

# OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation

Tong Wu<sup>2</sup>, Jiarui Zhang<sup>1,3</sup>, Xiao Fu<sup>1</sup>, Yuxin Wang<sup>1,4</sup>, Jiawei Ren<sup>5</sup>, Liang Pan<sup>5</sup>, Wayne Wu<sup>1</sup>, Lei Yang<sup>1,3</sup>, Jiaqi Wang<sup>1</sup>, Chen Qian<sup>1</sup>, Dahua Lin<sup>1,2✉</sup>, Ziwei Liu<sup>5✉</sup>

<sup>1</sup>Shanghai Artificial Intelligence Laboratory, <sup>2</sup>The Chinese University of Hong Kong, <sup>3</sup>SenseTime Research, <sup>4</sup>Hong Kong University of Science and Technology, <sup>5</sup>S-Lab, Nanyang Technological University

{wt020, dlin}@ie.cuhk.edu.hk, zjr954@163.com, ywangom@connect.ust.hk,

wuwenyan0503@gmail.com, {fuxiao, wangjiaqi, qianchen}@pjlab.org.cn,

yanglei@sensetime.com, jiawei011@e.ntu.edu.sg, {liang.pan, ziwei.liu}@ntu.edu.sg



Figure 1. **OmniObject3D** is a large vocabulary 3D object dataset with massive high-quality real-scanned 3D objects and rich annotations. It supports various research topics, e.g., perception, novel view synthesis, neural surface reconstruction, and 3D generation.

## Abstract

Recent advances in modeling 3D objects mostly rely on synthetic datasets due to the lack of large-scale real-scanned 3D databases. To facilitate the development of 3D perception, reconstruction, and generation in the real world, we propose **OmniObject3D**, a large vocabulary 3D object dataset with massive high-quality real-scanned 3D objects. OmniObject3D has several appealing properties: **1) Large Vocabulary:** It comprises 6,000 scanned objects in 190 daily categories, sharing common classes with pop-

ular 2D datasets (e.g., ImageNet and LVIS), benefiting the pursuit of generalizable 3D representations. **2) Rich Annotations:** Each 3D object is captured with both 2D and 3D sensors, providing textured meshes, point clouds, multi-view rendered images, and multiple real-captured videos. **3) Realistic Scans:** The professional scanners support high-quality object scans with precise shapes and realistic appearances. With the vast exploration space offered by OmniObject3D, we carefully set up four evaluation tracks: **a)** robust 3D perception, **b)** novel-view synthesis, **c)** neural surface reconstruction, and **d)** 3D object generation. Extensive studies are performed on these four benchmarks, revealing

✉Corresponding authors. <https://omniobject3d.github.io/>

*new observations, challenges, and opportunities for future research in realistic 3D vision.*

## 1. Introduction

Sensing, understanding, and synthesizing realistic 3D objects is a long-standing problem in computer vision, with rapid progress emerging in recent years. However, a majority of the technical approaches rely on unrealistic synthetic datasets [9, 26, 96] due to the absence of a large-scale real-world 3D object database. However, the appearance and distribution gaps between synthetic and real data cannot be compensated for trivially, hindering their real-life applications. Therefore, it is imperative to equip the community with a large-scale and high-quality 3D object dataset from the real world, which can facilitate a variety of 3D vision tasks and downstream applications.

Recent advances partially fulfill the requirements while still being unsatisfactory. As shown in Table 1, CO3D [74] contains 19k videos capturing objects from 50 MS-COCO categories, while only 20% of the videos are annotated with accurate point clouds reconstructed by COLMAP [77]. Moreover, they do not provide textured meshes. GSO [21] has 1k scanned objects while covering only 17 household classes. AKB-48 [49] focuses on robotics manipulation with 2k articulated object scans in 48 categories, but the focus on articulation leads to a relatively narrow semantic distribution, failing to support general 3D object research.

To boost the research on general 3D object understanding and modeling, we present **OmniObject3D**: a large-vocabulary 3D object dataset with massive high-quality, real-scanned 3D objects. Our dataset has several appealing properties: **1) Large Vocabulary:** It contains 6,000 high-quality textured meshes scanned from real-world objects, which, to the best of our knowledge, is the largest among real-world 3D object datasets with accurate 3D meshes. It comprises 190 daily categories, sharing common classes with popular 2D and 3D datasets (*e.g.*, ImageNet [19], LVIS [33], and ShapeNet [9]), incorporating most daily object realms (See Figure 1 and Figure 2). **2) Rich Annotations:** Each 3D object is captured with both 2D and 3D sensors, providing textured 3D meshes, sampled point clouds, posed multi-view images rendered by Blender [17], and real-captured video frames with foreground masks and COLMAP camera poses. **3) Realistic Scans:** The object scans are of high fidelity thanks to the professional scanners, bearing precise shapes with geometric details and realistic appearance with high-frequency textures.

Taking advantage of the vast exploration space offered by OmniObject3D, we carefully set up four evaluation tracks: **a)** robust 3D perception, **b)** novel-view synthesis, **c)** neural surface reconstruction, and **d)** 3D object generation. Extensive studies are performed on these benchmarks: First, the high-quality, real-world point clouds in OmniOb-

Table 1. **A comparison between OmniObject3D and other commonly-used 3D object datasets.**  $R^{lvis}$  denotes the ratio of the 1.2k LVIS [33] categories being covered.

Dataset	Real	Full Mesh	Video	# Objs	# Cats	$R^{lvis}$ (%)
ShapeNet [9]		✓		51k	55	4.1
ModelNet [96]		✓		12k	40	2.4
3D-Future [26]		✓		16k	34	1.3
ABO [16]		✓		8k	63	3.5
Toys4K [83]		✓		4k	105	7.7
CO3D V1 / V2 [74]	✓		✓	19 / 40k	50	4.2
DTU [1]	✓	✓		124	NA	0
ScanObjectNN [87]	✓			15k	15	1.3
GSO [21]	✓	✓		1k	17	0.9
AKB-48 [49]	✓	✓		2k	48	1.8
<b>Ours</b>	✓	✓	✓	6k	190	10.8

ject3D allow us to perform *robust 3D perception* analysis on both out-of-distribution (OOD) styles and corruptions, two major challenges in point cloud OOD generalization. Furthermore, we provide massive 3D models with multi-view images and precise 3D meshes for *novel-view synthesis* and *neural surface reconstruction*. The broad diversity in shapes and textures offers a comprehensive training and evaluation source for both scene-specific and generalizable algorithms. Finally, we equip the community with a database for large vocabulary and realistic *3D object generation*, which pushes the boundary of existing state-of-the-art generation methods to real-world 3D objects. The four benchmarks reveal new observations, challenges, and opportunities for future research in realistic 3D vision.

## 2. Related Works

**3D Object Datasets.** The acquisition of a large-scale realistic 3D database is usually expensive and challenging. Many widely-used 3D datasets prefer to collect synthetic CAD models from online repositories [9, 83, 96], for example, ShapeNet [9] has 51,300 3D CAD models in 55 categories, and ModelNet40 [96] consists of 12,311 models in 40 categories. Recent works, *e.g.*, 3D-FUTURE [26] and ABO [16], introduce high-quality CAD models with rich geometric details and informative textures. However, due to the inevitable gap between synthetic and real objects, the community is still eager for a large-scale 3D object dataset in the real world. DTU [1] and BlendedMVS [102] are photo-realistic datasets designed for multi-view stereo benchmarks, while they are small in scale and lack category annotations. ScanObjectNN [87] is a real-world point cloud object dataset based on scanned indoor scenes, containing around 15,000 objects with colored point clouds in 15 categories. However, the point clouds are incomplete and noisy, and multiple objects usually co-exist in one scene. GSO [21] has 1,030 scanned objects with fine geometries and textures in 17 household items, and AKB-48 [49] focuses on robotics manipulation with 2,037 articulated object models in 48 articulated object categories. However,



Figure 2. **Semantic distribution of the OmniObject3D dataset.** It covers 190 daily categories with a long-tailed distribution, sharing common classes with popular 2D and 3D datasets.

the relatively narrow semantic scope of GSO and AKB-48 hinders their applications for more general 3D research. CO3D v1 [74] contains 19,000 object-centric videos, while only 20% of them are annotated with accurate point clouds reconstructed by COLMAP [77], and they do not provide meshes or textures. In contrast, the proposed OmniObject3D dataset comprises 6,000 3D objects scanned by professional devices with meshes, textures, and multi-view photos in 190 categories, fulfilling the requirements of a wide range of research objectives. A detailed comparison is presented in Table 1.

**Robust 3D Perception.** Robustness to out-of-distribution (OOD) data is important in point cloud perception. Two main challenges include OOD styles (*e.g.*, differences between CAD models and real-world objects) and OOD corruptions (*e.g.*, random point jittering or missing due to sensory inaccuracy). A branch of works [13, 45, 71, 92] studies the OOD corruptions and proposes standard corruption test suites [75, 85], while they fail to take account of OOD styles. Another branch of works [3, 74] evaluates the sim-to-real domain gap by training models on clean synthetic datasets [96] and testing them on noisy real-world test sets [87], while OOD styles and corruptions cannot be disentangled under this setting for an independent analysis. In this work, we leverage high-quality, real-world point clouds from OmniObject3D to systematically measure the robustness against the OOD style and OOD corruptions, providing the first benchmark for fine-grained evaluation of the point cloud perception robustness.

**Neural Radiance Field and Neural Surface Reconstruction.** Neural radiance field (NeRF) [60] represents a scene with a fully-connected deep network (MLPs), which takes in hundreds of sampled points along each camera ray and outputs the predicted color and density. We can synthesize the image of an unseen view from a trained model via volume rendering. Inspired by the success of NeRF, massive follow-up efforts have been made to improve its quality [5, 6, 59, 89] and efficiency [10, 25, 62, 84]. A branch of works [11, 36, 52, 74, 91, 105] has also explored the generalization ability of NeRF-based frameworks, where they

aim to learn priors from deep image features across multiple scenes. Beyond novel view synthesis, another trend of approaches [18, 66, 90, 95, 103] proposes to combine neural radiance field with implicit surface representations like Signed Distance Function (SDF), and they achieve accurate and mask-free surface reconstruction from multi-view images. Since dense camera views of scenes are sometimes unavailable, recent advances explore surface reconstruction from sparse views. They address the problem by exploiting generalizable priors cross scenes for a generic surface prediction [54] or taking advantage of the estimated geometry cues estimated by pre-trained networks [106]. OmniObject3D can serve as a large-scale benchmark with realistic photos and accurate meshes for both training and evaluation. The high diversity in shape and appearance offers an opportunity for pursuing more generalizable and robust novel view synthesis and surface reconstruction methods.

**3D Object Generation.** Early approaches [27, 35, 55, 82, 94] extend 2D generation frameworks to 3D voxels with a high computational cost. Some others adopt different 3D data formulations, *e.g.*, point cloud [2, 61, 101, 109], octree [39], and implicit representations [14, 57]. However, it is non-trivial to generate complex and textured surfaces. Recent advances [12, 28, 70] explore the generation for textured 3D meshes, where GET3D [29] is a state-of-the-art approach that generates diverse meshes with rich geometry and textures in two branches. It is a promising but challenging task to train generative models on a large vocabulary and realistic dataset. We evaluate GET3D on our dataset and reveal several challenges and future opportunities.

In supplementary materials, we present more detailed discussions on related works for different tracks.

### 3. The OmniObject3D Dataset

In this section, we describe the data collection, processing, and annotation pipeline of OmniObject3D. We also introduce the statistics and distribution of it.

### 3.1. Data Collection, Processing, and Annotation

**Category List Definition.** In order to collect a large amount of 3D objects that are both commonly-distributed and highly diverse, we first pre-define a category list according to several popular 2D and 3D datasets [9, 19, 33, 47, 48, 80, 96]. We cover most of the categories that lie within the application scope of the scanners and also dynamically expand the list with reasonable new classes that are absent from the current list during collection. We end up with 190 widely-spread categories, which ensures a library with rich texture, geometry, and semantic information.

**Object Collection Pipeline.** We then collect a variety of objects from each category and use professional 3D scanners to obtain high-resolution textured meshes. Specifically, we use the Shining 3D scanner<sup>1</sup> and Artec Eva 3D scanner<sup>2</sup> for objects in different scales. The scanning time varies with the properties of the object: it takes around 15 minutes to scan a small rigid object with a simple geometry (*e.g.*, an apple, a toy), while it takes up to an hour to obtain a qualified 3D scan for non-rigid, complex, or large objects (*e.g.*, a bed, a kite). For around 10% of the objects, we conduct common manipulations (*e.g.*, taking a bite, cutting in pieces) to conform the natural instincts of them. The 3D scans can faithfully retain the real-world scale of each object, but their poses are not strictly aligned. We thus pre-define a canonical pose for each category and manually align the objects within a category. We then check the quality of each scan, and around 83% high-quality ones out of the total collection are finally reserved in the dataset.

**Image Rendering and Point Cloud Sampling.** To support a variety of research topics like point cloud analysis, neural radiance fields, and 3D generation, we render multi-view images and sample point clouds based on the collected 3D models. We use Blender [17] to render object-centric and photo-realistic multi-view images, together with accurate camera poses. The images are rendered from 100 random viewpoints sampled on the upper hemisphere at 800 × 800 pixels. We also produce high-resolution mid-level cues like depth and normal for more research use. We then uniformly sample multi-resolution point clouds from each 3D model using the Open3D toolbox [111], with  $2^n$  ( $n \in \{10, 11, 12, 13, 14\}$ ) points in each point cloud, respectively. Besides the data existing in the dataset, we also provide a data generation pipeline. One can easily obtain new data with self-defined camera distributions, lighting, and point sampling methods to meet different requirements.

**Video Capturing and Annotation.** After scanning each object, we capture its video with an iPhone 12 Pro mobile phone. The object is placed on or beside a calibration board, and each video covers a full 360° range around it.

<sup>1</sup><https://www.einscan.com/>

<sup>2</sup><https://www.artec3d.cn/>

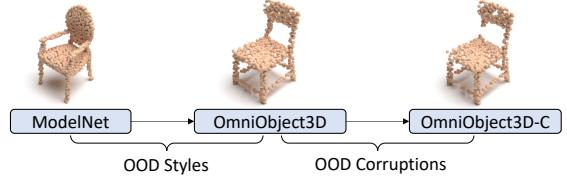


Figure 3. OmniObject3D provides the first clean real-world point cloud object dataset and allows fine-grained analysis on robustness to OOD styles and OOD corruptions.“-C”: corrupted by common corruptions described in [75]

Square corners on the calibration board can be recognized by the QR Codes beside it, and we then filter out the blurry frames whose recognized corners are less than 8. We uniformly sample 200 frames, and then COLMAP [77], a well-known SfM pipeline, is applied to annotate the frames with camera poses. Finally, we use the scales of the calibration board in both the SfM coordinate space and the real world to recover the absolute scale of the SfM coordinate system. We also develop a two-stage matting pipeline based on the U<sup>2</sup>Net [73] and FBA [24] matting model to generate the foreground masks for all the frames. Please refer to supplementary materials for more implementation details.

### 3.2. Statistics and Distribution

With 6,000 3D models in 190 categories, OmniObject3D exhibits a long-tailed distribution with an average of around 30 objects in each category, as shown in Figure 2. It shares many common categories with several famous 2D and 3D datasets [9, 19, 33, 47, 48, 80, 96]. For example, it covers 85 categories in ImageNet [19] and 130 categories in LVIS [33], which leads to the highest  $R^{lvis}$  in Table 1. Most of the categories are covered by the Open Images [47] image-level labels. It bears a huge diversity in shapes and appearances. The vast semantic and geometrical exploration spaces enable a wide range of research objectives.

## 4. Experiments

### 4.1. Robust 3D Perception

Object-level point cloud classification is one of the most fundamental tasks in point cloud perception. In this section, we show how OmniObject3D boosts robustness analysis of point cloud classification by disentangling the two critical out-of-distribution (OOD) challenges introduced in Sec. 2, *i.e.*, OOD styles and OOD corruptions.

Existing robustness evaluation utilizes clean synthetic datasets, *e.g.*, ModelNet [96], for training and sets up two kinds of test sets for evaluation:

1) *Noisy real-world datasets*, *e.g.*, ScanObjectNN [87], which are cropped from real-world scenes. They are employed to measure the robustness of the sim-to-real domain gap. However, the gap couples both OOD styles and OOD corruptions simultaneously as the cropped point clouds are

**Table 2. Point cloud perception robustness analysis on OmniObject3D with different architecture designs.** Models are trained on the ModelNet-40 dataset, with OA<sub>Clean</sub> to be their overall accuracy on the standard ModelNet-40 test set. OA<sub>Style</sub> on OmniObject3D evaluates the robustness to OOD styles. mCE on the corrupted OmniObject3D-C evaluates the robustness to OOD corruptions. Blue shadings indicate rankings. †: results on ModelNet-C [75]. Full results are presented in the supplementary materials.

	mCE <sup>†</sup> ↓	OA <sub>Clean</sub> ↑	OA <sub>Style</sub> ↑	mCE ↓
DGCNN [92]	1.000	0.926	0.448	1.000
PointNet [71]	1.422	0.907	0.466	0.969
PointNet++ [72]	1.072	0.930	0.407	1.066
RSCNN [51]	1.130	0.923	0.393	1.076
SimpleView [30]	1.047	<b>0.939</b>	0.476	0.990
GDANet [99]	<b>0.892</b>	0.934	<b>0.497</b>	<b>0.920</b>
PACConv [98]	1.104	0.936	0.403	1.073
CurveNet [97]	0.927	<b>0.938</b>	<b>0.500</b>	<b>0.929</b>
PCT [32]	0.925	0.930	0.459	0.940
RPC [75]	<b>0.863</b>	0.930	0.472	0.936

always noisy, making it impossible to analyze the two robustness challenges independently.

2) *Corrupted synthetic datasets*, e.g., ModelNet-C [75], which are artificially perturbed on top of clean synthetic datasets. The evaluation allows for detailed corruption analysis, but they do not reflect the robustness to OOD styles.

None of the existing robustness benchmarks allows for analyzing the robustness to both OOD styles and OOD corruptions in fine granularity. OmniObject3D, on the other hand, as the first clean real-world point cloud object dataset, can help to address the issue. For models trained on ModelNet, we first evaluate their performance on OmniObject3D to examine OOD-style robustness. Then, we create OmniObject3D-C by corrupting OmniObject3D with common corruptions described in [75] to examine the OOD-corruption robustness. We show a complete robustness evaluation scheme in Figure 3. For evaluation metrics, we use the overall accuracy (OA) on OmniObject3D to measure the OOD-style robustness and use DGCNN normalized mCE [75] to measure the OOD-corruption robustness.

We benchmark ten state-of-the-art point cloud classification models in Table R1. We observe that 1) performance on a clean test set has little correlation with OOD-style robustness. For example, SimpleView [30] achieves the best OA<sub>Clean</sub> but mediocre OA<sub>Style</sub>; 2) Advanced point grouping, e.g., curve-based point grouping in CurveNet [97] and frequency-based point grouping in GDANet [99], are robust not only to OOD corruptions as pointed out in [75], but also to OOD styles; 3) OOD style + OOD corruption is a more challenging setting. In particular, RPC, the most robust architecture to OOD corruptions [75], shows inferior mCE. In summary, robust point cloud perception models against both OOD styles and OOD corruptions are still under-explored. Our dataset sheds new light on a comprehensive understanding of point cloud perception robustness.

**Table 3. Single-scene novel view synthesis results.** Three metrics and their standard deviation (SD) across the training set.

Method	PSNR (↑) / SD	SSIM (↑) / SD	LPIPS (↓) / SD
NeRF [60]	34.01 / 3.46	0.953 / 0.029	0.068 / 0.061
mip-NeRF [5]	39.86 / 4.58	0.974 / 0.013	0.084 / 0.048
Plenoxels [25]	<b>41.04</b> / 6.84	<b>0.982</b> / 0.031	<b>0.030</b> / 0.031

**Table 4. Cross-scene novel view synthesis results on 10 categories.** ‘Cat.’ and ‘All\*’ denote training on each category and training on all categories except the 10 test ones, respectively.

Method	Train	PSNR (↑)	SSIM (↑)	LPIPS (↓)	$\mathcal{L}_1^{\text{depth}}$ (↓)
		All*	Cat.	Cat.-ft.	$\mathcal{L}_1^{\text{depth}}$ (↓)
MVSNeRF [11]	All*	17.49	0.544	0.442	0.193
	Cat.	17.54	0.542	0.448	0.230
	All*-ft.	25.70	0.754	0.251	0.081
	Cat.-ft.	25.52	0.750	0.264	<b>0.076</b>
IBRNet [91]	All*	19.39	0.569	0.399	0.423
	Cat.	19.03	0.551	0.415	0.290
	All*-ft.	<b>26.89</b>	<b>0.792</b>	<b>0.215</b>	0.081
	Cat.-ft.	25.67	0.760	0.238	0.099
pixelNeRF [105]	All*	22.16	0.692	0.342	0.109
	Cat.	20.65	0.676	0.348	0.195

See more results in the supplementary materials.

## 4.2. Novel View Synthesis

In this section, we study several representative methods on OmniObject3D for novel view synthesis (NVS) in two settings: 1) training on a single scene with densely captured images, which is the standard setting for NeRF [60]; 2) learning priors across scenes from our dataset to explore the generalization ability of NeRF-style models.

**Single-Scene NVS.** We select three objects in each category for the experiments, randomly sampling 1/8 images as the hold-out test set. We involve NeRF [60], mip-NeRF [5], and a voxel-based system named Plenoxels [25] for comparison. As in Table 3, we find that Plenoxels achieve the best performance on average for PSNR, SSIM [93], and LPIPS [108]. There exists a clear margin for LPIPS between Plenoxels and the other two methods since voxel-based methods are especially good at modeling high-frequency appearance. We also present the standard deviation (SD) of results across all the training samples, where Plenoxels are relatively unstable compared to NeRF and mip-NeRF. We observe that Plenoxels introduce artifacts when encountering concave geometry (e.g., bowls, chairs) and suffer from an inaccurate density field modeling when the foreground object is dark. MLP-based methods are relatively robust against these difficult cases. In a nutshell, our dataset provides a library with a variety of shapes and appearances, allowing a comprehensive evaluation of different NVS methods. See the supplementary for more results with the iPhone videos, detailed analysis, and visualizations.

**Cross-Scene NVS.** We conduct extensive experiments on novel view synthesis from sparse inputs by pixelNeRF [105], IBRNet [91] and MVSNeRF [11] in Table 4.

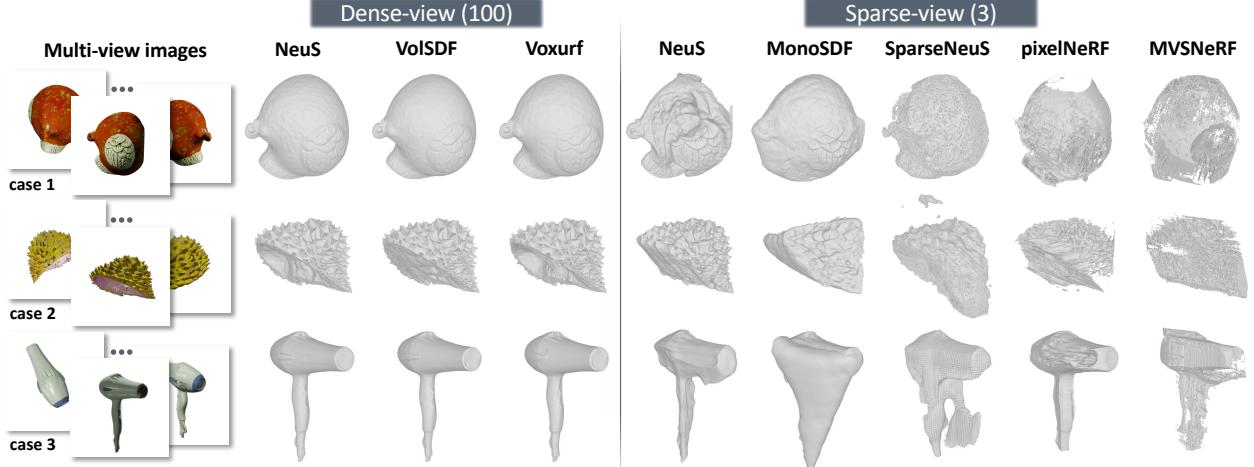


Figure 4. Neural surface reconstruction results for both dense-view and sparse-view settings.

Table 5. Dense-view surface reconstruction results.

Method	Chamfer Distance $\times 10^3$ ( $\downarrow$ )			
	Hard	Medium	Easy	Avg
NeuS [90]	9.26	5.63	3.46	6.09
VolSDF [103]	10.06	<b>4.94</b>	2.86	5.92
Voxurf [95]	<b>9.01</b>	4.98	<b>2.58</b>	<b>5.49</b>
Avg	9.44	5.19	2.97	5.83

We select 10 categories with the most various scenes as the test set. All metrics are averaged over 300 images. In the generalization setting, although not trained on the test category, MVSNeRF<sub>All\*</sub> is comparable to MVSNeRF<sub>Cat</sub>, IBRNet<sub>All\*</sub> and pixelNeRF<sub>All\*</sub> even outperforms the corresponding ‘Cat’s in all terms of visual metrics, especially on regular-shaped objects such as squash and apple. It confirms that OmniObject3D serves as an information-rich dataset that is beneficial for obtaining a strong generalizable prior on unseen scenes. Moreover, it is noteworthy that MVSNeRF and pixelNeRF with ‘All\*’ generate better underlying depth than those with ‘Cat.’, inferring generalizable methods can implicitly learn geometric cues though only trained from appearance in our dataset. It is reasonable that (1) IBRNet suffers more severely than the others in geometry under the scarcity of source context (only 3 views) as it is more suitable for dense-view generalization that complies with its view interpolation module. (2) MVSNeRF lags behind pixelNeRF on visual performance as we take 10 test frames widely distributed around the object in 360° by FPS sampling algorithm, where cost volume will be inaccurate on large-range viewpoint change. After further finetuning IBRNet for only around 10 minutes on each test scene, IBRNet<sub>All\*-ft</sub> achieves the best view synthesis results, comparable to test-time optimized NVS methods on nearby views. It is promising to utilize the large-scale and category-prosperous OmniObject3D, to build a benchmark suite for evaluating diverse cross-scene NVS methods.

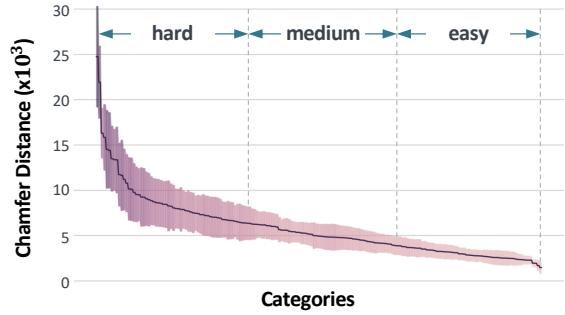


Figure 5. Performance distribution of dense-view surface reconstruction. The averaged results of the three methods is imbalanced. The colored area denotes a smoothed range of results.

### 4.3. Neural Surface Reconstruction

Precise surface reconstruction from multi-view images enables a broad range of applications. For a single scene with dense-view images, algorithms are expected to conduct accurate, robust, and efficient surface reconstruction. When only sparse-view images are available, it is crucial to learn generalizable priors from a set of scenes or use other geometric cues to assist reconstruction. Accordingly, we study the two settings for surface reconstruction methods.

**Dense-View Surface Reconstruction.** Under this setting, we include three representative methods. NeuS [90] and VolSDF [103] are two well-known approaches that bridge neural volume rendering with implicit surface representation. We also involve a voxel-based method called Voxurf [95], which leverages a hybrid representation to achieve acceleration and fine geometry reconstruction.

Previous approaches in this task mainly perform evaluations and comparisons on 15 scenes from the DTU [1] dataset, which is not comprehensive and robust enough to demonstrate the ability of the methods in different scenarios. In comparison, we select three objects per category to run each of the three methods above, leading to over 1,500 reconstructions in total. We calculate the Chamfer

Table 6. Sparse-view (3-view) surface reconstruction results.

Method	Train	Chamfer Distance $\times 10^3$ ( $\downarrow$ )			
		Hard	Medium	Easy	Avg
NeuS [90]	Single	29.35	27.62	24.79	27.3
MonoSDF [106]	Single	35.14	35.35	32.76	34.68
SparseNeuS [54]	1 cat.	34.05	31.32	31.14	32.36
	10 cats.	30.75	30.11	28.37	29.87
	All cats.	<b>26.13</b>	<b>26.08</b>	<b>22.13</b>	<b>25.00</b>
	Easy	28.39	26.65	23.76	26.48
	Medium	27.38	26.66	23.08	25.87
	Hard	27.42	26.95	24.63	26.47
MVSNeRF [11]	All cats.	56.68	48.09	48.70	51.16
pixelNeRF [105]	All cats.	63.31	59.91	61.47	61.56

Distance (CD) between the reconstructed surface and the ground truth. The distribution of the results is shown in Figure 5. The average curve is imbalanced: hard categories usually include low-textured, concave, or complex shapes (*e.g.*, bowls, vases, kennels, cabinets, and durians). We thus split the categories into three levels of “difficulty” based on the average curve, and the level-wise results are presented in Table 5. We can observe a clear margin among results in different levels for each method, indicating the split subsets to be generic and faithful.

**Sparse-View Surface Reconstruction.** Dense-captured images of a scene are sometimes not available, so we also study the sparse-view scenario here. The following methods are included: NeuS [90] with sparse-view input; MonoSDF [106], which takes in geometry cues estimated by pre-trained models; SparseNeuS [54], a generic surface prediction pipeline that learns generalizable priors; pixelNeRF [105] and MVSNeRF [11] from Sec. 4.2, whose geometries are extracted from the density field. For NeuS, MonoSDF, and pixelNeRF, we use Farthest Point Sampling (FPS) to sample views that are most widely distributed; for SparseNeuS and MVSNeRF, we conduct FPS among the nearest 30 camera poses from a random reference view. We sample 3 views in all the experiments.

The quantitative and qualitative comparisons are shown in Table 6 and Figure 4, respectively. We observe apparent artifacts in all the sparse-view reconstructed results. Among them, SparseNeuS trained on enough data demonstrates the best quantitative performance on average, and the pre-division on the train set does not result in a noticeable difference across difficulty levels. NeuS with sparse-view input achieves a surprisingly good performance. As shown in Figure 4, the FPS sampling equips NeuS with a coherent global shape for thin structures like case 3, while it encounters severe local geometry ambiguity like case 1. MonoSDF, on the contrary, partially overcomes the issue of ambiguity via the assistance of predicted geometry cues in case 1. However, it relies heavily on the accuracy of the estimated depth and normal and thus easily fails when the

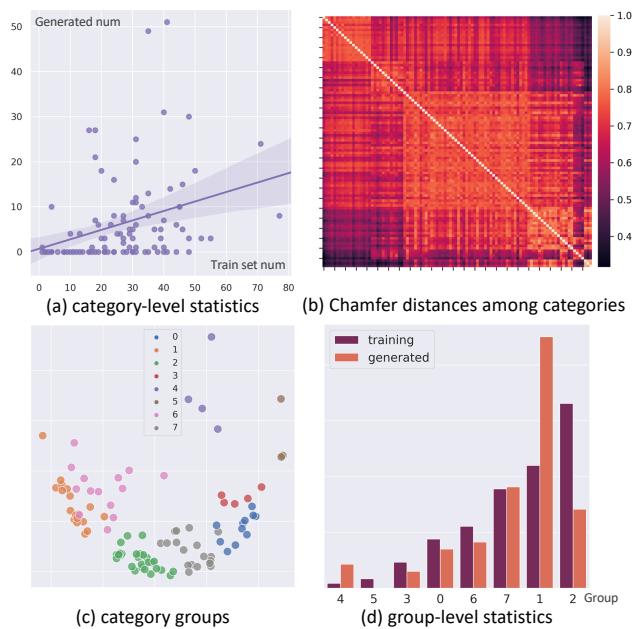


Figure 6. The category distribution of the generated shapes. (a) shows a weak positive correlation between the number of generated shapes and training shapes per category. (b) visualizes the correlation matrix among different categories by Chamfer Distance between their mean shapes. (c) visualizes categories being clustered into eight groups by KMeans. (d) presents a clear training-generation relation in the group-level statistics.

estimation is inaccurate (*e.g.*, case 2 and 3). The surfaces extracted from generalized NeRF models, *i.e.*, pixelNeRF and MVSNeRF, are of relatively low quality.

In brief, the challenging problem of sparse-view surface reconstruction has not been solved well currently. OmniObject3D is a promising database to study generalizable surface reconstruction pipelines as well as strategies for a robust usage of estimated geometry cues for this track.

#### 4.4. 3D Object Generation

In this section, we adopt a state-of-the-art generative model that directly generates explicit textured 3D meshes, namely GET3D [29]. GET3D is originally evaluated on six categories (*Car*, *Chair*, *Motorbike*, *Animal*, *House*, and *Human Body*) with independent models trained on each category. The number of shapes per category ranges from 337 to 7,497. In comparison, OmniObject3D contains many more categories with fewer objects in each. As a result, it is natural to train each model with multiple categories.

We first provide some qualitative results in Figure 7, where we show various generation results: The textures are rather realistic, and the shapes are coherent, enhanced by fine geometry details (*e.g.*, the lychee and pineapple). We explore the latent space of the model and show interpolation results in Figure 8. We can observe a smooth transition across instances that are semantically different. We

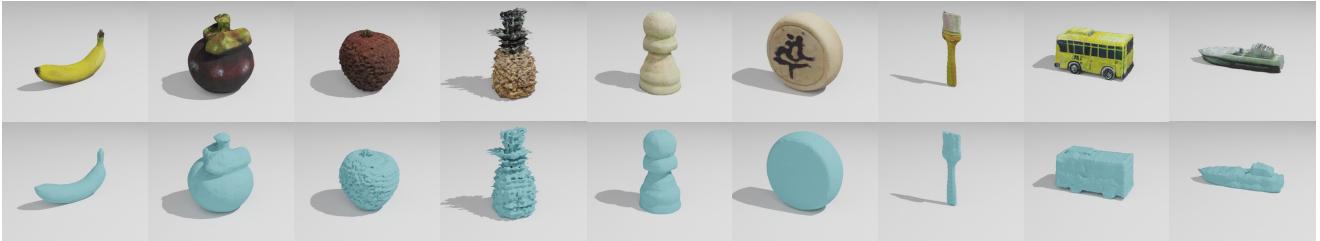


Figure 7. **Examples of the generated textured shapes rendered in Blender.** OmniObject3D enables GET3D with realistic generations across a wide range of categories.



Figure 8. **Shape interpolation.** We interpolate both geometry and texture latent codes from left to right.

would like to further analyze the performance of the generative model trained on OmniObject3D from three aspects, *i.e.*, *semantic distribution*, *diversity*, and *quality*.

**Semantic Distribution.** We randomly select 100 categories to train an unconditional model jointly. We randomly generate 1,000 textured meshes at inference time and ask human experts to label them. Shapes with ambiguity are not counted. Figure 6 (a) shows that the generated shapes per category are highly imbalanced, exhibiting a weak positive correlation with the training shape numbers. Actually, the categories are not independent but rather highly correlated. We calculate the “mean shape” for each category and visualize the Chamfer Distance among them in Figure 6 (b), which indicates that they can be further grouped. Regarding each matrix row as a feature vector, we use KMeans to cluster them into eight groups (Figure 6 (c)) and carry out *group-level statistics* in Figure 6 (d). It demonstrates a clear trend that the number of generated shapes increases along with or even faster than the number of training shapes in the group, revealing an enlarged semantic bias during generation. However, it also depends on the inner-group divergence. For example, Group 2 (883 shapes in 27 categories) has the largest number of training samples, while the high variation among its categories prevents it from dominating the generated shapes; Group 1 (587 shapes in 18 categories) has a relatively small divergence, which becomes the most popular in the generated shapes. We present more details in the supplementary material.

**Diversity and Quality.** We select four representative data subsets for training and evaluation, namely *fruits*, *furniture*, *toys*, and *Rand-100*. We randomly split each subset into training (80%) and testing (20%). We leverage three evaluation metrics: for geometry, we use Chamfer Distance (CD) to compute the Coverage score (Cov) and Minimum Matching Distance (MMD), which focus on the diversity

Table 7. **Quantitative evaluations on different data splits.**

Split	#Objs	#Cats	Cov (% ↑)	MMD (↓)	FID (↓)	FID <sup>ref</sup>
Furniture	265	17	<b>67.92</b>	4.27	87.39	58.40
Fruits	610	17	46.72	3.32	105.31	87.15
Toys	339	7	55.22	<b>2.78</b>	122.77	41.40
Rand-100	2951	100	61.70	3.89	<b>46.57</b>	<b>8.65</b>

and quality of the shapes, respectively; for texture, we adopt the widely-used FID [37]. But the FID metrics with different test splits are not directly comparable, suffering from a large variance when the test set is small. We thus introduce FID<sup>ref</sup> for reference, which is the FID between the train and test set. The results are shown in Table 7. *Furniture* suffers from the lowest quality (MMD) since the small train set with 17 categories is a difficult training source. *Fruits* has the same number of categories while being 2.3 times larger in scale, and some fruits share a very similar structure, leading to relatively higher quality and lower diversity (Cov). *Toys* achieve the best quality by training on only 7 categories. *Rand-100* is the most difficult case, and we can observe a trade-off between quality and diversity. Both the FID and FID<sup>ref</sup> are high for the first three subsets due of the small testing sets, while only ‘Rand-100’ is relatively low.

In brief, training and evaluating generative models on a large-vocabulary and realistic dataset is a promising but challenging task. We reveal crucial problems like the semantic distribution bias and varying exploration difficulties in different groups. OmniObject3D serves as a great database for further examination in this area.

## 5. Conclusion and Outlook

We introduce OmniObject3D, a large vocabulary 3D object dataset with massive high-quality real-scanned 3D objects, including 6,000 objects from 190 categories. It provides rich annotations, including textured 3D meshes, sampled point clouds, posed multi-view images rendered by Blender, and real-captured video frames with foreground masks and COLMAP camera poses. We set up four evaluation tracks, revealing new observations, challenges, and opportunities for future research in realistic 3D vision.

We will regulate the usage of our data to avoid potential negative social impacts.

**Acknowledgement.** This project is funded by Shanghai

AI Laboratory, CUHK Interdisciplinary AI Research Institute, the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, Hong Kong RGC Theme-based Research Scheme 2020/21 (No. T41-603/20- R), OpenXD-Lab, the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative.

## A. Additional Information of OmniObject3D

We first provide a full category list with the number of objects for each class in Figure S1. Most of the categories have [10, 40] objects. The dataset includes objects that have undergone common manipulations, as shown in Figure S3 (b). For each object, the raw data includes a textured 3D mesh and several surrounding videos. To demonstrate the completeness and high quality of our scanned objects, we compare the quality between the COLMAP sparse reconstruction and the textured mesh from the scanner in Figure S3 (c). Given a high-fidelity 3D scan, we can render realistic and high-resolution multi-view images with modern graphics engines like the Blender [17], where we also save the corresponding depth and normal maps (Figure S2) for different research usage. We also provide the users with posed frames from the real-captured videos following [74]. We leverage the calibration board and COLMAP [77] to recover the poses of selected frames with a real-world scale, as described in the main text, and then we develop a matting pipeline based on a two-stage U<sup>2</sup>-Net [73] model together with a post-processing FBA [24] model. In detail, we first utilize the Rembg<sup>3</sup> tool on image frames to remove backgrounds from different categories and choose 3,000 good results as the pseudo segmentation labels. We then refine our pipeline by fine-tuning with the pseudo labels to boost its segmentation ability on objects. We show some examples and failure cases of our segmentation pipeline in Figure S3 (a).

## B. Related Works

We have briefly discussed the related works for the four benchmarks in the main text, and we conduct a more comprehensive discussion here.

**Robust Point Cloud Perception.** Robustness to out-of-distribution (OOD) data has been an important topic in point cloud perception since point clouds are widely employed in safety-critical applications, *e.g.*, autonomous driving. In particular, OOD styles (*e.g.*, different styles in CAD models and real-world objects) and OOD corruptions (*e.g.*, missing points) are two main challenges to point cloud perception robustness. A line of work [13, 45, 71, 92] evaluates

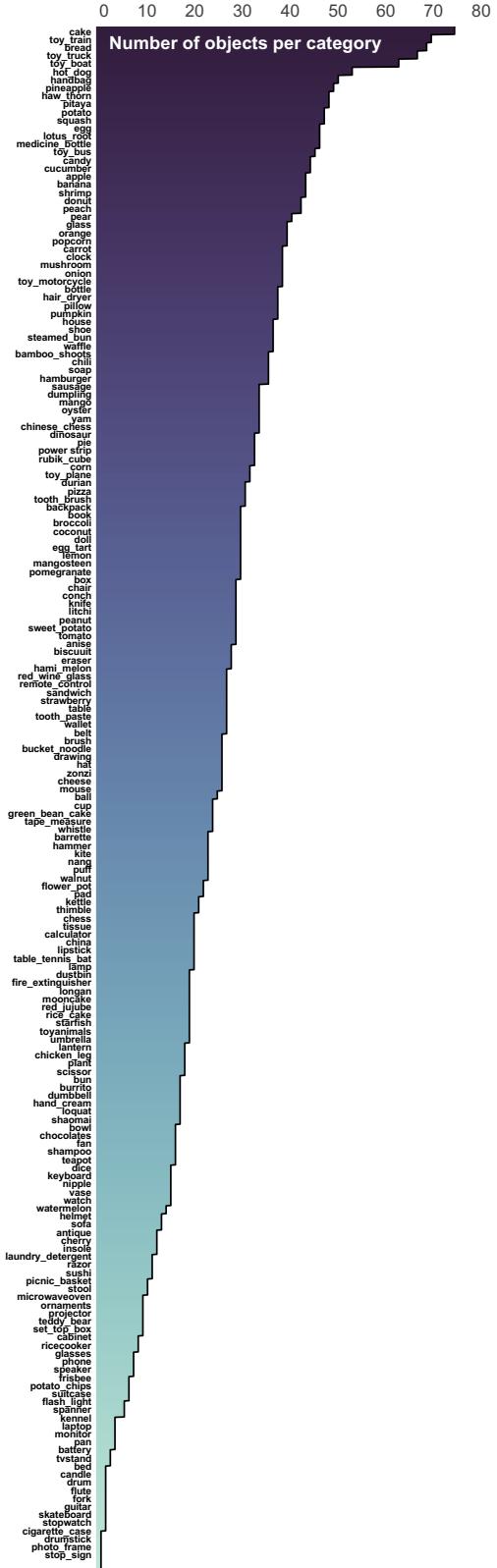


Figure S1. A full class list with number of objects per category.

<sup>3</sup><https://github.com/danielgatis/rembg>

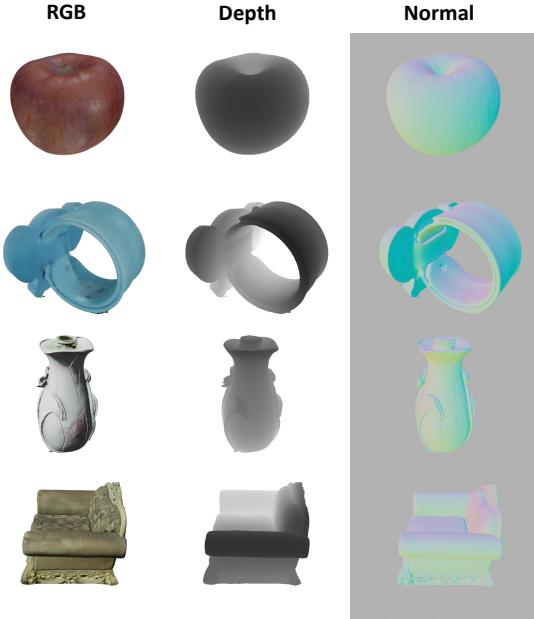


Figure S2. Examples of the Blender [17] rendered results.

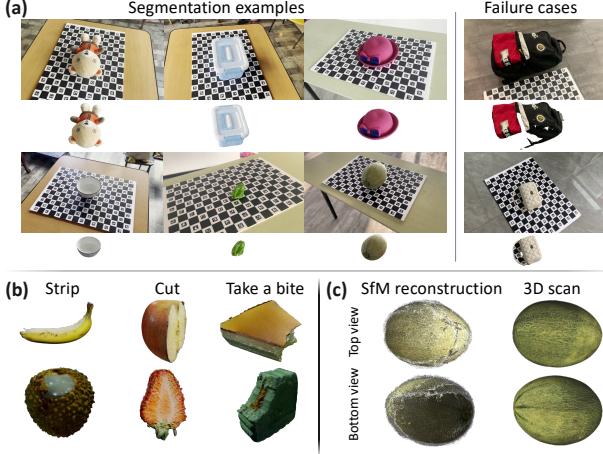


Figure S3. Examples of the segmentation (a), manipulation (b), and reconstruction (c). In (c), the missing bottom of the SfM reconstruction from video frames is due to its touch with the table.

the robustness to OOD corruptions by adding corruptions like random jittering and rotation to clean test sets. Recent work works [75, 85] further systematically anatomize the corruptions and propose a standard corruption test suite. However, they fail to take account of OOD styles. Another line of work [3, 74] evaluates the sim-to-real domain gap by testing models trained on clean synthetic datasets (*e.g.*, ModelNet-40 [96]) on noisy real-world test sets (*e.g.*, ScanObjectNN [87]). However, the sim-to-real gap couples OOD styles and OOD corruptions at the same time, which makes the results hard to analyze. In this work, we use OmniObject3D dataset to provide high-quality real-world point cloud to measure the OOD style robustness, and apply sys-

tematic corruptions on top of it to measure the OOD corruptions robustness. We hence provide the first point cloud perception benchmark that allows fine-grained evaluation of the robustness on both OOD styles and corruptions.

**Neural Radiance Field.** Neural radiance field (NeRF) [60] represents a scene with a fully-connected deep network (MLPs), which takes in hundreds of sampled points along each camera ray and outputs the predicted color and density. Novel views of the scene are synthesized by projecting the colors and densities into an image via volume rendering. Inspired by the success of NeRF, a massive follow-up effort has been made to improve its quality [5, 6, 59, 89], and efficiency [10, 25, 62, 84]. A branch of works [11, 52, 74, 91, 105] has also explored the generalization ability of NeRF-based frameworks. PixelNeRF [105], MVSNeRF [11], IBR-Net [91], and NeuRay [52] reconstruct the radiance field with a mere forward pass during inference via training on cross-scenes. NeRFormer [74], IBRNet [91], and GNT [88] leverage Transformers for generalizable NeRF.

**Neural Surface Reconstruction.** Implicit Neural Representations (INR) [4, 15, 40, 53, 58, 68, 76, 81, 86, 107] of 3D object geometry and appearance with neural networks have attracted increasing attention in recent years. Some approaches [44, 50, 65, 104] regard the color of an intersection point between the ray and the surface as the rendered color, namely surface rendering, and they typically rely on accurate object masks. Another trend of recent approaches [18, 66, 90, 95, 103] proposes to leverage neural radiance field with implicit surface representations like Signed Distance Function (SDF) for higher-quality and mask-free surface reconstruction from multi-view images. NeuS [90], VolSDF [103] reconstruct implicit surfaces with an SDF-based volume rendering scheme, and Voxsurf [95] leverages an explicit volumetric representation for acceleration. Since dense camera views of scenes are sometimes unavailable, SparseNeuS [54] and MonoSDF [106] explore surface reconstruction from sparse views. The former exploits generalizable priors cross scenes for a generic surface prediction, while the latter takes advantage of the estimated geometry cues predicted by pretrained networks.

OmniObject3D can serve as a large-scale benchmark with realistic photos and meshes for both training and evaluation. It bears a large vocabulary and high diversity in shape and appearance, offering an opportunity for pursuing more generalizable and robust novel view synthesis and surface reconstruction methods.

**3D Object Generation.** Recent advances in photorealistic 2D image generations [20, 23, 38, 41–43, 69] inspire the explorations of 3D content generation. Early approaches [27, 35, 55, 82, 94] extend 2D generation frameworks to 3D voxels with a high computational cost when generating high-resolution contents. Some other works adopt different 3D data formulations, *e.g.*, point cloud [2,

Table R1. **Point cloud perception robustness analysis on OmniObject3D with different architecture designs.** Models are trained on ModelNet-40 dataset. OA on OmniObject3D evaluates the robustness to OOD styles. mean Corruption Error (mCE) on the corrupted OmniObject3D-C evaluates the robustness to OOD corruptions. The blue cells denote best in each row, and the red cells denote the worst.

	OA <sub>Clean</sub> $\uparrow$	OA <sub>Style</sub> $\uparrow$		Scale	Jitter	Drop-G	Drop-L	Add-G	Add-L	Rotate	mCE $\downarrow$
DGCNN [92]	0.926	0.448		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
PointNet [71]	0.907	0.466		<u>0.925</u>	<b>0.858</b>	0.976	<b>0.816</b>	1.318	0.921	0.935	0.969
PointNet++ [72]	0.930	0.407		1.104	1.071	1.108	0.886	1.101	1.123	1.031	1.066
RSCNN [51]	0.923	0.393		1.115	1.078	1.144	0.997	1.042	1.079	1.025	1.076
SimpleView [30]	<b>0.939</b>	0.476		<u>0.940</u>	0.951	0.959	1.012	1.043	1.037	0.949	0.990
GDANet [99]	0.934	<u>0.497</u>		<b>0.887</b>	0.933	<u>0.923</u>	0.975	<b>0.884</b>	0.921	<u>0.882</u>	<b>0.920</b>
PACConv [98]	0.936	0.403		1.034	1.101	<u>1.032</u>	1.052	1.159	1.057	1.082	1.073
CurveNet [97]	<u>0.938</u>	<b>0.500</b>		0.930	<u>0.930</u>	<b>0.920</b>	0.869	0.929	0.997	<u>0.907</u>	<u>0.929</u>
PCT [32]	0.930	0.459		0.950	0.986	1.011	0.862	<u>0.921</u>	<u>0.912</u>	1.001	0.940
RPC [75]	0.930	0.472		0.947	0.940	0.967	<u>0.855</u>	0.999	<b>0.909</b>	0.915	0.936

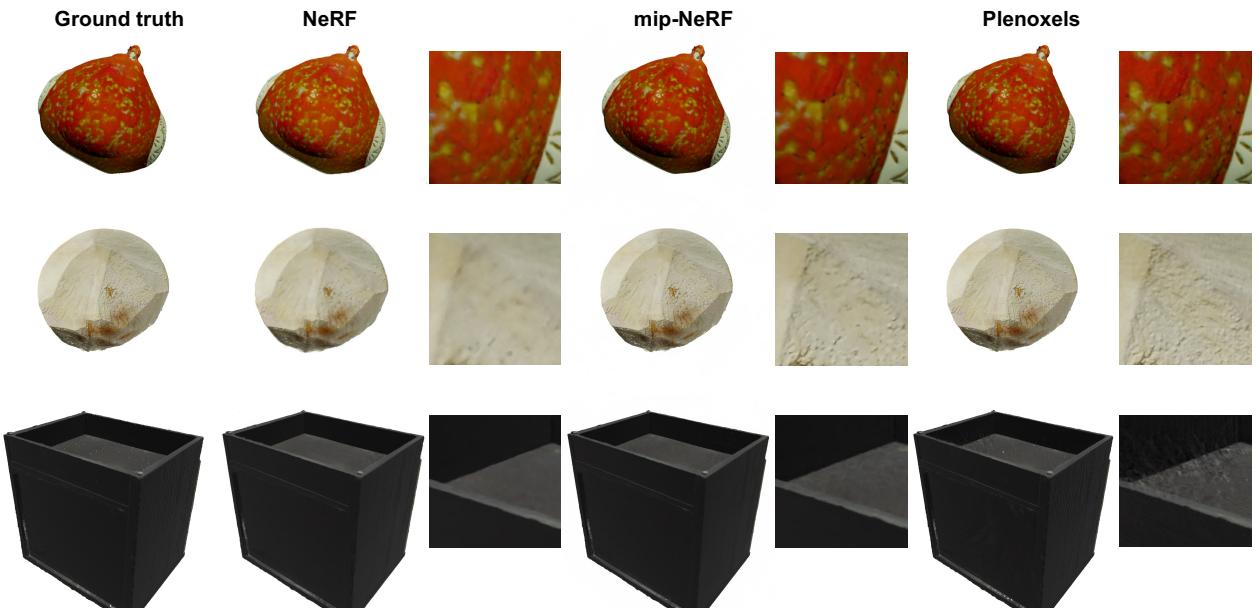


Figure S4. Qualitative comparisons of single-scene NVS methods in different rendered scenes from our dataset.

61, 101, 109] and octree [39] to generate coarse geometry. OccNet [57], IM-NET [14] generates the 3D meshes with implicit representation while extracting high-quality surfaces is non-trivial. Encouraged by NeRF [60], extensive works [7, 8, 31, 34, 64, 67, 78, 79, 100, 110] explore 3D-aware image synthesis rather than mesh generation. Aiming at generating textured 3D meshes, Textured3DGAN [70] and DIBR [12] deform template meshes, preventing them from complex shapes. PolyGen [63], SurfGen [56], and GET3D [28] generate meshes with arbitrary topology. Distinguishable from others, GET3D generates diverse meshes with rich geometry and textures. With the proposed OmniObject3D dataset, we extend the benchmarks of realistic 3D generation approaches to large vocabulary and massive objects, enabling the exploration of better generation quality and diversity.

## C. Additional Experimental Results

### C.1. Robust 3D Perception

Following ModelNet-C [75], we perform seven kinds of out-of-distribution (OOD) corruptions for study, including “Scale”, “Jitter”, “Drop Global/Local”, “Add Global/Local”, and “Rotate”. Please refer to their paper for a detailed illustration of each corruption type. We calculate the error under each corruption and the mean Corruption Error (mCE) is an average of the results. The full evaluation results corresponding to are shown in Table R1.

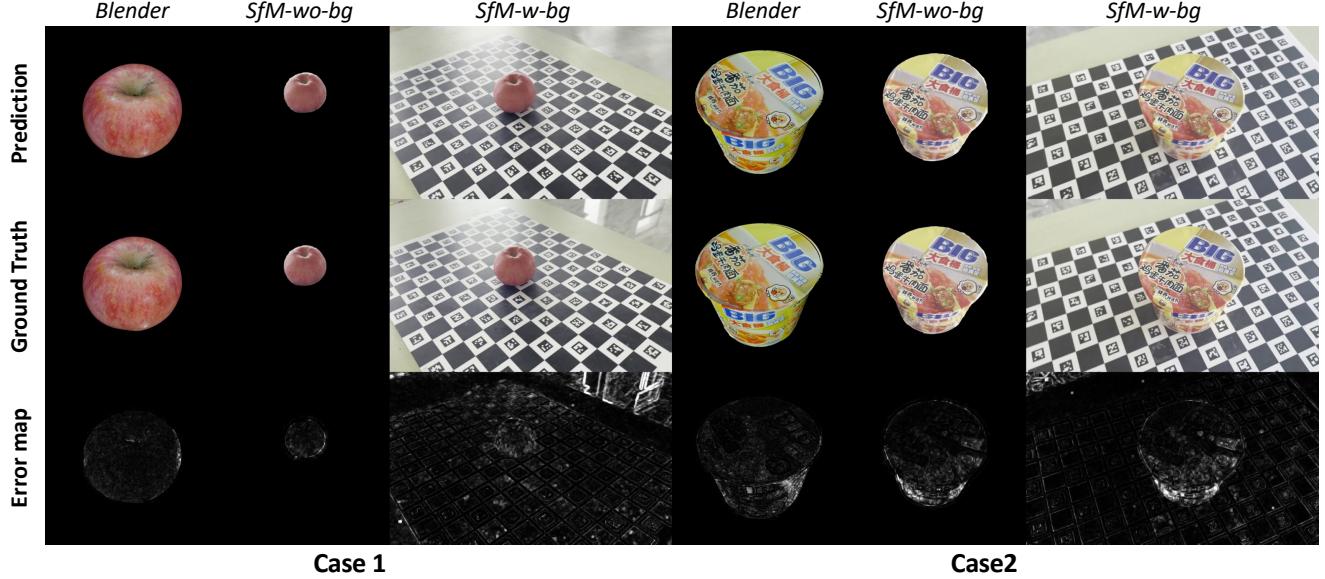


Figure S5. Qualitative comparisons of NVS on the same scenes with different data types.

**Table R2. Comparisons of 3 single-scene NVS methods on different data types.** For all the methods we involve, we can observe that the *Blender* setting performs the best; the *SfM-wo-bg* setting is a little bit worse due to the motion blur and potential inaccuracy in SfM pose estimation; the *SfM-w-bg* setting always achieves the lowest PSNR, as the background in the unbounded scene introduces further challenges.

Method	Data-type	PSNR ( $\uparrow$ )
NeRF [60]	SfM-w-bg	22.92
	SfM-wo-bg	24.70
	Blender	<b>28.07</b>
Mip-NeRF [5]	SfM-w-bg	23.29
	SfM-wo-bg	25.62
	Blender	<b>31.25</b>
Plenoxel [25]	SfM-w-bg	14.06
	SfM-wo-bg	19.18
	Blender	<b>28.07</b>

## C.2. Novel View Synthesis

### C.2.1 Single-Scene NVS

**Implementation Details.** We use the official code and default settings by NeRF [60], Mip-NeRF [6], and Plenoxels [25] in this section. For NeRF, we re-weight the foreground and background contents by 1:0.5 to avoid all-black output. For Plenoxel on the SfM data with background, we enable the background model provided by the official code to model the background area.

#### Qualitative Comparisons of NVS on rendered images.

We describe the performance of three representative methods in the main text, and we provide some qualitative com-

parisons here in Figure S4, accordingly. Plenoxels are especially good at modelling high-frequency textures (e.g., the coconut), while it is less robust than NeRF and mip-NeRF when dealing with dark textures and concave geometry, suffering from inaccurate geometry. Our dataset helps to provide a comprehensive evaluation of different methods.

**Comparisons of NVS on rendered images and iPhone videos.** We conduct qualitative and quantitative evaluations on novel view synthesis with several scenes under different data types, including *SfM-wo-bg*, *SfM-w-bg* and *Blender*. The *SfM-wo-bg* and *SfM-w-bg* settings use images sampled from iPhone videos and camera parameters generated by COLMAP. The difference between them is whether the background is included. The *Blender* data are rendered by Blender [17]. Since the image resolutions and foreground proportions are different among the data types, we calculate the PSNR metric only in the foreground area for *SfM-wo-bg* data and *Blender* data, whereas for *SfM-w-bg* data, every pixel in the image is included PSNR calculation.

Based on the qualitative comparisons in Figure S5, we observe that for both two selected scenes, the predicted novel view image under the *Blender* setting achieves the best visual quality, resulting in the highest PSNR in Table R2. When comparing the two SfM based data types, we find that the quality of the foreground object from the *SfM-wo-bg* data is only slightly better than the other, while the high background error under the *SfM-w-bg* setting leads to a significant drop in performance, as shown in Table R2. The experimental results shed light on how real-captured videos introduce further challenges to NeRF-like methods. We demonstrate that performing robust novel view synthesis with casually captured videos will be an important and

Table R3. **Cross-scene novel view synthesis results on 10 categories.** We evaluate our benchmarks on 3 unseen scenes per category with 3 source views. In each scene, we take 10 test frames widely distributed around the object by FPS sampling strategy.

Method	Train	Metric	toy train	bread	cake	toy boat	hot dog	wallet	pitaya	squash	handbag	apple
MVSNeRF [11]	All*	PSNR	15.90	16.80	15.47	16.28	15.84	20.58	18.69	17.81	18.02	19.55
		SSIM	0.501	0.548	0.522	0.519	0.497	0.534	0.490	0.576	0.564	0.681
		LPIPS	0.480	0.456	0.480	0.408	0.429	0.449	0.456	0.417	0.444	0.403
		$\mathcal{L}_1^{\text{depth}}$	0.182	0.155	0.249	0.253	0.127	0.261	0.178	0.187	0.229	0.113
	Cat.	PSNR	16.14	16.87	14.60	15.65	16.64	20.76	19.09	16.97	18.35	20.40
		SSIM	0.515	0.560	0.527	0.444	0.520	0.524	0.505	0.548	0.575	0.709
		LPIPS	0.475	0.463	0.488	0.433	0.431	0.464	0.449	0.435	0.444	0.399
		$\mathcal{L}_1^{\text{depth}}$	0.175	0.127	0.339	0.477	0.134	0.382	0.237	0.101	0.219	0.112
	All*-ft	PSNR	23.16	25.82	25.14	23.47	23.91	27.83	25.36	25.68	26.09	30.53
		SSIM	0.717	0.769	0.745	0.736	0.714	0.739	0.710	0.761	0.803	0.845
		LPIPS	0.281	0.224	0.263	0.228	0.248	0.293	0.227	0.255	0.280	0.215
		$\mathcal{L}_1^{\text{depth}}$	0.091	0.062	0.081	0.141	0.053	0.078	0.069	0.061	0.130	0.053
	Cat.-ft	PSNR	22.88	25.58	25.29	23.80	23.44	27.38	25.46	25.40	25.94	30.06
		SSIM	0.721	0.758	0.748	0.733	0.698	0.722	0.715	0.759	0.803	0.840
		LPIPS	0.283	0.243	0.262	0.226	0.280	0.318	0.229	0.277	0.283	0.244
		$\mathcal{L}_1^{\text{depth}}$	0.122	0.053	0.064	0.096	0.060	0.084	0.071	0.048	0.120	0.046
IBRNet [91]	All*	PSNR	17.90	19.08	17.09	17.89	17.77	23.13	20.11	20.25	18.36	22.36
		SSIM	0.526	0.599	0.538	0.530	0.516	0.579	0.511	0.632	0.530	0.726
		LPIPS	0.430	0.383	0.422	0.368	0.394	0.426	0.405	0.356	0.451	0.352
		$\mathcal{L}_1^{\text{depth}}$	0.379	0.327	0.610	0.357	0.338	0.419	0.388	0.392	0.847	0.175
	Cat.	PSNR	17.33	18.30	16.87	17.13	17.83	23.39	19.62	19.05	19.73	21.02
		SSIM	0.502	0.554	0.525	0.491	0.498	0.579	0.485	0.606	0.584	0.684
		LPIPS	0.449	0.415	0.446	0.394	0.413	0.427	0.420	0.376	0.443	0.371
		$\mathcal{L}_1^{\text{depth}}$	0.417	0.394	0.392	0.169	0.096	0.234	0.177	0.352	0.336	0.331
	All*-ft	PSNR	22.12	27.53	26.28	25.80	22.89	30.03	26.33	29.15	26.74	32.00
		SSIM	0.683	0.829	0.769	0.834	0.686	0.814	0.764	0.845	0.815	0.885
		LPIPS	0.298	0.177	0.238	0.152	0.267	0.211	0.199	0.177	0.268	0.163
		$\mathcal{L}_1^{\text{depth}}$	0.232	0.051	0.079	0.083	0.054	0.036	0.075	0.051	0.073	0.080
	Cat.-ft	PSNR	21.90	26.47	24.83	22.46	24.74	27.68	26.41	25.37	26.61	30.18
		SSIM	0.678	0.804	0.739	0.707	0.755	0.727	0.766	0.745	0.813	0.861
		LPIPS	0.301	0.195	0.261	0.233	0.210	0.280	0.197	0.254	0.266	0.184
		$\mathcal{L}_1^{\text{depth}}$	0.225	0.049	0.070	0.101	0.046	0.063	0.062	0.195	0.065	0.111
pixelNeRF [105]	All*	PSNR	19.77	21.54	20.77	20.15	20.93	24.73	21.78	23.48	21.30	27.18
		SSIM	0.647	0.701	0.690	0.661	0.671	0.666	0.606	0.748	0.696	0.833
		LPIPS	0.377	0.331	0.363	0.315	0.339	0.393	0.370	0.283	0.381	0.269
		$\mathcal{L}_1^{\text{depth}}$	0.142	0.131	0.141	0.109	0.073	0.085	0.114	0.065	0.175	0.061
	Cat.	PSNR	19.91	20.93	17.55	20.20	19.63	24.16	20.80	18.59	19.84	24.96
		SSIM	0.685	0.702	0.622	0.686	0.645	0.662	0.606	0.667	0.657	0.828
		LPIPS	0.332	0.330	0.426	0.275	0.348	0.392	0.367	0.342	0.420	0.249
		$\mathcal{L}_1^{\text{depth}}$	0.136	0.224	0.364	0.119	0.142	0.152	0.243	0.181	0.336	0.054

practical topic.

## C.2.2 Cross-Scene NVS

**Implementation Details.** We use the official codes to evaluate three benchmarks on 10 categories, *i.e.*, toy train, bread, cake, toy boat, hot dog, wallet, pitaya, squash, handbag, and apple. We split three scenes from each category as a test-set, and the remaining scenes are used as a train-set. During training, we randomly sample rays from scenes in the train-set of each category and use Adam [46] optimizer. For a fair comparison, we evaluate these methods with the

same source views, *i.e.*, 3 views from nearby 30 views (explained in Sec. C.3.2) by FPS sampling. Then in a scene with 100 rendered views, we exclude these 3 source views and select 10 test views from the remaining 97 views by FPS criteria again. For MVSNeRF, we pretrain the ‘All\*’ with total 300k iterations, and the ‘Cat.’ with 20k to 40k iterations depending on the number of scenes. In finetuning stage, we take 3 views as input and additional 13 views sampling for per-scene optimization. Each scene is finetuned for 15k iterations. For IBRNet, we pretrain the ‘All\*’ with 300k iterations, and the ‘Cat.’ with 50k iterations. After cross-scene training, we further finetune the model with 15k

Table R4. Unaligned Cross-scene novel view synthesis results of pixelNeRF-U [105] on 10 categories.

Train	Metric	toy train	bread	cake	toy boat	hot dog	wallet	pitaya	squash	handbag	apple
All*	PSNR	18.81 -0.96	19.92 -1.62	19.86 -0.91	19.54 -0.29	19.64 -1.29	20.31 -4.42	20.44 -1.34	20.74 -2.74	20.79 -0.51	21.21 -5.97
	SSIM	0.591 -0.056	0.625 -0.076	0.636 -0.054	0.626 -0.035	0.627 -0.044	0.628 -0.038	0.619 +0.013	0.631 -0.117	0.635 -0.061	0.650 -0.183
	LPIPS	0.432 -0.055	0.406 -0.075	0.405 -0.042	0.398 -0.083	0.397 -0.058	0.401 -0.008	0.405 -0.035	0.394 -0.111	0.397 -0.016	0.390 -0.121
	$\mathcal{L}_1^{\text{depth}}$	0.145 -0.003	0.118 +0.013	0.123 +0.018	0.132 -0.023	0.122 -0.049	0.120 -0.035	0.119 -0.005	0.113 -0.048	0.121 +0.054	0.117 -0.056
	PSNR	19.36 -0.55	19.03 -1.90	18.46 -0.91	18.45 -1.75	18.53 -1.10	19.41 -4.75	19.51 -1.29	19.34 -0.75	19.38 -0.46	19.58 -5.38
Cat.	SSIM	0.637 -0.048	0.636 -0.066	0.626 +0.004	0.624 -0.062	0.623 -0.022	0.625 -0.037	0.616 +0.010	0.614 -0.053	0.618 -0.039	0.631 -0.197
	LPIPS	0.392 -0.060	0.402 -0.072	0.415 +0.011	0.400 -0.125	0.396 -0.048	0.399 -0.007	0.403 -0.036	0.404 -0.062	0.408 +0.012	0.404 -0.155
	$\mathcal{L}_1^{\text{depth}}$	0.172 -0.036	0.219 +0.005	0.260 +0.104	0.262 -0.143	0.247 -0.105	0.251 -0.099	0.252 -0.009	0.286 -0.105	0.293 +0.043	0.276 -0.222

iterations on each test scene. For pixelNeRF, we train the ‘All\*’ with 400k iterations, and the ‘Cat.’ with 12k to 30k iterations depending on the number of scenes. All methods sample rays within a tight foreground bounding box around the object.

**Detailed Comparisons.** The full evaluation results are presented in Table R3. We additionally provide qualitative comparisons of 4 cases, each with rendered RGB and depth, as shown in Figure S6 (we leave an extra 15 pixels of each edge). We evaluate depth within the foreground masks. From the visualization, it may seem that methods w/ ‘Cat.’ generate more accurate contour than that w/ ‘All\*’, contradicting the statement that methods w/ ‘All\*’ can learn a better geometric cue than that w/ ‘Cat.’ in the main context. However, we find that within the masks, the depth of the former is generally more precise than that of the latter, obviously illustrated by “pitaya” (the third case) in pixelNeRF. It may raise an interesting research topic of how generic methods can perform both accurately in shape contour and geometry. After slightly finetuning MVSNeRF and IBRNet on a test scene, these methods achieve comparable performance with scene-specific methods, *e.g.*, NeRF.

**Results on Unaligned Coordinate System.** We additionally provide a more challenging setting by evaluating Cross-Scene NVS on an unaligned coordinate system rather than in a perfectly predefined canonical space. Specifically, we examine pixelNeRF-U [105], where the coordinate system of each object is randomly rotated by  $\theta$  ( $\sim 60^\circ \cdot \mathcal{N}(0, 1)$ ) in three axes and translated by  $[0.5, 0.5, 0.5] \cdot \mathcal{N}(0, 1)$ . As detailedly illustrated in Table R4 and Figure S7, the PSNR drops with All\*:22.16→21.20, Cat.: 20.65→19.58, particularly for apple and wallet, and the geometry also suffers except for bread, cake, and handbag, resulting in a generally more blurry and irregular-shaped appearance. We in-

fer that since xyz is fed into the network, the coordinates will implicitly store category-specific priors, *e.g.*, a specific sampled 3D location in canonical space will learn the prior of head or tail (other elements) of toy train. Thus the misalignment will tend to impair this learned variance of the rigid scene. In our experiment, we manually perform non-alignment in a regular mathematical manner, we believe this impairment will become more severe when applied to a naturally-unaligned coordinate system.

### C.3. Neural Surface Reconstruction

#### C.3.1 Dense-View Surface Reconstruction

**Implementation Details.** We use the publicly available code for NeuS [90] and VolSDF [103], and we use the code provided by the authors for Voxurf [95], training with for a standard number of iteration on each of them. For all the methods, we do not involve the mask loss as supervision. Each scene is trained on 100 views. We use the Chamfer Distance between the reconstructed surface and the ground truth mesh for evaluation. The distance is calculated in a normalized space (all coordinates lying within  $[-1, 1]$ ). We clip the distance by 0.1 to alleviate the huge effect of outliers. We will release the standard evaluation code.

**Qualitative Comparisons.** In the main text, we split the categories into three difficulty levels, namely *hard*, *medium*, and *easy*. Figure S8 shows some examples from each level. We observe that the “hard” examples usually suffer from dark and low-texture appearance (*e.g.*, the pan), concave geometry (*e.g.*, the vase and the kennel), and complex or thin structures (*e.g.*, the durian, the fork, and the toy train). The “medium” and “easy” cases usually have a simple geometry with proper texture. The wide exploration of geometry and textures of the dataset helps to provide a comprehensive

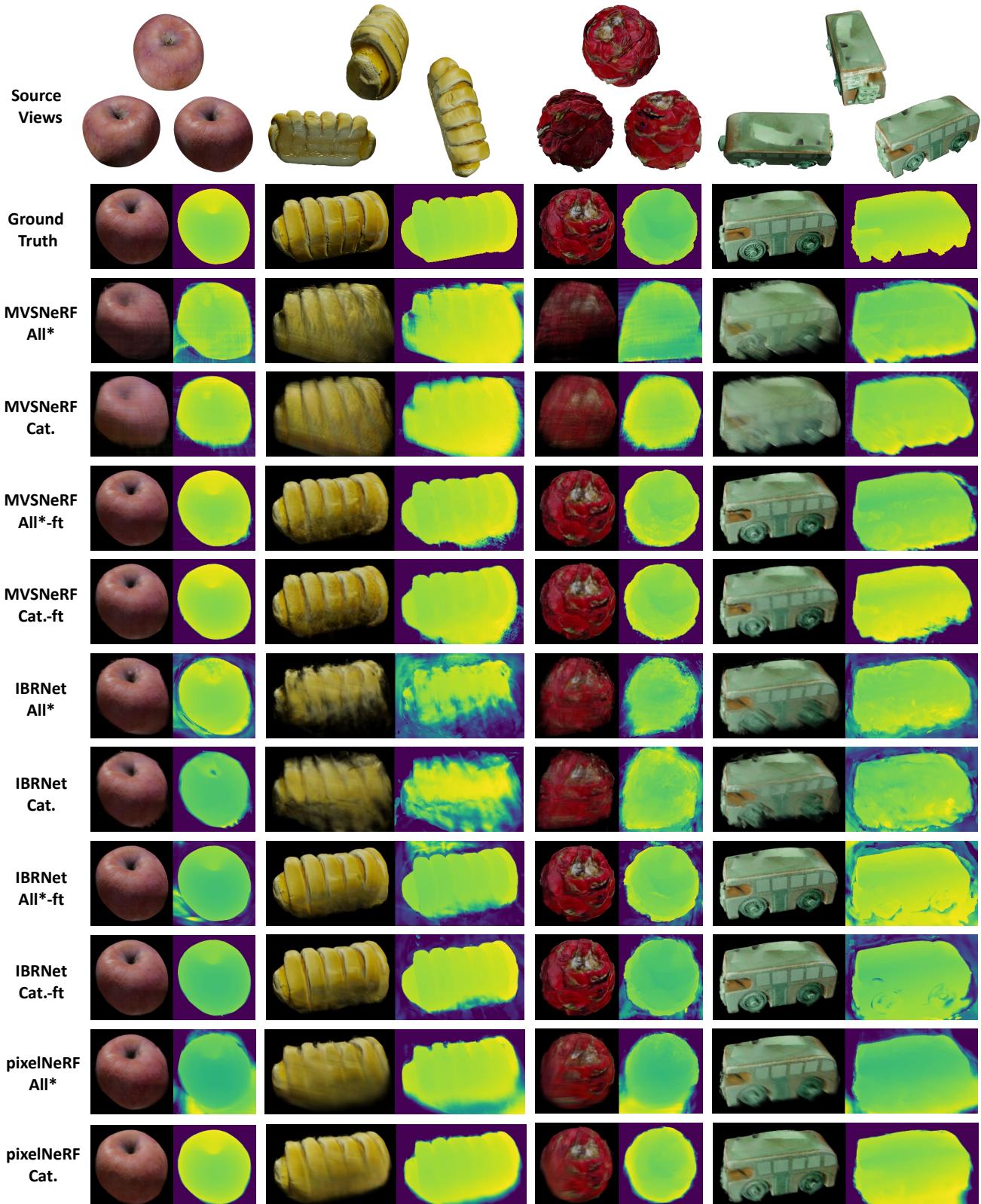


Figure S6. Qualitative comparisons of several cross-scene NVS methods in different scenes from our dataset.

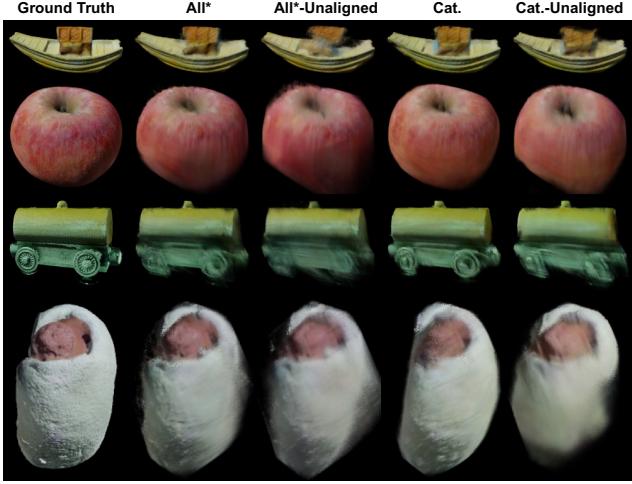


Figure S7. **Qualitative comparison of pixelNeRF-U and pixel-NeRF.** The former shows a more blurry and irregular-shaped appearance.

evaluation of different methods.

### C.3.2 Sparse-View Surface Reconstruction

**Implementation Details.** For NeuS [90] and MonoSDF [106], we use FPS sampling to sample 3 views from all the 100 views. We train 10k iterations for NeuS and 500 epochs for MonoSDF, both being reduced from the original setting due to the few-view input. For SparseNeuS [54], we fix the first three examples in each category as the testing set and skip them when training. We conduct FPS among the nearest 30 camera poses from a random reference view at inference time. The fine-tuning stage of SparseNeuS is not stable: the training usually collapses before convergence, and the issue also exists for the officially used DTU dataset. So we report the results via direct inference for all the experiments.

**Detailed Comparisons.** In Table 6 of the main text, we surprisingly find that NeuS can serve as a strong baseline under the sparse-view setting without bells and whistles. MonoSDF is enhanced by depth and normal estimations from pre-trained networks [22], and it claims a superior performance on DTU with only 3 views as input. However, MonoSDF does not seem to perform as well as NeuS in OmniObject3D.

As shown in Figure S9, the NeuS baseline with FPS sampling is especially good at dealing with thin structures: the wide-spread views together with the black backgrounds help to bound the geometry well. However, the depth estimation is especially inaccurate in these scenarios, which is probably caused by the gap between the training and testing images of the depth estimation network. Nevertheless, it shows great performance in some cases for maintaining

Table R5. **Sparse-view surface reconstruction results with a range of views.**

Method	Chamfer Distance $\times 10^3$ ( $\downarrow$ )			
	2 views	3 views	5 views	8 views
NeuS [90]	41.06	27.3	12.65	7.96
MonoSDF [106]	45.35	34.68	23.02	18.97

a coherent shape and adding some geometry details. It is an interesting problem to explore a robust usage of the estimated geometry cues under different circumstances.

**Sparse-view surface reconstruction with a range of view numbers.** In addition to the default setting of 3 views, we try a range of views (*i.e.*, 2, 3, 5, 8 views) with FPS sampling for NeuS [90] and MonoSDF [106], and the results are shown in Table. R5. For NeuS, we observe a significant improvement in accuracy as the view number increases from 2 to 8, but the 8-view setting (7.96) is still worse than the 100-view setting (6.09) with a clear margin. For MonoSDF, the improvement begins to slow down when lifting from 5 views to 8 views. This problem is probably due to the inaccurate depth guidance, as described above.

**View Selection Range for Cost Volume Initialization.** In MVSNeRF [11], due to occlusions, initialized local cost volume feature is inconsistent with large viewpoint changes, causing poor geometry extracted from the global density field. One naïve solution is to decrease the interval distance between source views. Although the constructed local feature will accordingly be more consistent as the occlusion region reduces, it will encode less source context. To make a trade-off between feature consistency and richness of encoded information, we conduct a comparison on how the extracted mesh will perform with the number of the nearest source views in FPS on 15 random categories from three levels of “difficulty”. We filter the categories with averaged CD  $\geq 0.04$ , whose geometries are too poor to rely on. Finally, we remain 5 classes as shown in Figure S10. The geometric quality shows a fluctuating trend of decreasing and then increasing with regard to the view range. As a result, we pick up “30” as a proper view selection range. Similarly, we find that “30” can also be applied to SparseNeuS [54] for cascaded geometry volume construction.

### C.4. 3D Object Generation

**Implementation Details.** We use the official code by GET3D [29] to train all the models. We prepare the multi-view image dataset by rendering 24 inward-facing multi-view images per object with Blender [17]. For the large subset with 100 categories, we train 7k iterations with MSE loss and Adam optimizer; we train 3k iterations on smaller subsets (*e.g.*, *furniture*, *fruits*, and *toys*).

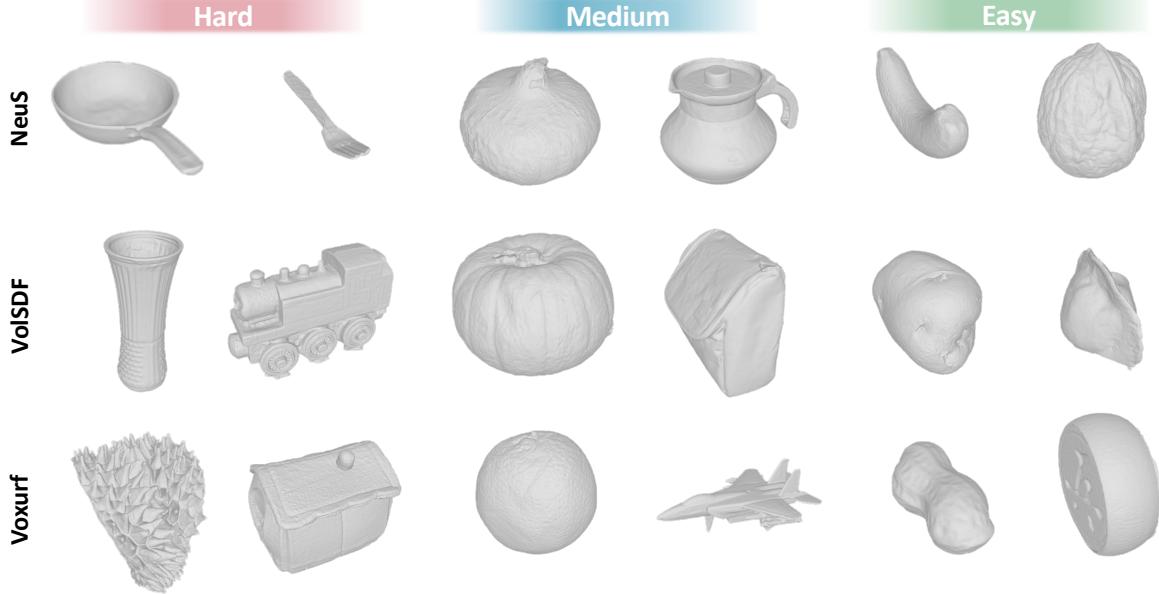


Figure S8. Examples from different difficulty levels in surface reconstruction.

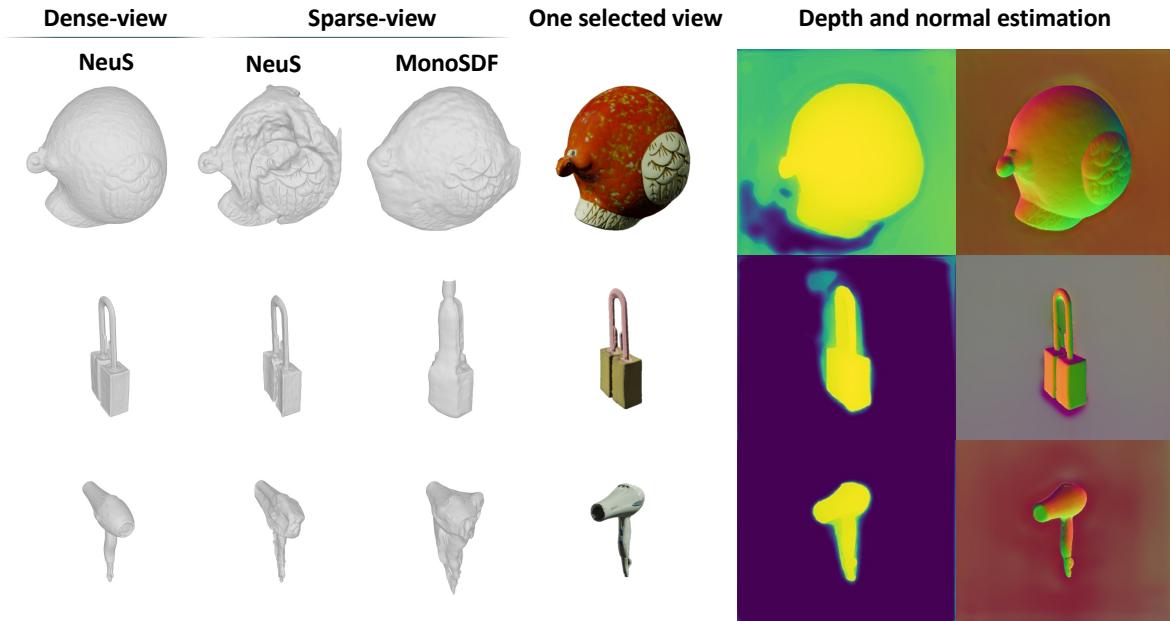


Figure S9. A comparison of sparse-view surface reconstruction between NeuS and MonoSDF. The estimated depth and normal maps used by MonoSDF are shown on the right.

**Additional Experimental Results and Discussions.** We study the semantic distribution in the main text, where we use KMeans to cluster 100 random categories into 8 groups, as shown in Figure S11. We can observe that Group 2 has the largest number of categories, while they suffer from a high inner-group divergence (*e.g.*, the peanut, handbag, mushroom, and hot dog). In contrast, Group 1 contains many fruits, vegetables, and some other categories that are similar in shape. The high inner-group similarity enables them to enhance the learning of each other, and Group 1

is finally able to dominate the generation distribution. The Group-level analysis reveals how cross-class relationships affect the generation distribution, which is a critical factor for generative models trained with large vocabulary datasets like OmniObject3D. We also provide the distribution of the four subsets used in this section in Figure S12.

Finally, we provide disentangled interpolation results in Figure S13 with geometry latent code and texture latent code, respectively. In the first row, the texture changes with a fixed shape, and the semantic changes accordingly. In

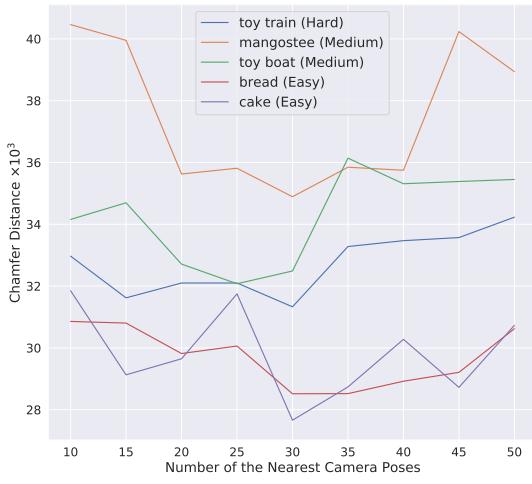


Figure S10. Geometric quality with regard to view selection range.

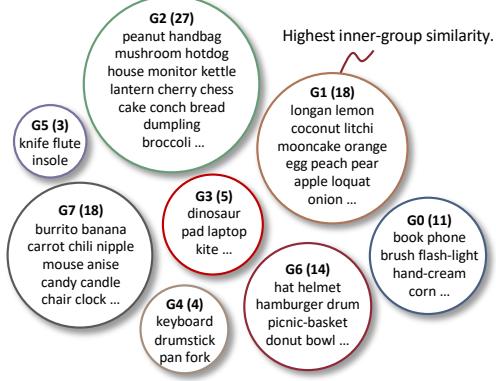


Figure S11. Categories in each group after the KMeans clustering. Categories in Group 1 are highly similar to each other, while those in Group 2 bear a high inner-group divergence.

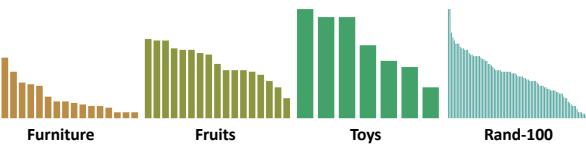


Figure S12. Distributions of the four subsets.

the second row, when the geometry changes, the texture is fixed at first while encountering a substantial change along with the geometry at the end. This indicates that the two factors are not fully disentangled, and the geometry code can sometimes affect the texture since the category, geometry, and texture are highly correlated with each other in the dataset. Meanwhile, we observe that complex textures (*e.g.*, the cover of a book) usually fail to be well generated, which is another challenging problem to be explored in the future.

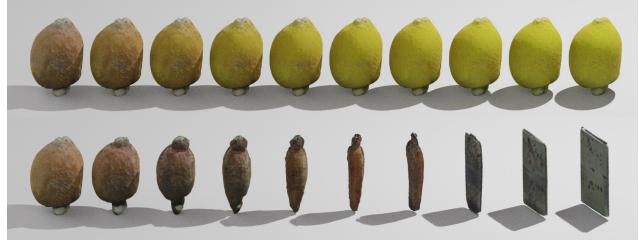


Figure S13. Shape Interpolation. In the first row, we keep the latent code of geometry fixed and interpolate the latent code of texture; in the second row, we keep the latent code of texture fixed and interpolate the latent code of geometry.

## References

- [1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjørholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision (IJCV)*, 120(2):153–168, 2016. [2](#), [6](#)
- [2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *Proceedings of the International Conference on Machine learning (ICML)*, pages 40–49, 2018. [3](#), [10](#)
- [3] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7822–7831, 2021. [3](#), [10](#)
- [4] Matan Atzmon, Niv Haim, Lior Yariv, Ofer Israelov, Haggai Maron, and Yaron Lipman. Controlling neural level sets. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, 2019. [10](#)
- [5] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. [3](#), [5](#), [10](#), [12](#)
- [6] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. [3](#), [10](#), [12](#)
- [7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. [11](#)
- [8] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

- and Pattern Recognition (CVPR)*, pages 5799–5809, 2021. 11
- [9] 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.org*, 1512.03012, 2015. 2, 4
- [10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3, 10
- [11] 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 (ICCV)*, pages 14124–14133, 2021. 3, 5, 7, 10, 13, 16
- [12] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, 2019. 3, 11
- [13] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees GM Snoek. Pointmixup: Augmentation for point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 330–345, 2020. 3, 9
- [14] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5939–5948, 2019. 3, 11
- [15] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5939–5948, 2019. 10
- [16] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21126–21136, 2022. 2
- [17] Blender Online Community. Blender - a 3d modelling and rendering package. 2018. 2, 4, 9, 10, 12, 16
- [18] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. *arXiv.org*, 2112.09648, 2021. 3, 10
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 2, 4
- [20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, volume 34, pages 8780–8794, 2021. 10
- [21] 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. *arXiv.org*, 2204.11918, 2022. 2
- [22] Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10786–10796, 2021. 16
- [23] Patrick Esser, Robin Rombach, and Bjorn Ommer. Tampering transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 10
- [24] Marco Forte and François Fleuret. F, b, alpha matting. *arXiv.org*, 2003.07711, 2020. 4, 9
- [25] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinzhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, 2022. 3, 5, 10, 12
- [26] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision (IJCV)*, 129(12):3313–3337, 2021. 2
- [27] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 402–411, 2017. 3, 10
- [28] 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. *arXiv.org*, 2209.11163, 2022. 3, 11
- [29] 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. In *Advances in Neural Information Processing Systems (NIPS)*, 2022. 3, 7, 16
- [30] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *Proceedings of the International Conference on Machine learning (ICML)*, pages 3809–3820, 2021. 5, 11
- [31] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenet: A style-based 3d-aware generator for high-resolution image synthesis. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 11
- [32] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021. 5, 11
- [33] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

- and Pattern Recognition (CVPR)*, pages 5356–5364, 2019. 2, 4
- [34] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14072–14082, 2021. 11
- [35] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato’s cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9984–9993, 2019. 3, 10
- [36] 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. 3
- [37] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017. 8
- [38] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 91–109, 2022. 10
- [39] Moritz Ibing, Gregor Kobsik, and Leif Kobbelt. Octree transformer: Autoregressive 3d shape generation on hierarchically structured sequences. *arXiv.org*, 2111.12480, 2022. 3, 11
- [40] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1261, 2020. 10
- [41] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 34, pages 852–863, 2021. 10
- [42] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 10
- [43] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 10
- [44] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4287–4297, 2021. 10
- [45] Siheyon Kim, Sanghyeok Lee, Dasol Hwang, Jaewon Lee, Seong Jae Hwang, and Hyunwoo J Kim. Point cloud augmentation with weighted local transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–557, 2021. 3, 9
- [46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv.org*, 1412.6980, 2014. 13
- [47] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981, 2020. 4
- [48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 4
- [49] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14809–14818, 2022. 2
- [50] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2028, 2020. 10
- [51] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8895–8904, 2019. 5, 11
- [52] 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 (CVPR)*, pages 7824–7833, 2022. 3, 10
- [53] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv.org*, 1906.07751, 2019. 10
- [54] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. *arXiv.org*, 2206.05737, 2022. 3, 7, 10, 16
- [55] Sebastian Lunz, Yingzhen Li, Andrew Fitzgibbon, and Nate Kushman. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv.org*, 2002.12674, 2020. 3, 10
- [56] Andrew Luo, Tianqin Li, Wen-Hao Zhang, and Tai Sing Lee. Surfgen: Adversarial 3d shape synthesis with explicit surface discriminators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16238–16248, 2021. 11
- [57] 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 (CVPR)*, pages 4460–4470, 2019. 3, 11
- [58] 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 (CVPR)*, pages 4460–4470, 2019. 10
- [59] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16190–16199, 2022. 3, 10
- [60] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 3, 5, 10, 11, 12
- [61] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy J. Mitra, and Leonidas J. Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv.org*, 1908.00575, 2019. 3, 10
- [62] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 2022. 3, 10
- [63] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *Proceedings of the International Conference on Machine learning (ICML)*, pages 7220–7229, 2020. 11
- [64] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, 2021. 11
- [65] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3504–3515, 2020. 10
- [66] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5589–5599, 2021. 3, 10
- [67] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, 2022. 11
- [68] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representa-
- tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 10
- [69] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019. 10
- [70] Dario Pavlo, Jonas Kohler, Thomas Hofmann, and Aurelien Lucchi. Learning generative models of textured 3d meshes from real-world images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13879–13889, 2021. 3, 11
- [71] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: deep learning on point sets for 3d classification and segmentation. *corr abs/1612.00593* (2016). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 3, 5, 9, 11
- [72] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017. 5, 11
- [73] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020. 4, 9
- [74] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021. 2, 3, 9, 10
- [75] Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *Proceedings of the International Conference on Machine learning (ICML)*, 2022. 3, 4, 5, 10, 11
- [76] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigima, 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 (ICCV)*, pages 2304–2314, 2019. 10
- [77] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2, 3, 4, 9
- [78] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 20154–20166, 2020. 11
- [79] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *arXiv.org*, 2206.07695, 2022. 11

- [80] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8430–8439, 2019. 4
- [81] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, 2019. 10
- [82] Edward J Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *Proceedings of the Conference on Robot Learning (CoRL)*, pages 87–96, 2017. 3, 10
- [83] Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1798–1808, 2021. 2
- [84] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5459–5469, 2022. 3, 10
- [85] Saeid Asgari Taghanaki, Jieliang Luo, Ran Zhang, Ye Wang, Pradeep Kumar Jayaraman, and Krishna Murthy Jatavallabhula. Robustpointset: A dataset for benchmarking robustness of point cloud classifiers. *arXiv.org*, 2011.11572, 2020. 3, 10
- [86] Briac Toussaint, Maxime Genisson, and Jean-Sébastien Franco. Fast Gradient Descent for Surface Capture Via Differentiable Rendering. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 1–10, 2022. 10
- [87] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1588–1597, 2019. 2, 3, 4, 10
- [88] Mukund Varma, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, and Madras. Is attention all nerf needs? *arXiv.org*, 2207.13298, 2022. 10
- [89] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490, 2022. 3, 10
- [90] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NIPS)*, volume 34, pages 27171–27183, 2021. 3, 6, 7, 10, 14, 16
- [91] 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 (CVPR)*, pages 4690–4699, 2021. 3, 5, 10, 13
- [92] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5):1–12, 2019. 3, 5, 9, 11
- [93] 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 (TIP)*, 13(4):600–612, 2004. 5
- [94] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29, 2016. 3, 10
- [95] Tong Wu, Jiaqi Wang, Xingang Pan, Xudong Xu, Christian Theobalt, Ziwei Liu, and Dahua Lin. Voxurf: Voxel-based efficient and accurate neural surface reconstruction. *arXiv.org*, 2208.12697, 2022. 3, 6, 10, 14
- [96] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 2, 3, 4, 10
- [97] Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 915–924, 2021. 5, 11
- [98] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3173–3182, 2021. 5, 11
- [99] Mutian Xu, Junhao Zhang, Zhipeng Zhou, Mingye Xu, Xiaojuan Qi, and Yu Qiao. Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 35, pages 3056–3064, 2021. 5, 11
- [100] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18430–18439, 2022. 11
- [101] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4541–4550, 2019. 3, 10
- [102] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition (CVPR)*, pages 1790–1799, 2020. 2
- [103] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NIPS)*, volume 34, pages 4805–4815, 2021. 3, 6, 10, 14
- [104] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 2492–2502, 2020. 10
- [105] 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 (CVPR)*, pages 4578–4587, 2021. 3, 5, 7, 10, 13, 14
- [106] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. In *Advances in Neural Information Processing Systems (NIPS)*, 2022. 3, 7, 10, 16
- [107] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6525–6534, 2021. 10
- [108] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 5
- [109] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5826–5835, 2021. 3, 10
- [110] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv.org*, 2110.09788, 2021. 11
- [111] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv.org*, 1801.09847, 2018. 4