view lacks sufficient information. (b) The input view contains ambiguous or complicated structures.

performance of our method in certain cases. Addressing these challenges and refining the reliability of Zero123's predictions remain areas for further investigation and improvement.

We have also noticed slight artifacts on the back side of our generated results. As one of the first works in combining view-conditioned 2D diffusion models with generalizable multi-view reconstruction, we believe that there is still ample room for exploring more advanced reconstruction techniques and incorporating additional regularizations. By doing so, we expect to significantly mitigate the minor artifacts and further enhance results in the future.