

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

Tong Wu^{1,2}, Jiarui Zhang^{1,3}, Xiao Fu¹, Yuxin Wang^{1,4}, Jiawei Ren⁵, Liang Pan⁵,
 Wayne Wu¹, Lei Yang^{1,3}, Jiaqi Wang¹, Chen Qian¹, Dahua Lin^{1,2}, Ziwei Liu⁵

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



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 popular 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 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

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

rapid progress emerging in recent years. However, a majority of the technical approaches rely on unrealistic synthetic datasets [6, 19, 64] 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 [48] 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 [50]. Moreover, they do not provide textured meshes. GSO [16] has 1k scanned objects while covering only 17 household classes. AKB-48 [33] 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 [15], LVIS [25], and ShapeNet [6]), 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 [13], 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 OmniObject3D 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 diver-

Table 1. **A comparison between OmniObject3D and other commonly-used 3D object datasets.** R^{vis} denotes the ratio of the 1.2k LVIS [25] categories being covered.

Dataset	Real	Full Mesh	Video	# Objs	# Cats	R^{vis} (%)
ShapeNet [6]		✓		51k	55	4.1
ModelNet [64]		✓		12k	40	2.4
3D-Future [19]		✓		16k	34	1.3
ABO [12]		✓		8k	63	3.5
Toys4K [53]		✓		4k	105	7.7
CO3D V1 / V2 [48]	✓		✓	19 / 40k	50	4.2
DTU [1]	✓	✓		124	NA	0
ScanObjectNN [56]	✓			15k	15	1.3
GSO [16]	✓	✓		1k	17	0.9
AKB-48 [33]	✓	✓		2k	48	1.8
Ours	✓	✓	✓	6k	190	10.8

sity 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 [6, 53, 64], for example, ShapeNet [6] has 51,300 3D CAD models in 55 categories, and ModelNet40 [64] consists of 12,311 models in 40 categories. Recent works, e.g., 3D-FUTURE [19] and ABO [12], 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 [69] are photo-realistic datasets designed for multi-view stereo benchmarks, while they are small in scale and lack category annotations. ScanObjectNN [56] 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 [16] has 1,030 scanned objects with fine geometries and textures in 17 household items, and AKB-48 [33] focuses on robotics manipulation with 2,037 articulated object models in 48 articulated object categories. However, the relatively narrow semantic scope of GSO and AKB-48 hinders their applications for more general 3D research. CO3D v1 [48] contains 19,000 object-centric videos, while only 20% of them are annotated with accurate point clouds reconstructed by COLMAP [50], and they do not provide meshes or textures. In contrast, the proposed OmniOb-



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.

ject3D 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 [10, 30, 45, 60] studies the OOD corruptions and proposes standard corruption test suites [49, 55], while they fail to take account of OOD styles. Another branch of works [3, 48] evaluates the sim-to-real domain gap by training models on clean synthetic datasets [64] and testing them on noisy real-world test sets [56], 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) [40] 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 [4, 5, 39, 57] and efficiency [7, 18, 42, 54]. A branch of works [8, 27, 35, 48, 59, 71] 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 [14, 43, 58, 63, 70] 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 im-

ages. 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 [36] or taking advantage of the estimated geometry cues estimated by pre-trained networks [72]. 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 [20, 26, 37, 52, 62] extend 2D generation frameworks to 3D voxels with a high computational cost. Some others adopt different 3D data formulations, *e.g.*, point cloud [2, 41, 68, 74], octree [29], and implicit representations [11, 38]. However, it is non-trivial to generate complex and textured surfaces. Recent advances [9, 21, 44] explore the generation for textured 3D meshes, where GET3D [22] 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 [6, 15, 25, 31, 32, 51, 64]. 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 ¹ and Artec Eva 3D scanner ² 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 [13] 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 [75], with 2ⁿ ($n \in \{10, 11, 12, 13, 14\}$) 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. 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 [50], 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.

¹<https://www.einscan.com/>

²<https://www.artec3d.cn/>

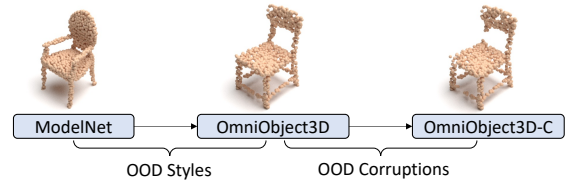


Figure 3. OmniObject3D provides the first clean real-world point cloud object dataset and allows fine-grained analysis on robustness to OOD styles and OOD corruptions. “-C”: corrupted by common corruptions described in [49]

We also develop a two-stage matting pipeline based on the U²Net [47] and FBA [17] 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 [6, 15, 25, 31, 32, 51, 64]. For example, it covers 85 categories in ImageNet [15] and 130 categories in LVIS [25], which leads to the highest R^{vis} in Table 1. Most of the categories are covered by the Open Images [31] 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 [64], for training and sets up two kinds of test sets for evaluation:

1) *Noisy real-world datasets*, *e.g.*, ScanObjectNN [56], 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 always noisy, making it impossible to analyze the two robustness challenges independently.

2) *Corrupted synthetic datasets*, *e.g.*, ModelNet-C [49], 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 cor-

Table 2. **Point cloud perception robustness analysis on OmniObject3D with different architecture designs.** Models are trained on the ModelNet-40 dataset, with OA_{Clean} to be their overall accuracy on the standard ModelNet-40 test set. OA_{Style} 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 [49]. Full results are presented in the supplementary materials.

	mCE† ↓	OA_{Clean} ↑	OA_{Style} ↑	mCE ↓
DGCNN [60]	1.000	0.926	0.448	1.000
PointNet [45]	1.422	0.907	0.466	0.969
PointNet++ [46]	1.072	0.930	0.407	1.066
RSCNN [34]	1.130	0.923	0.393	1.076
SimpleView [23]	1.047	0.939	0.476	0.990
GDANet [67]	<u>0.892</u>	0.934	<u>0.497</u>	0.920
PACConv [66]	1.104	0.936	0.403	1.073
CurveNet [65]	0.927	<u>0.938</u>	0.500	<u>0.929</u>
PCT [24]	0.925	0.930	0.459	0.940
RPC [49]	0.863	0.930	0.472	0.936

ruptions 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 [49] 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 [49] to measure the OOD-corruption robustness.

We benchmark ten state-of-the-art point cloud classification models in Table 2. We observe that 1) performance on a clean test set has little correlation with OOD-style robustness. For example, SimpleView [23] achieves the best OA_{Clean} but mediocre OA_{Style} ; 2) Advanced point grouping, *e.g.*, curve-based point grouping in CurveNet [65] and frequency-based point grouping in GDANet [67], are robust not only to OOD corruptions as pointed out in [49], 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 [49], 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. 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 [40]; 2) learning priors across scenes from our dataset to explore

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 [40]	34.01 / 3.46	0.953 / 0.029	0.068 / 0.061
mip-NeRF [4]	39.86 / 4.58	0.974 / 0.013	0.084 / 0.048
Plenoxels [18]	41.04 / 6.84	0.982 / 0.031	0.030 / 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 (↓)	$\mathcal{L}_1^{\text{depth}}$ (↓)
MVSNeRF [8]	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	0.076
IBRNet [59]	All*	19.39	0.569	0.399	0.423
	Cat.	19.03	0.551	0.415	0.290
	All*-ft.	26.89	0.792	0.215	0.081
	Cat.-ft.	25.67	0.760	0.238	0.099
pixelNeRF [71]	All*	22.16	0.692	0.342	0.109
	Cat.	20.65	0.676	0.348	0.195

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 [40], mip-NeRF [4], and a voxel-based system named Plenoxels [18] for comparison. As in Table 3, we find that Plenoxels achieve the best performance on average for PSNR, SSIM [61], and LPIPS [73]. 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 [71], IBRNet [59] and MVSNeRF [8] in Table 4. 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_{All*} is comparable to MVSNeRF_{Cat.}, IBRNet_{All*} and pixelNeRF_{All*} 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

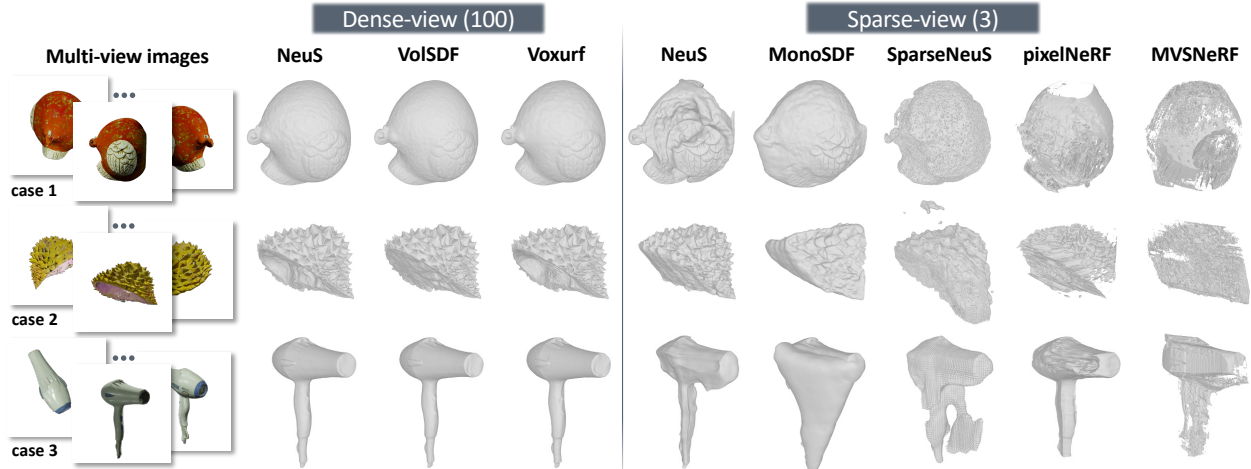


Figure 4. Neural surface reconstruction results for both dense-view and sparse-view settings.

Table 5. Dense-view surface reconstruction results.

Method	Chamfer Distance $\times 10^3$ (\downarrow)			
	Hard	Medium	Easy	Avg
NeuS [58]	9.26	5.63	3.46	6.09
VolSDF [70]	10.06	4.94	2.86	5.92
Voxurf [63]	9.01	4.98	2.58	5.49
Avg	9.44	5.19	2.97	5.83

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_{All*-ft} 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.

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

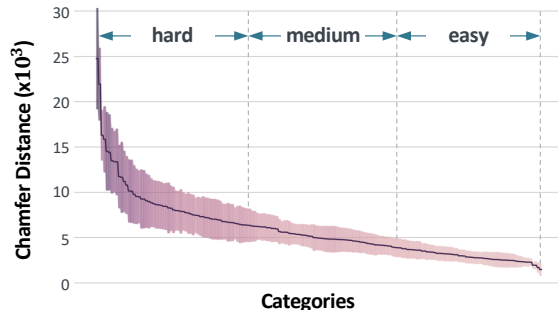


Figure 5. Performance distribution of dense-view surface reconstruction. The averaged results of the three methods is imbalanced. The colored area denotes a smoothed range of results.

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 [58] and VolSDF [70] are two well-known approaches that bridge neural volume rendering with implicit surface representation. We also involve a voxel-based method called Voxurf [63], 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 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

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

Method	Train	Chamfer Distance $\times 10^3$ (\downarrow)			
		Hard	Medium	Easy	Avg
NeuS [58]	Single	29.35	27.62	24.79	27.3
MonoSDF [72]	Single	35.14	35.35	32.76	34.68
SparseNeuS [36]	1 cat.	34.05	31.32	31.14	32.36
	10 cats.	30.75	30.11	28.37	29.87
	All cats.	26.13	26.08	22.13	25.00
	Easy	28.39	26.65	23.76	26.48
	Medium	27.38	26.66	23.08	25.87
MVSNeRF [8]	All cats.	56.68	48.09	48.70	51.16
	pixelNeRF [71]	63.31	59.91	61.47	61.56

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 [58] with sparse-view input; MonoSDF [72], which takes in geometry cues estimated by pre-trained models; SparseNeuS [36], a generic surface prediction pipeline that learns generalizable priors; pixelNeRF [71] and MVSNeRF [8] 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 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.

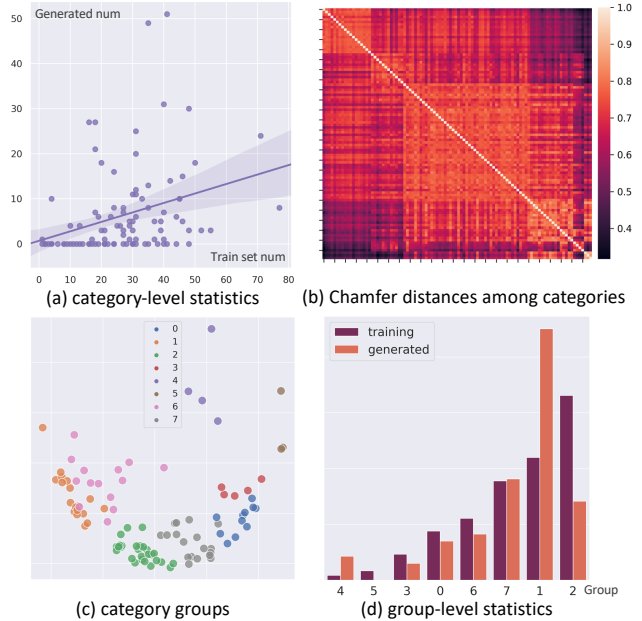


Figure 6. **The category distribution of the generated shapes.** (a) shows a weak positive correlation between the number of generated shapes and training shapes per category. (b) visualizes the correlation matrix among different categories by Chamfer Distance between their mean shapes. (c) visualizes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

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 [22]. 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 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



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.

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 and quality of the shapes, respectively; for texture, we adopt the widely-used FID [28]. 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^{ref} 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

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

Split	#Objs	#Cats	Cov (% \uparrow)	MMD (\downarrow)	FID (\downarrow)	FID^{ref}
Furniture	265	17	67.92	4.27	87.39	58.40
Fruits	610	17	46.72	3.32	105.31	87.15
Toys	339	7	55.22	2.78	122.77	41.40
Rand-100	2951	100	61.70	3.89	46.57	8.65

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^{ref} 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.

References

- [1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm 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
- [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
- [4] 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
- [5] 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
- [6] 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, 3, 4
- [7] 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
- [8] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnrf: 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
- [9] 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
- [10] 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
- [11] 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
- [12] 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
- [13] Blender Online Community. Blender - a 3d modelling and rendering package. 2018. 2, 4
- [14] 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
- [15] 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, 3, 4
- [16] 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
- [17] Marco Forte and François Pitié. F, b, alpha matting. *arXiv.org*, 2003.07711, 2020. 4
- [18] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong 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
- [19] 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
- [20] 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
- [21] 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
- [22] 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
- [23] 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
- [24] 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

- [25] 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, 3, 4
- [26] 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
- [27] 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
- [28] 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
- [29] Moritz Ibing, Gregor Kobsik, and Leif Kobbelt. Octree transformer: Autoregressive 3d shape generation on hierarchically structured sequences. *arXiv.org*, 2111.12480, 2022. 3
- [30] Sihyeon 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
- [31] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981, 2020. 3, 4
- [32] 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. 3, 4
- [33] 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
- [34] 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
- [35] 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
- [36] 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
- [37] 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
- [38] 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
- [39] 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
- [40] 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
- [41] 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
- [42] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multi-resolution hash encoding. *ACM Transactions on Graphics*, 2022. 3
- [43] 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
- [44] Dario Pavllo, 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
- [45] 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
- [46] 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
- [47] 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
- [48] 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

- [49] 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
- [50] 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, 4
- [51] 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. 3, 4
- [52] 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
- [53] 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
- [54] 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
- [55] 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
- [56] 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
- [57] 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
- [58] 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
- [59] 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
- [60] 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
- [61] 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
- [62] 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
- [63] 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
- [64] 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
- [65] 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
- [66] 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
- [67] 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
- [68] 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
- [69] 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
- [70] 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
- [71] 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

- [72] 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](#)
- [73] 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](#)
- [74] 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](#)
- [75] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv.org*, 1801.09847, 2018. [4](#)