

Slice3D: Multi-Slice, Occlusion-Revealing, Single View 3D Reconstruction

Yizhi Wang¹, Wallace Lira¹, Wenqi Wang², Ali Mahdavi-Amiri¹, Hao Zhang¹

¹Simon Fraser University ²Tsinghua University

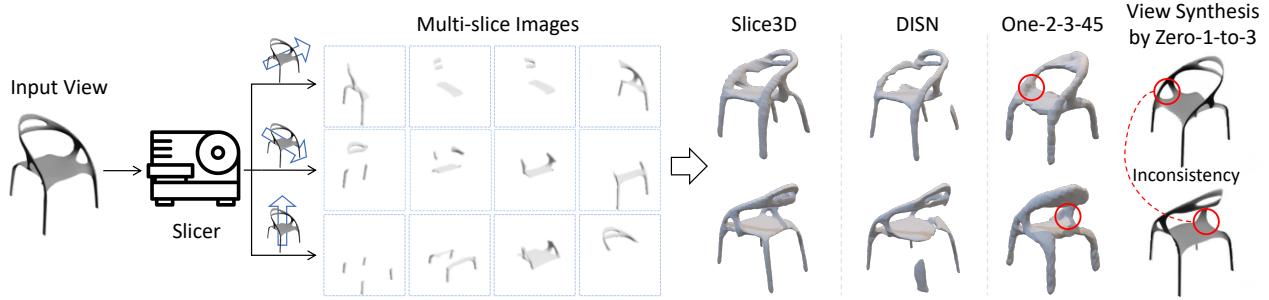


Figure 1. Our single-view 3D reconstruction method, Slice3D, predicts *multi-slice images* to reveal occluded parts without changing the camera (in contrast to *multi-view synthesis*), and then lifts the slices into a 3D model. This leads to better recovery of occluded parts than both encoder-decoder based methods such as DISN [67] and 3D reconstructions from diffusion-synthesized multi-view images (via Zero-1-to-3 [30]) such as One-2-3-45 [29]. The key challenge of the latter is multi-view consistency, or lack thereof, as shown on the right.

Abstract

We introduce *multi-slice reasoning*, a new notion for single-view 3D reconstruction which challenges the current and prevailing belief that *multi-view synthesis* is the most natural conduit between single-view and 3D. Our key observation is that *object slicing* is more advantageous than altering views to reveal occluded structures. Specifically, *slicing* can *peel through* any *occluder* without *obstruction*, and in the limit (infinitely many slices), it is guaranteed to *unveil all hidden object parts*. We realize our idea by developing *Slice3D*, a novel method for single-view 3D reconstruction which first predicts *multi-slice images* from a single *RGB* image and then integrates the slices into a 3D model using a coordinate-based transformer network for signed distance prediction. The slice images can be regressed or generated, both through a *U-Net* based network. For the former, we inject a learnable slice indicator code to designate each decoded image into a spatial slice location, while the slice generator is a denoising diffusion model operating on the entirety of slice images stacked on the input channels. We conduct extensive evaluation against state-of-the-art alternatives to demonstrate superiority of our method, especially in recovering complex and severely occluded shape structures, amid ambiguities. All *Slice3D* results were produced by networks trained on a single *Nvidia A40 GPU*, with an inference time less than 20 seconds.

1. Introduction

Single-view 3D reconstruction has been one of the extensively studied problems in computer vision. Despite the rapid advances on learning-based approaches to tackle this problem, including the latest attempts to leverage large foundational models and powerful diffusion-based generators, the most fundamental challenge still remains: how to faithfully reconstruct *occluded parts* from just one view?

Conventional encoder-decoder based methods, e.g., [9, 17, 26, 45, 55, 67], rely on direct 3D supervision to learn mappings from global or local image features to 3D, but such features from just a single view often do not have the capacity to handle severe occlusion. Also, recent evi-

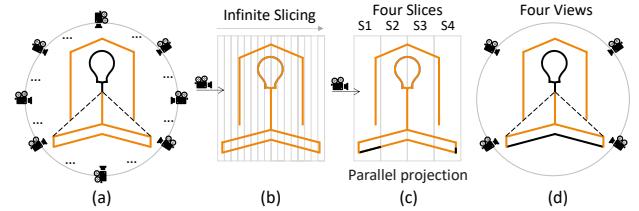


Figure 2. Multi-view vs. multi-slice for occlusion revelation, with orange edges showing revealed parts and black edges showing occluded parts. In the limit, multi-view (i.e., infinitely many views) may still leave undisclosed parts (a), e.g., the bulb, while infinitely many slices would guarantee complete structure revelation (b). With limited slices (c) and views (d), four each, the multi-slice approach tends to reveal more shape structures.

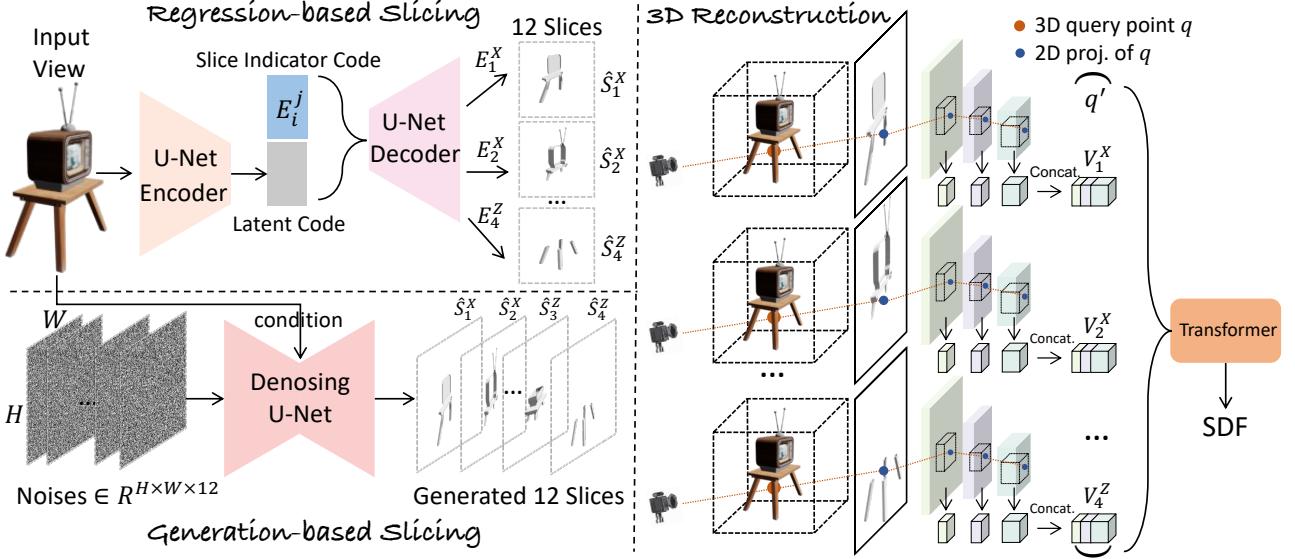


Figure 3. Pipeline of Slice3D, for single-view 3D reconstruction from multi-slice images. The illustrated example shows 4 slices per X , Y , Z direction. The slice images can be obtained by either a regressor or a generator (details in Section 3.2). Regression is carried out by a U-Net encoder-decoder, where the decoder takes an extra *slice indicator code* that is appended to the encoder output, to produce the corresponding slice image. There is one indicator code for each of the $3 \times 4 = 12$ slices, that is optimized during training. Our generator is a denoising diffusion model that is trained to produce a concatenation of slice images following the distribution of the ground-truth (GT) images, when conditioned on the single-view input. All images are of dimension $H \times W$. Inference from random noises may output multiple plausible results. The 3D reconstruction stage takes as input the predicted slice images to produce an implicit field; see Section 3.3. We first extract hierarchical CNN features from the input slices, then project a 3D query point onto the 2D planes of the feature maps, and finally gather all the query-related features and send them into a transformer [60] to predict the signed distance at the query point.

dence suggests that many of these methods are essentially performing image classification or retrieval [57]. As such, occlusion recovery is by and large a “hit or miss”. Differentiable rendering enables 3D reasoning from images without 3D supervision [7, 39], but image-space rendering losses alone do not account for occluded structures.

Recently, more and more works model the reconstruction problem as a single-view conditioned *generative* task, aiming to learn a distribution of unseen structures. Typically, *multi-view* images are generated from single view via diffusion. This is followed by multi-view reconstruction, e.g., via neural radiance fields (NeRF) [6, 29, 43], to obtain 3D outputs. However, multi-view consistency is a reoccurring challenge (see Fig. 1), especially over disparate views.

In this paper, we introduce a novel approach to single-view 3D reconstruction which is *occlusion-revealing* (see Fig. 2) by first predicting *multi-slice images* from the single-view input and then integrating the slices into a 3D model, as illustrated in Fig. 3. In Fig. 4, we show slice images provided during training along a set of slicing directions. Each slice represents the projected image, in the input camera view, of one slice of the 3D *mesh* model. For the TV example shown, one can observe that a completely occluded leg, in the camera view, is revealed in the slice images.

The motivations behind our multi-slice reconstruction

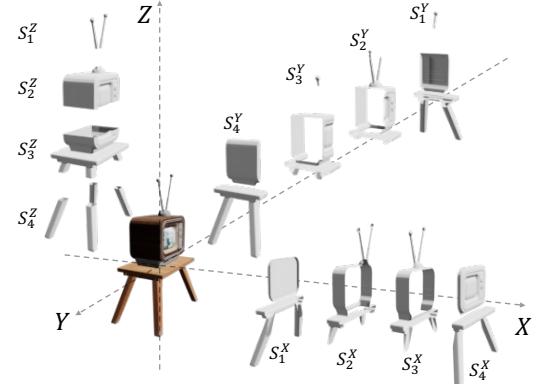


Figure 4. Our slice image prediction networks are trained using occlusion-revealing multi-slice images produced from a GT 3D object *mesh* along a predetermined set of slicing directions (the X , Y , Z axes of the object bounding box in the example above), while keeping the hollows as holes in the images, through which the *back faces* of the mesh can be visible. The 3D mesh may be textured, but the slice images are all rendered without texture.

are three-fold. First, slicing through a volume or 3D mesh is a natural way to reveal occluded parts. Second, with the slice images forming an intermediate representation, our approach breaks the difficult problem of mapping from single view to 3D into two comparatively simpler tasks: slice pre-

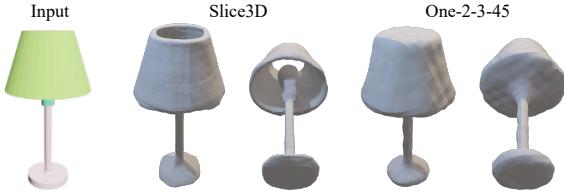


Figure 5. Single-view reconstruction of severely occluded parts in a lamp (e.g., the bulb) by Slice3D vs. One-2-3-45 [29], a state-of-the-art multi-view approach. The multi-view images in [29] were obtained by Zero-1-to-3 [30], which was trained on renderings of 800K 3D models [10]. It is conceivable that with a limited number of views (12, the same as our slice count), the bulbs were barely visible due to occlusion by the lamp shades and bases.

diction followed by 3D reconstruction from slices.

More importantly, when contrasting our multi-*slice* representation with the popular multi-view approach for occlusion revelation, we observe that slicing in the limit (i.e., with infinitely many slices) is *guaranteed* to unveil all hidden structures, while the same cannot be said about multi-view captures, as illustrated in Fig. 2 (a-b), for a 2D lamp example. In practice, when slice and view counts are both limited, the slicing approach is more occlusion-revealing in general since it can peel through any occluders without any obstruction, unlike the multi-view case; see (c-d). On the technical front, neither our slice prediction step nor the 3D reconstruction step is concerned with camera view changes or multi-view consistency, which facilitates the use of convolutional features. In Fig. 5, we show the above contrasting characteristics of the multi-slice vs. multi-view approaches manifested in a real 3D lamp reconstruction.

As outlined in Fig. 3, our slice images can be produced by either a regressive network or a generative one, both employing a U-Net architecture, with network losses measured against ground-truth (GT) images obtained by slicing 3D models from the training set; see Fig. 4. In the regressive network, we inject a learnable slice indicator code into the encoder-decoder setup to designate each decoded image into the proper spatial slice location. The slice generator is a denoising diffusion model operating on separate channels for different slice images. In the second stage, 3D reconstruction from the obtained slice images, we design a coordinate-based transformer to obtain an implicit field.

Our network is coined Slice3D. We conduct extensive experiments on ShapeNet [5] and Objverse [10] for evaluation. Quantitative and qualitative comparisons are made to representative methods including encoder-decoders capturing local image features [67], autoregressive latent models [37], generative model inversion [43], and three recent methods with diffusion + NeRF [6, 29, 34], demonstrating superiority of our method in terms of reconstruction quality, generalizability, and handling of ambiguities.

2. Related work

The earliest inspiration for our multi-slice 3D reconstruction idea was GANHopper by Lira et al. [27]. In the context of unsupervised image-to-image translation, the key idea of their work is to *gradually* execute the translation between disparate domains by making small “hops,” rather than completing it in one go. We were then motivated to tackle the single-view 3D reconstruction problem as a domain translation task, hopping through the multi-slice images so as to bridge the gap between 2D and 3D.

Neural fields. Neural implicit functions [65] are learned by a neural network to predict such properties as color, occupancy, opacity, and surface normals. Earlier works such as IM-Net [8], occupancy networks [35], and DeepSDF [42] represent a 3D shape as a latent code and send it to a decoder formed by MLPs to predict occupancy or SDF values for any query point sampled in 3D. NeRF [36] and its variants [15] represent a 3D scene using a neural radiance field for novel view synthesis. Many works on neural 3D reconstruction [15, 65] take on single- or multi-view images as well as point clouds as input. In contrast, the reconstruction module in Slice3D is built on multi-slice images.

Single-view reconstruction from image features. An early work to exploit neural image features for 3D reconstruction was Pixel2Mesh [61], which uses a graph representation. Subsequently, neural implicit fields have been widely adopted for single-view reconstruction. IM-Net [8] first extracts a global latent code, from an input view, to learn a 3D implicit field. Auto-SDF [37] and ShapeFormer [68] employ vector quantization (VQ) [59] to improve the representation of latent features. More recent works [23, 38, 70] consider single-view reconstruction as a 3D generation task conditioned on the latent code of the input view. However, a global latent code is insufficient to capture finer details on a 3D object, often leading to overly smoothed or retrieved outcomes that do not faithfully represent the input. To alleviate this problem, both DISN [67] and D²IM-Net [26] opted to learn local images features, along with global ones, to improve the recovery of fine-grained topological (e.g., thin) features and surface details, respectively.

3D reconstruction from synthesized multi-views. Conventional 3D reconstruction approaches [14] involve leveraging multi-view reasonings. With recent advances in generative AI, a prevailing practice in single-view 3D reconstruction is to first employ generative models such as GANs [16, 24] and diffusion models [11, 21, 46, 52] to generate several novel views from the input view, and then feed them to a multi-view 3D reconstruction module such as NeRF. NeRF-based surface reconstruction techniques [40, 62, 69] can be further employed to output 3D mesh models. EG3D [3] and NeRF-from-Image [43] build NeRFs or its variants upon the outputs of a style-GAN [24], while more recent

works [4, 6, 18, 29–31, 34, 43, 44, 53, 64, 66] are favoring diffusion models such as Stable Diffusion (SD) [46]. To date, Zero-1-to-3 [30] represents the state of the art on single- to multi-view synthesis; it is built on SD and fine-tuned by 800K 3D models from Objaverse [10].

To alleviate the multi-view consistency problem, which is an reoccurring challenge for the image-space generative methods, recent methods [31, 32, 44, 54, 73] aim to enhance the consistency by performing spatial attention between different views, which dramatically increases the computational overhead. Nevertheless, the consistency of two disparate views is still far from satisfactory. A deeper reason is that standard convolutional neural networks (CNNs) are not inherently rotation-invariant, making it difficult to build correspondences between significantly different views.

Slice representation. There have been techniques to use object slices to reconstruct a 3D shape, e.g., as inspired by Magnetic Resonance Imaging (MRI). Early works in computer graphics employ either parallel cross-sectional curves, e.g., [25], or non-parallel curve networks, e.g., [28], as inputs. Recently, with the advent of machine learning, OReX [49] and Cut-and-Approximate [41] develop neural reconstruction models to obtain 3D meshes, also from a sparse collection of 2D planar cross-sections, while Neural-Slice [22] uses slicing to obtain 3D shapes from 4D tetrahedral meshes. However, none of these techniques, nor others that we are aware of, address the problem of single-view 3D reconstruction since their inputs all consist of 3D data such as curve slices or point clouds. Moreover, our method learns to predict much *fewer* slice *images* and use them as input to a transformer-based reconstruction module.

Dip transform. Dip transform [2] is an ingenious way of acquiring occluded structures of a 3D object by measuring fluid displacement when the object is repeatedly oriented and dipped into the fluid. Similar to slicing, their method was also designed to “reach” occluded parts that optical captures via lines of sight cannot. However, both the approaches to occlusion revelation (liquid penetration vs. slicing) and the target applications (3D acquisition vs. single-view 3D reconstruction) are completely different.

Multi-plane image (MPI). Introduced by Zhou et al. [74], an MPI representation for a scene consists of a set of fronto-parallel planes at fixed depths, with each image intended to capture scene appearance and visibility at a specific depth. Conceptually, both MPIs and our slice images provide a layered scene representation for occlusion revelation. However, MPIs [13, 19, 58, 71, 74] have only been employed for novel view synthesis, not 3D reconstruction, and the prediction networks were trained on *multi-view images*, e.g., from Youtube videos [74]. The generated MPIs only reveal a small amount of occluded scenes to account for view shifts. In contrast, our method directly cuts through a 3D object to reveal all the occluded parts at several depth ranges.

3. Method

Our method is based on a slicing representation; see Section 3.1 and Fig. 4. Given a single-view RGB image $I \in \mathbb{R}^{H \times W \times 3}$ of a 3D object, we first learn a slicing network (Section 3.2) which takes I as input and outputs its corresponding multi-slice images. Then, an SDF prediction network is trained (Section 3.3) to leverage image features from the slices to produce an implicit field of the object.

3.1. Slicing Representation

As shown in Fig. 4, given a 3D object *mesh*, we slice it along the X , Y , and Z axes of its bounding box and render the sliced segments. The slice images can optionally include textures or have holes filled. By default, we remove the textures and keep the holes, which could reveal back faces of the mesh, and overall, it produces the best experimental results. The rendered slice images consist of $\{\hat{S}_i^j | 1 \leq i \leq N_s, j \in \{X, Y, Z\}\}$, where $\hat{S}_i^j \in \mathbb{R}^{H \times W \times 1}$, j denotes the axis and N_s is the number of slices along each axis. In our current implementation, we produce $N_s = 4$ slices per direction. Compared to planar cross-sections or tomography scans which consist of densely stacked planar images, our slicing representation is more compact with each slice image revealing more object semantics.

3.2. Slice Regression and Generation

Given an image I , observing a 3D shape from a single view, we aim to train a slicing network F_s that can produce the slice images from I : $F_s(I) = \{\hat{S}_i^j\}$. We have two different designs for such a network: regressive and generative. The former aims to regress GT slices, assuming deterministic occluded parts based on the partial observation of the input view. The latter aims to *generate* slices that are consistent with the input view, considering various possible occluded parts that do not necessarily match the GT slices.

3.2.1 Regression-based Slicing

We employ a U-Net [47] with an encoder and a decoder to regress the slices from I . As shown in the top-left of Fig. 3, an indicator code $E_i^j \in \mathbb{R}^{N_e}$ is concatenated with the latent code of I to indicate which slice to output:

$$\hat{S}_i^j = F_d([F_e(I); E_i^j]), \quad (1)$$

where F_e and F_d are the encoder and decoder of F_s , respectively, and $[;]$ denotes concatenation.

The indicator codes $\{E_i^j\}$ are randomly initialized and jointly optimized with the whole network. The loss function \mathcal{L}_{reg} is the reconstruction error measured by L_1 -Norm and perceptual loss [72] (\mathcal{L}_p) for all slices:

$$\mathcal{L}_{reg} = \frac{\sum_{i,j} \|\hat{S}_i^j - S_i^j\|_1 + \mathcal{L}_p(\hat{S}_i^j, S_i^j)}{|S|}, \quad (2)$$

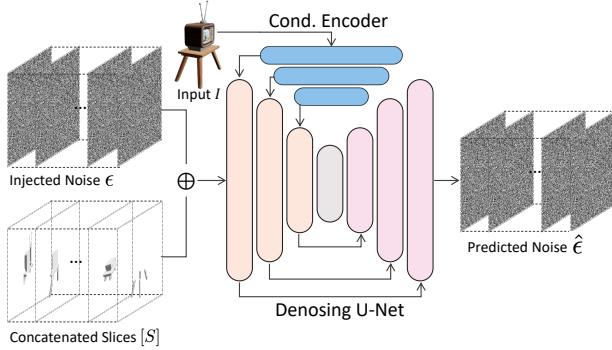


Figure 6. The pipeline of generation-based slicing via DDPM. \oplus denotes the injection of noise to slice images.

where $|S| = 3 \cdot N_s$ is the total number of slices.

3.2.2 Generation-based Slicing

Single-view occlusion revelation is clearly ill-posed, with potential ambiguities leading to multiple plausible 3D reconstructions. To this end, we consider slicing as an image-conditioned generation task and develop a generation-based slicing via Denoising Diffusion Probabilistic Models (DDPM) [21], whose detailed pipeline is shown in Fig. 6.

The key challenge of applying DDPM lies in how to generate *consistent* slices. Rather than diffusing each single slice, we propose to concatenate all slices into an aggregated structure $[S]$, along either the color dimension ($[S] \in \mathbb{R}^{H \times W \times |S|}$) or a spatial dimension such as height ($[S] \in \mathbb{R}^{|S| \cdot H \times W \times 1}$). The concatenation of slices instructs DDPM to learn the correlation of slices, yielding consistent generation results. By default, we concatenate along the color dimension. The comparison of these two concatenation ways can be found in the supplementary.

Following the procedures of DDPM, we first inject a noise $\epsilon \sim \mathcal{N}(0, 1)$ to $[S]$ to obtain $[\bar{S}]$. The noised slices $[\bar{S}]$ are then sent to a U-Net based denoising network F_{dn} that outputs the predicted noise conditioned on I :

$$\hat{\epsilon} = F_{dn}([\bar{S}], I). \quad (3)$$

In practice, the injection and prediction of noise ϵ involve an iteration step t , which we omit here for brevity. To better condition I for denoising diffusion, we employ a pre-trained VGG to extract multi-layer features from I and add them into the corresponding layers in the encoder of F_{dn} (see the blue part of Fig. 6). The loss function is the denoising loss: $\mathcal{L}_{gen} = \|\hat{\epsilon} - \epsilon\|_2^2$. After training, we can sample random noises to generate a set of slice images $\{\hat{S}_i^j\}$.

Compared to regression-based slicing, generation-based slicing can produce multiple plausible results rather than a single deterministic result, but consumes much longer inference time due to the nature of diffusion models.

3.3. Learning Implicit Field from Slices

We learn an implicit field from slice images, by projecting a 3D query point onto each slice image, extracting query-specific features, and then aggregating them to predict a signed distance. For regression-based slicing, we train the SDF prediction network on the predicted slices from the previous section thus the whole pipeline can be trained end-to-end. For generation-based slicing, we train the SDF prediction network on GT slice images since the generated slices may not match the GT signed distances.

Specifically, for a query point $q \in \mathbb{R}^3$, we project it to a location $\tilde{q} \in \mathbb{R}^2$ on the 2D image plane according to the camera parameters Θ . For slice \hat{S}_i^j (regression) or S_i^j (generation), we first extract its feature maps, then retrieve features from each feature map corresponding to location \tilde{q} , and finally concatenate them to obtain the image feature vector $V_i^j \in \mathbb{R}^{1 \times 1 \times N_c}$. The retrieval is accomplished via bi-linear interpolation over the features grids that encompass the location of \tilde{q} . We design a transformer network [60] F_t which learns to aggregate the feature vectors from all slices to predict the signed distance \hat{d} : $\hat{d} = F_t([q', V_1^X, \dots, V_{N_s}^Z])$, where $q' \in \mathbb{R}^{N_c}$ is linearly projected from q to be the same dimension with V_i^j . The optimization objective is to minimize the \mathcal{L}_1 distance between predicted and GT signed distance: $\mathcal{L}_{imp} = \|\hat{d} - d\|_1$. In regression-based slicing, \mathcal{L}_{imp} and \mathcal{L}_{reg} are jointly optimized.

As in DISN [67], we extract from only the input view I to obtain a feature vector V . For query points located around the occluded parts of the object (e.g., the query point in Fig. 3), the feature V is irrelevant to q since it is extracted from the front parts that block q , often leading to erroneous SDF predictions. In contrast, one or more of our slice images could reveal the occluded parts, yielding accurate features fed to the transformer to predict the correct SDF. After learning the implicit field, we apply the Marching Cube algorithm [33] to obtain triangle meshes.

3.4. Camera Pose Estimation

We follow the strategy in DISN [67] to estimate the camera poses. First, assuming a fixed set of intrinsic parameters, we only need to predict a translation vector $\Theta_t \in \mathbb{R}^3$ and a rotation matrix $\Theta_r \in \mathbb{R}^{3 \times 3}$. Then, a CNN, e.g., VGG-16 [50], is trained to estimate $\hat{\Theta}_t$ and $\hat{\Theta}_r$ from input view I . Afterwards, we sample a point cloud $P \in \mathbb{R}^{N_p \times 3}$ from the object along with its camera-aligned version P' . Finally, the loss function of the CNN is to align P' with P using $\hat{\Theta}_t$ and $\hat{\Theta}_r$, i.e., $\mathcal{L}_{cam} = \frac{1}{N_p} \|P - (P' \Theta_r + \Theta_t)\|_2^2$.

4. Results and Evaluation

Dataset. We utilize two well-known datasets, ShapeNet [5] and Objaverse [10], to train and test our network. For ShapeNet, our method is applied to Chairs and Air-

planes separately, using the train/validation/test split as in SHREC16 [48]. Objaverse is a challenging dataset, where many objects may not belong to a clear category. In our experiment, we randomly selected $\approx 40K$ shapes from Objaverse to keep resource and time costs manageable, while ensuring diversity of the collection. We follow a 95%-5% train-test split over the 40K selected shapes, ensuring that every training model we use is also from the training set of Zero-1-2-3 [30]. Note that Zero-1-2-3 used a train-test split of 99%-1% over the entire 800K models in Objaverse.

Rendering Slice Images. We employ the *bisect* function in Blender [1] to cut an object mesh into slices and then render them. By default, we keep internal cavities and holes produced by slicing. Since many 3D models in Objaverse are not in, or have no, canonical poses and feature random rotations, slicing them along the default X , Y , and Z directions is not suitable. To this end, we opted for *camera-aligned* slicing instead. Specifically, we set the camera position, introduce a random rotation to the object, position it in front of the camera, and utilize the camera’s axes for slicing. Under this setting, estimating the rotation of objects is unnecessary; see more details in the supplementary material.

Implementation Details. When preparing training data, we use the Vega FEM Library [51] for SDF computation, same as DISN [67] and AutoSDF [37]. For each object, the diagonal length of the object’s 3D bounding box is normalized to 1. We set H and W to 128, N_e to 128, N_s to 4, and N_p to 1,024. We train our model on *only one* Nvidia A40 GPU.

Evaluation Metrics. We use Chamfer \mathcal{L}_2 Distance (CD), F-score%1 (F1) [56], and Hausdorff Distance (HD) to evaluate reconstruction results. Besides these 3D metrics, we also evaluate using 2D metrics, PSNR, SSIM [63], and LPIPS [72], on 24 rendered views of reconstructed meshes.

4.1. Comparisons to State Of The Art (SOTA)

We compare our method with SOTA methods including DISN [67], AutoSDF [37], NeRF-From-Image [43], SSDNeRF [6], RealFusion [34], and One-2-3-45 [29]. Note that when showing visual results, we always present two views to remove any view bias towards the input.

ShapeNet. Figs. 7 and 8 demonstrate the effectiveness of our approach in reconstructing obscured sections of objects while precisely restoring visible ones. In contrast, DISN [67] can barely reconstruct the occluded parts. AutoSDF [37] and NeRF-From-Image [43] rely on latent features for reconstruction, which can yield rather clean results but they do not necessarily respect the input image, as the chair results from Fig. 7 evidently showed. SSDNeRF [6] tends to produce erroneous geometric details and the surfaces are not as smooth as from other methods, especially for the airplane examples in Fig. 8. Generally speaking, One-2-3-45 [29] yields rough and bulky geometries that clearly lack details. Additionally, our method outperforms

Datasets Metrics	ShapeNet Chairs			ShapeNet Airplanes		
	CD \downarrow	F1 \uparrow	HD \downarrow	CD \downarrow	F1 \uparrow	HD \downarrow
DISN [67]	3.25	2.34	7.84	3.48	2.81	7.62
Auto-SDF [37]	25.0	0.77	15.8	-	-	-
NeRF-Img [43]	19.5	1.27	13.4	16.0	1.41	14.0
One-2-3-45 [29]	12.5	1.01	11.7	23.4	0.75	12.6
Ours	2.80	2.43	6.75	2.99	2.83	6.45
DISN* [67]	3.02	2.42	7.86	3.19	2.96	7.56
SSDNeRF* [6]	8.98	2.06	10.9	10.4	2.00	10.4
Ours*	2.67	2.53	6.70	2.81	2.96	6.38

Table 1. Quantitative results of single-view 3D reconstruction on ShapeNet Chairs and Airplanes. ‘*’ denotes using ground-truth (GT) camera poses. Slice3D (ours) uses regression-based slicing.

Method	CD \downarrow	F1 \uparrow	HD \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
DISN [67]	28.2	1.47	23.0	21.9	8.79	15.2
One-2-3-45 [29]	35.4	0.87	29.8	19.8	8.55	19.8
Ours (G)	25.0	1.51	16.4	22.3	8.84	13.4
Ours (R)	16.6	1.53	13.4	22.6	8.88	12.8

Table 2. Quantitative results of single-view 3D reconstruction on the Objaverse dataset. ‘R’ and ‘G’ denote regression-based and generation-based slicing in Slice3D, respectively.

the other methods in the reconstruction of unconventional inputs, as shown by the two airplanes in Fig. 8.

Tab. 1 shows quantitative comparisons to the above methods on the two ShapeNet categories. When our method operates without using GT camera poses and relies on estimated poses, its performance sees only a marginal degradation. Notably, for Chair, our method with estimated poses even outperforms DISN and SSDNeRF with GT poses. It is also notable that One-2-3-45 tends to output dilated shapes. Although it falls behind DISN in quantitative results, it usually generates more complete shapes and more accurate topologies than DISN; see Figs. 7 and 8.

Objaverse. Since AutoSDF [37], NeRF-From-Image [43], and SSDNeRF [6] are all category-specific methods, they cannot be trained on Objaverse which consists of many 3D objects with no clear categories. Thus, we mainly compare our method with DISN [67], RealFusion [34], and One-2-3-45 [29]; see Fig. 9 for a qualitative comparison.

Note that RealFusion [34] and One-2-3-45 [29] both leverage a significantly larger volume of training data compared to our method, as they both rely on a pre-trained Stable Diffusion [46] model. Despite this disparity in the scale of the training sets, our method still outperforms these two methods, as well as DISN; see Fig. 9. Quantitatively, Tab. 2 demonstrates our method’s across-the-board advantage over DISN and One-2-3-45 (more so on Objaverse than on ShapeNet), in both 3D and 2D metrics, especially CD, HD, and LPIPS. These results also show that our method possesses the capacity to generalize across common shapes rather than being confined to specific categories.

Real-world Image Input. To further stress-test the methods,

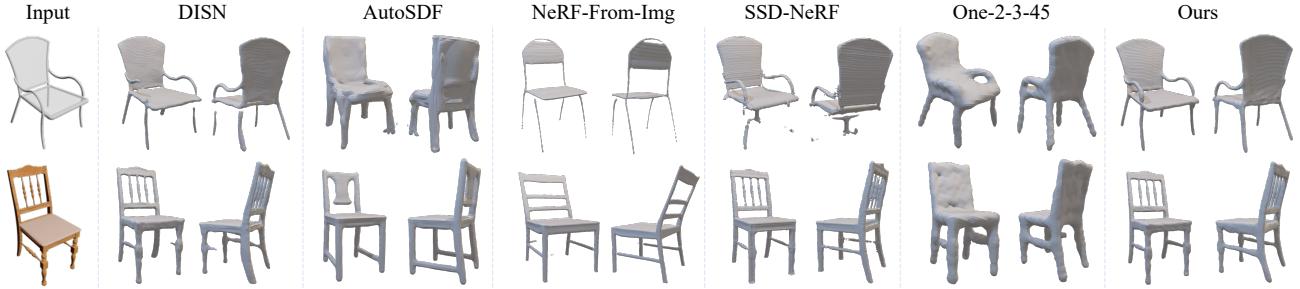


Figure 7. Visual comparison between single-view 3D reconstruction methods on two ShapeNet chairs. DISN and our method (based on regressive slicing) utilize the same estimated camera parameters. Two different views are displayed to remove view bias.

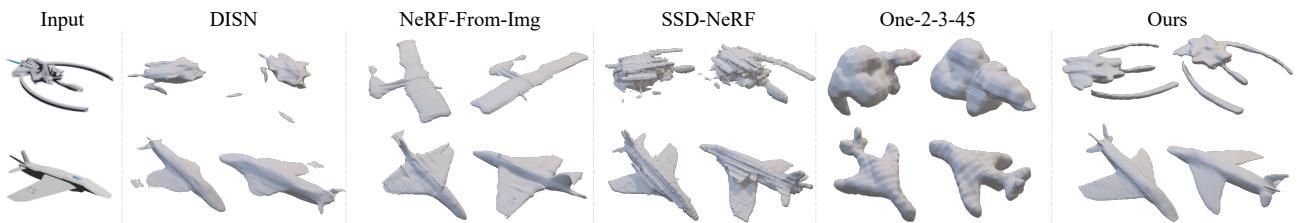


Figure 8. Visual comparison between single-view 3D reconstruction methods on two somewhat unconventional airplanes from ShapeNet.

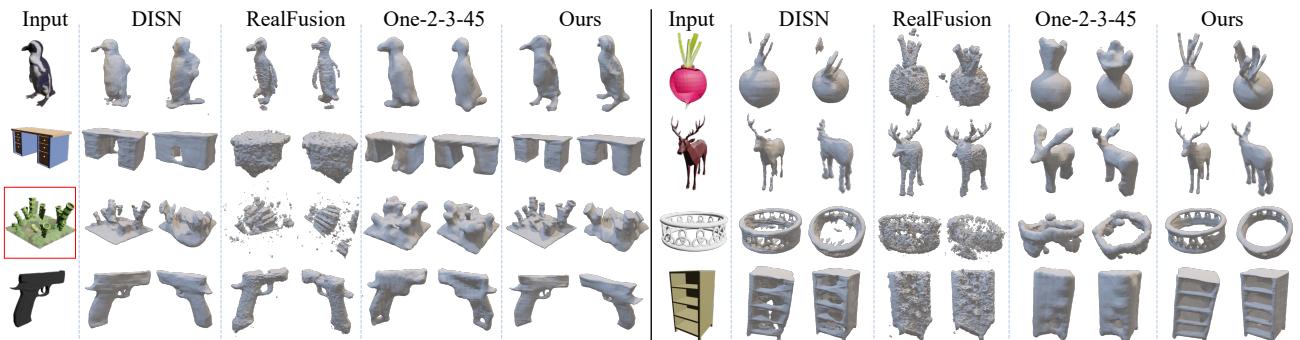


Figure 9. Visual comparison between cross-category reconstruction methods on Objaverse. Our method better recovers overall geometries, fine details, and is more faithful to the inputs. Except for the seaweed (squared), all the other models are in the *training* set of Zero-1-to-3.

we perform Google image search on object keywords (e.g., “chair”, “airplane”, etc.) and take the top-1 image with occlusions as input. In addition, we use some examples from Google Scanned Objects (GSO) [12] dataset. Fig. 10 shows that our method performs the best in recovering occluded parts while preserving fine details, e.g., the turbines of the airplane, the heel, and the opening of the boot.

Multi-slice vs. Multi-view Reconstruction. Fig. 11 compares between multi-slice and multi-view reconstructions. We demonstrate two different sets of multi-slices and multi-views generated from our method and One-2-3-45 [29]. Our generated slice images depict two distinct configurations of chair legs, maintaining high consistency across slices. In contrast, the multiple views synthesized in One-2-3-45 do encounter inconsistency issues, compromising an accurate 3D reconstruction. We show three slices in each set and

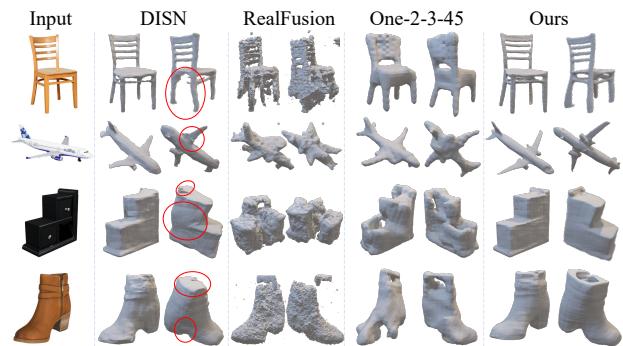


Figure 10. On real-world image inputs. Top two from Google image search and the others from GSO [12]. Zoom in for details.

more can be found in the supplementary material.

Training and Inference Costs. Unlike RealFusion [34] and

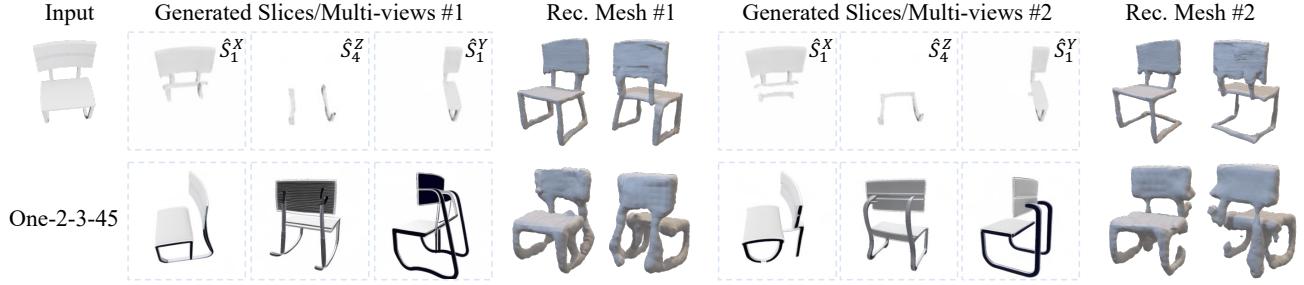


Figure 11. Multi-slice vs. multi-view reconstructions amid ambiguities in the chair legs. Both One-2-3-45 (bottom) and Slice3D (top) can produce multiple results. Our results are both plausible from consistent slices, while One-2-3-45 suffers from multi-view inconsistencies.

One-2-3-45 [29], our method does not depend on a large pre-trained model [46]. NeRF-based methods tend to have long inference times, as many of them require per-test optimization. Our method has much faster inference times: **10 or 20 seconds** with slice regression or generation, compared to 90 minutes for RealFusion and 45 seconds for One-2-3-45 and SSDNeRF, while delivering superior results. With similar inference times to DISN, Auto-SDF, and NeRF-From-Img, our method has better reconstruction quality. See supplementary material for more details.

4.2. Ablation Study

Generative vs. Regressive Slicing. As shown in the first row of Fig. 11, for generation-based slicing, we can produce multiple plausible results for a given input view. Tab. 2 shows that regressive slicing outperforms generative slicing when compared to GT. However, it is notable that the 3D outputs may differ from GT when input views have ambiguous occluded parts. Regressive training struggles to handle such one-to-many mappings, leading to generation outperforming regression in reconstructing occluded parts.

Number of Slices. We test the performance of our method with regression-based slicing by varying N_s between 2, 4, and 8, on ShapeNet Chairs. From Tab. 3, it is evident that larger N_s leads to better reconstruction quality. However, $N_s = 8$ will dramatically increase the computational cost of our method since its complexity is $O(N_s^2)$, as pair-wised attentions are performed on the slices in the transformer. As a trade-off between reconstruction quality and computational complexity, we set $N_s = 4$ by default.

Tab. 3 also covers several other ablation studies.

Learned v.s. One-hot Slice Indicator. Learned indicator codes (the default setting) achieve slightly better performance than using one-hot codes to indicate slice locations. This can be seen by comparing numbers from the first half row of Tab. 3 to those in the last row of Tab. 1.

Slices v.s. Depth Images. Since our slice images offer layered information, it is logical to perform a comparison against depth images as supplementary information. We compared the performance of DISN when concatenating

Setting	CD↓	F1↑	HD↓	Setting	CD↓	F1↑	HD↓
One-hot ind.	2.68	2.53	6.82	only X	2.88	2.39	7.74
DISN+Depth	2.99	2.43	7.81	only Y	2.75	2.45	6.84
$N_s = 2$	2.90	2.46	7.31	only Z	3.08	2.38	7.60
$N_s = 4$	2.67	2.53	6.70	Filled	2.73	2.50	7.37
$N_s = 8$	2.56	2.56	6.93	w/ textures	2.82	2.50	7.70

Table 3. Ablation studies on ShapeNet Chairs.

estimated depth images with the input view. Results are shown in the “DISN+Depth” half row, which indicates that depth images can only provide marginal improvement.

Three Axes v.s. One Axis. We also investigated the option of slicing along only one axis. Not surprisingly, Tab. 3 shows that this underperforms against using three axes.

Solid v.s Hollow Slices. Tab. 3 also demonstrates that not filling the holes improves the performance as more information about the inner/outer surfaces is revealed this way.

Using Textured Slices. We can optionally keep object textures in the slice images. Experimental results show that removing the textures in the slice images can achieve better reconstruction. The reasons could be: 1) learning to produce slice images without textures could help the model better distinguish geometries from textures; 2) for some 3D models the color of inner and outer surfaces could be different which may complicate the learning process.

5. Discussion and Future Work

Our multi-slice idea represents a significant departure from conventional approaches to single-view 3D reconstruction. Through conceptual illustrations (Fig. 5) and extensive experiments, we demonstrate its compelling properties in recovering occluded object structures. Note again that all the Slice3D results on Objaverse [10] were obtained by networks trained on a very *small subset* (40K or 5%) of the 800K 3D models therein, whereas One-2-3-45 [29] utilizes Zero-1-to-3 [30] for multi-view synthesis and the latter was fine-tuned on the entirety of the 800K. Yet our results (Table 2) show that Slice3D outperforms One-2-3-45 across the board, suggesting that our model generalizes better, likely owing to its superior occlusion-revealing abilities.

Our current implementation only uses four slices per di-

rection, which is quite coarse. Since there is no camera rotation, parts of an object far away from the camera are in smaller scale, e.g., the rear leg of the chair in Fig. 10 and of the buck in Fig. 9. This can hinder our network’s ability to extract detailed features. More refined and gradual slicing over those parts may be needed to improve matters.

The foremost limitation of slicing however, is that realistically, it can only be executed at scale on *digital* 3D models, not physical ones. On the other hand, the multi-view approach can be applied to physical 3D objects by capturing photos only, without ever needing 3D supervision, as in RealFusion [34]. However, this is not without a high cost: 90 minutes by RealFusion vs. 45 seconds by One-2-3-45 (requiring 3D fine-tuning for multi-view synthesis) and less than 20 seconds by Slice3D. We envision synthesizing 3D models, not many but carefully crafted few, for the multi-slice approach as an intriguing direction to explore.

Slice3D: Multi-Slice, Occlusion-Revealing, Single View 3D Reconstruction

Supplementary Material

6. Rendering Details

Given a 3D model, we first normalize the length of its body diagonal to 1. Then, we put the 3D model in front of the camera with a distance of 1.2, i.e., the camera is pointed to the centre of the bounding box of the object with a distance of 1.2. Next, we randomly give a rotation, scaling and translation to the object. The rotation is done by introducing a pair of elevation and azimuth randomly sampled from the ranges of $[-10^\circ, 40^\circ]$ and $[0^\circ, 360^\circ]$, respectively. When introducing scaling and translation, we restrict the 3D model to be always inside $[-0.5, 0.5]^3$, from where the query points are sampled in both training and inference.

This rendering rule is implemented in both ShapeNet [5] and Objaverse [10]. Note that positions of an object and a camera are relative. If the object is assumed to be fixed, we can tailor the rendering specifics by adjusting the camera through rotation and translation. In Section 3.4 of our main paper, we assume the object is fixed and then estimate rotations and translations of the camera.

7. Visualization of Hole-Filled Slices

As our ablation studies already showed (Tab. 3), operating on slice images *without* the holes filled provides more information about both the inside and outside structures of an object, yielding better reconstruction quality. In Fig. 12, we visually compared slice images with (left) vs. without holes filled.

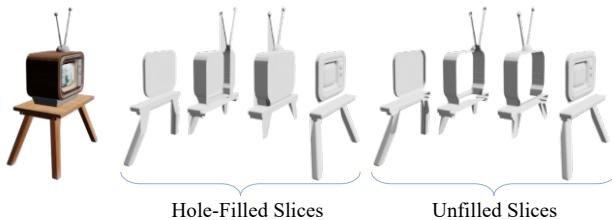


Figure 12. Slicing with and without filling holes. The demonstrated images are sliced along axis X of the object.

8. Details on Two Slicing Directions

As mentioned in Section 4, we utilize camera-aligned slicing for Objaverse dataset where many 3D models feature arbitrary orientations, i.e., they do not have clear canonical poses. In other cases, e.g., for ShapeNet shapes, we rely on the canonical poses of the objects and perform slicing along the X , Y , and Z directions with respect to those canonical poses. Fig. 13 illustrates and contrasts these two choices for

the slicing directions, using a chair example from ShapeNet dataset.

The 3D models in ShapeNet possess canonical poses (e.g., chairs consistently have a front orientation) and they are aligned along the default X , Y , and Z axes. Consequently, objects are always sliced along their default X , Y , and Z axes regardless of the rendered views.

For 3D models in Objaverse, we use camera-aligned slicing. As described in Sec. 6, the camera is fixed and an object will be randomly rotated, scaled, and translated. We slice the bounding boxes of the object along the axes in the camera world.

9. Whether to Estimate Camera Poses

For our method, camera pose estimation is necessary only when we want to output 3D shapes in their canonical poses. In ShapeNet, we follow most existing methods [6, 37, 43, 67] to produce 3D shapes in their canonical poses, which involves camera pose estimation. However, in Objaverse, where many objects do not have canonical poses and exhibit random orientations, we avoid estimating the rotations and translations of objects, and reconstruct the 3D shapes as they are in the camera world.

10. Training Cost and Inference Speed

Tab. 4 makes a detailed comparison of training cost and inference speed among single-view- reconstruction methods. Our method does not rely on big pre-trained model like Stable Diffusion [46]. Compared to multi-view based methods, our method runs much faster in the process of producing 3D meshes from slice images because we employ neural signed distance filed instead of NeRFs whose optimization is time-consuming.

11. Concatenation of Slice Images in Diffusion

We perform DDPM [21] on the entirety of slice images that can be stacked either on the color dimension or a spatial dimension. Note that concatenating along a spatial dimension significantly increases complexity because of the self attentions operated on the spatial dimensions in a diffusion network. By default, we stack them on the color channel to reduce the training and inference time. Tab. 5 provides a quantitative comparison of these two concatenation methods, revealing that concatenating along a spatial dimension achieves better performance than the color dimension. This outcome is logical as the former can model the spatial correspondence of different slice images throughout the dif-

Method	Pre-trained model	Training data	Training GPU & time	Infer. speed
DISN [67]	VGG [50]	SPN-Chair	1 × A40 for 1 day	<5s
AutoSDF [37]	ResNet-18 [20]	SPN-Chair	unknown	<10s
NeRF-Img [43]	VGG-16 [50]	SPN-Chair	4 × A100	<30s
SSDNeRF [6]	N/A	SPN-Chair	2 × 3090 for 6 days	45s-1min
Ours	VGG-16 [50]	SPN-Chair	1 × A40 for 2 days	<10s(R)/<20s(G)
Real-Fusion [34]	SD [46]	Per-case opt.	N/A	90min
One-2-3-45 [29]	Zero-1-to-3 [30]	Objv-40k*	2 × A10 for 6 days	≈45s
Ours	VGG-16 [50]	Objv-40k	1 × A40 for 3 days	<10s(R)/<20s(G)

Table 4. Training cost and inference speed of single-view-reconstruction methods. ‘SPN-Chair’ denotes ShapeNet Chair dataset. ‘opt.’ denotes optimization. ‘Objv-40k’ denotes a subset from Objaverse with around 40k 3D models. ‘R’ and ‘G’ denote regression-based and generation-based slicing, respectively. Note that the first stage of One-2-3-45 (i.e., Zero-1-to-3) is trained with nearly the whole Objaverse dataset. In the second stage, it is trained with a subset in the scale of 40k 3D models. The inference speed is tested on a single Nvidia-A40-GPU for all methods.

Method	CD↓	F1↑	HD↓
Ours (G-C)	25.0	1.51	16.4
Ours (G-S)	20.0	1.51	14.1

Table 5. Quantitative results of single-view 3D reconstruction on the Objaverse dataset. ‘G-C’ and ‘G-S’ denote concatenating the slice images along the color and a spatial dimension, respectively.

fusion network. Given sufficient computational resources, prioritizing concatenation along a spatial dimension is recommended.

12. View Inconsistency Problem

As mentioned in our main paper, recent methods [31, 32, 44, 54, 73] aim to enhance the consistency of synthesized views by performing spatial attention across different views. In Fig. 14, a comparison is made between Slice3D and Sync-Dreamer [31] from these works. The findings indicate that despite the utilization of expensive spatial attentions, the challenge of maintaining view consistency persists. This further substantiates the advantages of employing multi-slices over multi-views.

13. Visualization of Predicted Slice Images

The predicted slice images for the examples in Fig. 7 and 8 can be found in Fig. 15 and Fig. 16, respectively. Notably, our slice3D can produce slice images with a high level of consistency.

14. More Visual Results

More visual results and comparisons are provided in Fig. 17 and 18 for ShapeNet, Fig. 19 for Objaverse, and Fig. 20 for Google Scanned Objects (GSO) [12]. As apparent, our results respect the geometric details better than the other techniques while they do not suffer from unwanted artifacts or noise. Also compared to other techniques such as AutoSDF, it better respects the input image and does not re-

trieve a model that might look clean and noise-free but it is far from the input image (e.g., Fig. 17; first two rows).

15. Image Resolution

Due to limited computing resources, the resolution of our input images and slice images is set to only 128. We plan to increase the resolution to 256 or 512 in the future and produce 3D meshes with better quality and details.

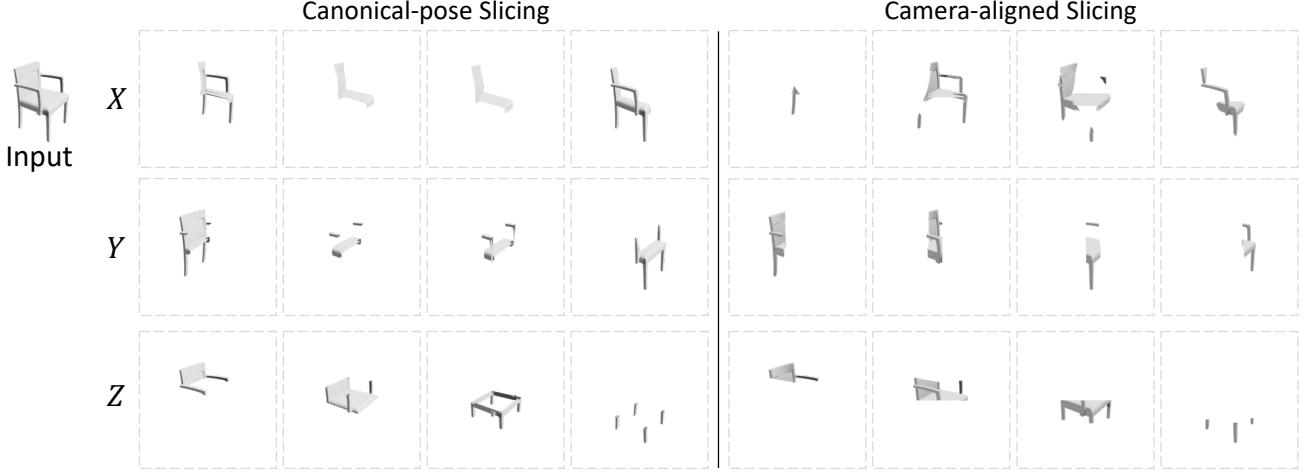


Figure 13. Canonical-pose slicing v.s. camera-aligned slicing. In canonical-pose slicing, the slicing directions are determined by the canonical pose of the object. In camera-aligned slicing, the slicing directions are determined by the orientation of the camera.

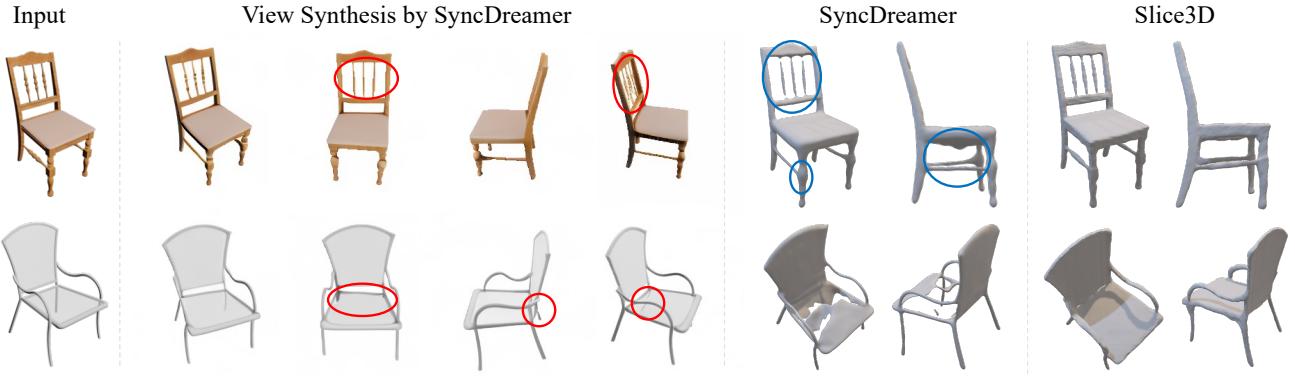


Figure 14. Visual comparison against SyncDreamer [31], which aims to enhance the consistency by performing spatial attention across different views. The red circles highlight the inconsistency across different synthesized views. For example, the pillars in the first chair and the rear legs in the second chair. The blue circles highlight the artifacts in the 3D mesh.



Figure 15. Visualization of the predicted (regressed) slice images from the input views.



Figure 16. Multi-slice generation results and the meshes resulted from them.

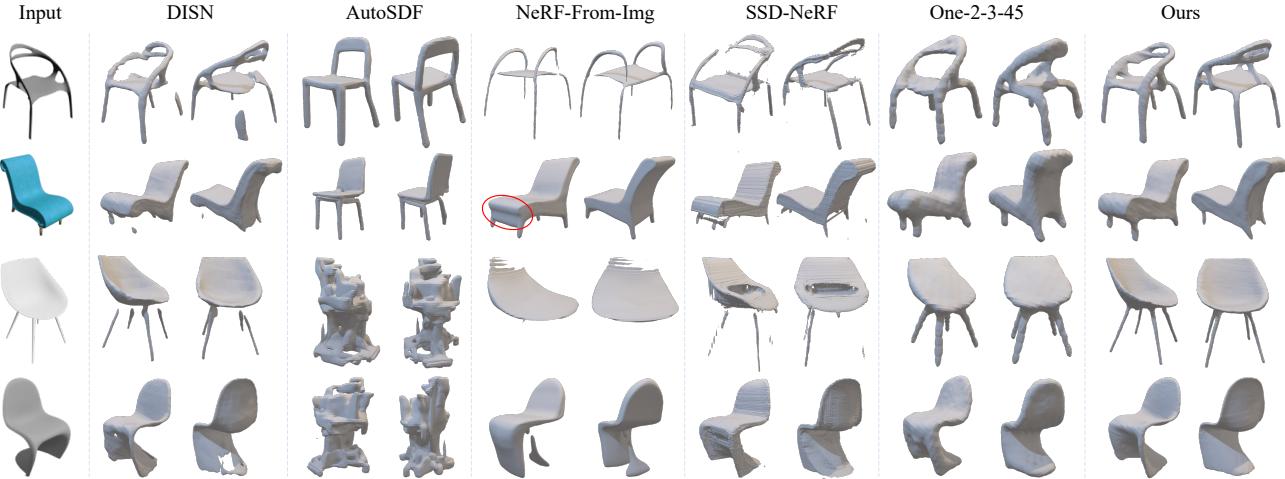


Figure 17. More visual comparison between single-view 3D reconstruction methods on ShapeNet chairs. DISN and our method (based on regressive slicing) utilize the same estimated camera parameters. Two different views are displayed to remove view bias.

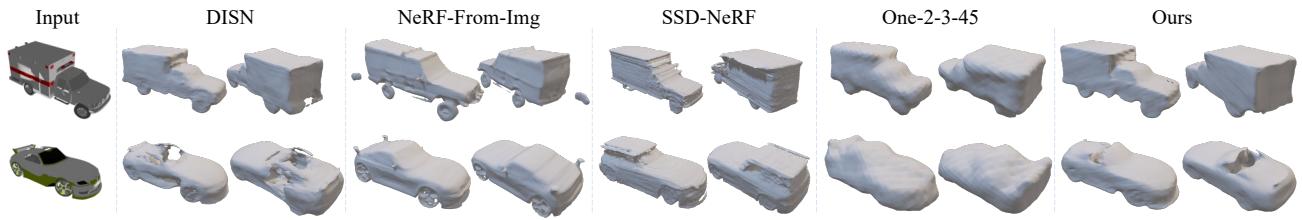


Figure 18. Visual comparison between single-view 3D reconstruction methods on two ShapeNet cars.

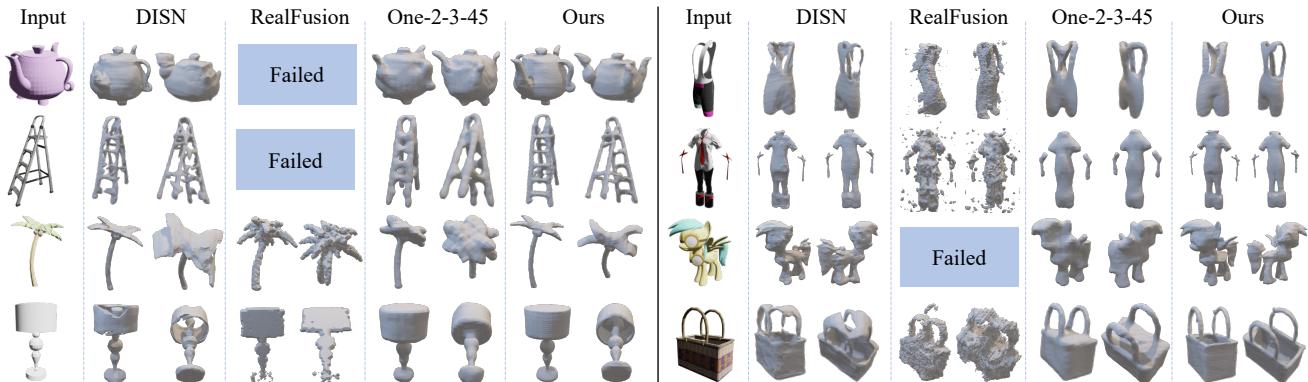


Figure 19. More results on Objaverse. “Failed” denotes no meaningful results after several optimizations of RealFusion.

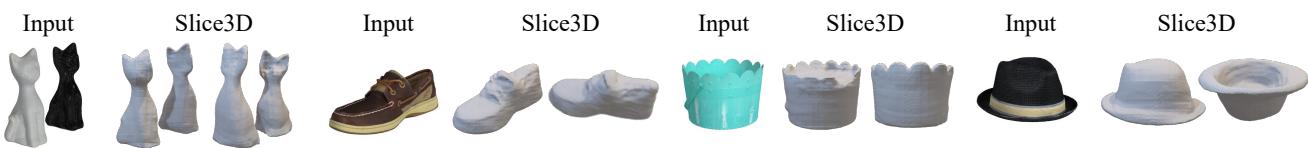


Figure 20. More results on GSO [12] dataset.

References

- [1] Bisect - blender manual. <https://docs.blender.org/manual/en/2.80/modeling/meshes/editing/subdividing/bisect.html>. Accessed: 2023-11-15. 6
- [2] Kfir Aberman, Oren Katzir, Qiang Zhou, Zegang Luo, Andrei Sharf, Chen Greif, Baoquan Chen, and Daniel Cohen-Or. Dip transform for 3D shape reconstruction. *ACM Transactions on Graphics*, 36(4):Article 79, 2017. 4
- [3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 3
- [4] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 4
- [5] 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. 3, 5, 10
- [6] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion NeRF: A unified approach to 3d generation and reconstruction. *arXiv preprint arXiv:2304.06714*, 2023. 2, 3, 4, 6, 10, 11
- [7] Wenzheng Chen, Jun Gao, Huan Ling, Edward J. Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in Neural Information Processing Systems*, 2019. 2
- [8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948, 2019. 3
- [9] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *CVPR*, 2020. 1
- [10] 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*, pages 13142–13153, 2023. 3, 4, 5, 8, 10
- [11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3
- [12] 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. 7, 11, 13
- [13] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019. 4
- [14] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 3
- [15] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022. 3
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3
- [17] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. In *CVPR*, 2018. 1
- [18] 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 *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 4
- [19] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 4
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 11
- [21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 5, 10
- [22] Chenbo Jiang, Jie Yang, Shuai He, Yu-Kun Lai, and Lin Gao. Neuralslice: Neural 3d triangle mesh reconstruction via slicing 4d tetrahedral meshes. 2023. 4
- [23] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3
- [24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 3
- [25] Shay Kels and Nira Dyn. Reconstruction of 3d objects from 2d cross-sections with the 4-point subdivision scheme adapted to sets. *Computers & Graphics*, 35:741–746, 2011. 4
- [26] Manyi Li and Hao Zhang. D²IM-Net: Learning detail disentangled implicit fields from single images. In *CVPR*, 2021. 1, 3
- [27] Wallace Lira, Johannes Merz, Daniel Ritchies, Daniel Cohen-Or, and Hao Zhang. GANHopper: Multi-hop gan for unsupervised image-to-image translation. In *ECCV*, 2020. 3
- [28] Lu Liu, Chandrajit Bajaj, Joseph O. Deasy, Daniel A. Low, and Tao Ju. Surface reconstruction from non-parallel curve networks. *Computer Graphics Forum*, 27:155–163, 2008. 4
- [29] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh

- in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 1, 2, 3, 4, 6, 7, 8, 11
- [30] 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. 1, 3, 4, 6, 8, 11
- [31] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image, 2023. 4, 11, 12
- [32] 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. *arXiv preprint arXiv:2310.15008*, 2023. 4, 11
- [33] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 5
- [34] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, pages 8446–8455, 2023. 3, 4, 6, 7, 9, 11
- [35] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 3
- [36] 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. 3
- [37] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *CVPR*, pages 306–315, 2022. 3, 6, 10, 11
- [38] 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. 3
- [39] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 2
- [40] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, pages 5589–5599, 2021. 3
- [41] Azimkhon Ostonov. Cut-and-approximate: 3d shape reconstruction from planar cross-sections with deep reinforcement learning. *arXiv preprint arXiv:2210.12509*, 2022. 4
- [42] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 3
- [43] Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In *CVPR*, pages 4391–4401, 2023. 2, 3, 4, 6, 10, 11
- [44] 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. *arXiv preprint arXiv:2306.17843*, 2023. 4, 11
- [45] Stephan R. Richter and Stefan Roth. Matryoshka networks: Predicting 3d geometry via nested shape layers. In *CVPR*, 2018. 1
- [46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3, 4, 6, 8, 10, 11
- [47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4
- [48] Manolis Savva, Fisher Yu, Hao Su, M Aono, B Chen, D Cohen-Or, W Deng, Hang Su, Song Bai, Xiang Bai, et al. Shrec16 track: largescale 3d shape retrieval from shapenet core55. In *Proceedings of the eurographics workshop on 3D object retrieval*, 2016. 6
- [49] Haim Sawdayee, Amir Vaxman, and Amit H. Bermano. Orex: Object reconstruction from planar cross-sections using neural fields. In *CVPR*, pages 20854–20862, 2023. 4
- [50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5, 11
- [51] Fun Shing Sin, Daniel Schroeder, and Jernej Barbić. Vega: non-linear fem deformable object simulator. In *Computer Graphics Forum*, pages 36–48. Wiley Online Library, 2013. 6
- [52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3
- [53] 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. 4
- [54] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023. 4, 11
- [55] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, 2017. 1
- [56] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, pages 3405–3414, 2019. 6
- [57] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 2

- [58] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 4
- [59] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3
- [60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5
- [61] Nanyang Wang, Yinda Zhang, Zhiwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, pages 52–67, 2018. 3
- [62] 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. 3
- [63] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [64] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 4
- [65] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Comput. Graph. Forum*, 2022. 3
- [66] 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*, pages 4479–4489, 2023. 4
- [67] 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. 1, 3, 5, 6, 10, 11
- [68] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *CVPR*, pages 6239–6249, 2022. 3
- [69] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 3
- [70] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *arXiv preprint arXiv:2301.11445*, 2023. 3
- [71] Mingfang Zhang, Jinglu Wang, Xiao Li, Yifei Huang, Yoichi Sato, and Yan Lu. Structural multiplane image: Bridging neural view synthesis and 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16707–16716, 2023. 4
- [72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 4, 6
- [73] Minda Zhao, Chaoyi Zhao, Xinyue Liang, Lincheng Li, Zeng Zhao, Zhipeng Hu, Changjie Fan, and Xin Yu. Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion prior. *arXiv preprint arXiv:2308.13223*, 2023. 4, 11
- [74] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 4