RGBD Objects in the Wild:
Scaling Real-World 3D Object Learning from RGB-D Videos

Hongchi Xia[1*] Yang Fu[2*] Sifei Liu[3] Xiaolong Wang[2]
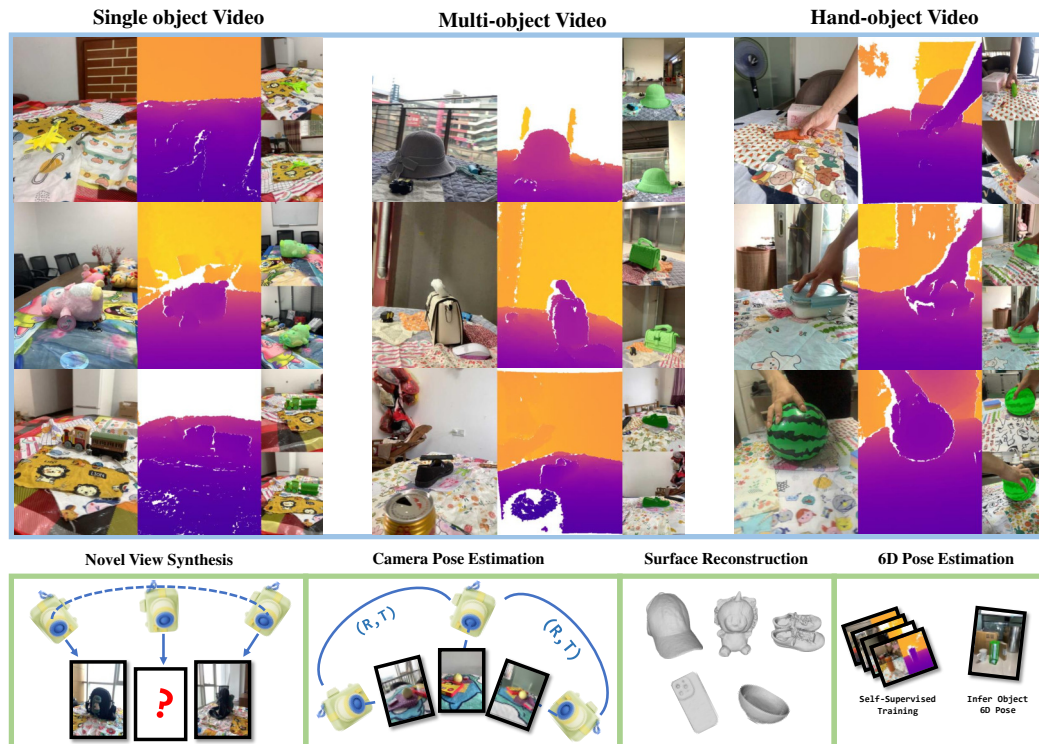[1]Shanghai Jiao Tong University [2]UC San Diego [3]NVIDIA

Figure 1. **WildRGB-D Dataset** contains almost 8500 recorded objects and nearly 20000 RGBD videos in 46 common categories with corresponding object masks and 3D point clouds.

Abstract

We introduce a new RGB-D object dataset captured in the wild called WildRGB-D. Unlike most existing real-world object-centric datasets which only come with RGB capturing, the direct capture of the depth channel allows better 3D annotations and broader downstream applications. WildRGB-D comprises large-scale category-level RGB-D object videos, which are taken using an iPhone to go around the objects in 360 degrees. It contains around 8500 recorded objects and nearly 20000 RGB-D videos across 46 common object categories. These videos are taken with diverse cluttered backgrounds with three setups to cover as many real-world scenarios as possible: (i) a single object in one video; (ii) multiple objects in one video; and (iii) an object with a static hand in one video. The dataset is annotated with object masks, real-world scale camera poses, and reconstructed aggregated point clouds from RGBD videos. We benchmark four tasks with WildRGB-D including novel view synthesis, camera pose estimation, object 6d pose estimation, and object surface reconstruction. Our experiments show that the large-scale capture of RGB-D objects provides a large potential to advance 3D object learning. Our project page is https://wildrgbd.github.io/.

1. Introduction

The recent advancement of computer vision has been largely relying on the scaling of training data [26, 53]. The same success in data-driven approaches has been recently adopted to 3D object modeling with new large 3D object-

---
*Equal contribution.

Figure 2. **The camera poses trajectory in WildRGB-D Dataset.** We visualize the corresponding camera in each scene of our dataset, showing that our dataset is featured in 360 degree full and dense multi-view camera poses.

centric dataset collection [25, 40, 52]. Most of the large datasets are synthetic 3D data [5, 14, 22, 76] and a mix of synthetic data and real-world object scans [16], given it is much less labor intensive for scaling by rendering from simulation. However, it remains a big challenge to apply the model trained in simulation data to the real world. This is not only because the synthetic data has less realistic texture and shape, but also due to it is very hard to model the cluttered background and the natural light comes with it in simulation.

To make deep learning with 3D objects applicable in the real world, researchers have made efforts to collect real-world multiview object data [2, 54]. For example, the CO3D dataset [54] contains 19K object videos across 50 categories. However, due to the lack of depth, they require the use of COLMAP [56] to provide 3D annotations, which only works for 20% of the collected videos. Collecting the depth channel part of the data is not only useful for more accurate 3D ground-truth annotations, but also provides very useful information for downstream applications such as object 6D pose estimation and novel view synthesis. The OmniObject3D dataset [75] provides both object videos and a separate scanning of the objects. However, the collected videos do not come with the depth channel inputs and they are mostly taken with clean backgrounds. The Wild6D dataset [23] is one of the few recent efforts to collect RGB-D object videos taken in the wild. However, it only contains 6 categories of data and covers relatively smaller ranges of object views.

In this paper, we propose to collect a new dataset that contains large-scale RGB-D object videos across diverse object categories and presented in the wild. Our dataset, namely WildRGB-D, covers 8500 tabletop objects across 44 categories in 20K videos. The videos are taken using iPhones to go around the objects in 360 degrees (see Figure 2 for visualization). Examples of the dataset are shown

in Figure 1. There are three types of videos: (i) *Single object video* where there is only one object presented on the table; (ii) *Multi-object video* where there are multiple objects presented at the same time; and (iii) *Hand-object video* where there is a static human hand grasping the object. More video types add variety, creating occlusion for objects in scenes, which are worthy study cases in some tasks. The collection of the WildRGB-D dataset not only considers the cluttered background in the real world, but also the common scenarios where the objects are occluded by human hands.

We perform automatic annotations for WildRGB-D. With RGB-D capturing, we can apply the Simultaneous Localization and Mapping [19, 58] (SLAM) algorithms, and exploit the RGB images and depth information from the depth sensor of mobile phones to reconstruct the *3D camera poses* in real-world scale and aggregated *3D point clouds*. Additionally, center *object segmentation masks* can be attained by bounding-box detection using text prompt of object category in Grounding-DINO [41], segmentation using Segment-Anything [33] and mask tracking using XMem [11], which are largely integrated into [12, 78].

To exploit the potential of our dataset, we benchmark it in four downstream tracks:

(i) Novel view synthesis. We evaluate various algorithms based on NeRF [47] which is optimized in a single scene, and generalizable NeRF which is trained in a category-level. With the help of depth information when training NeRFs, we can achieve consistently improved results. This offers a new platform for evaluating view synthesis approaches using RGB or RGB-D data.

(ii) Camera pose estimation. We adopt different pose estimation approaches [35, 86] to evaluate their capability of estimating relative camera poses in a sparse setting. We validate their generalizable ability through training on a partial of all categories and testing on unseen ones. We observe remarkable generalization performance on unseen

Dataset	Real	Multi-View	Depth Src.	3D GT	Video	# Cats	# Objs
ShapeNet [5]		none	CAD	mesh	none	55	51k
ModelNet [76]		none	CAD	mesh	none	40	12k
3D-Future [22]		none	CAD	mesh	none	34	16k
ABO [14]		none	CAD	mesh	none	63	8k
DTU [1]	✓	limited	COLMAP	mesh	RGB	N/A	124
CO3D [54]	✓	full	COLMAP	pcl	RGB	50	19k
MVImgNet [84]	✓	limited	COLMAP	pcl	RGB	238	220k
Objectron [2]	✓	limited	COLMAP	pcl	RGB	9	15k
GSO [18]	✓	none	scanner	mesh	none	17	1k
OmniObject3D [75]	✓	full	scanner	mesh	RGB	190	6k
Choi et al. [13]	✓	limited	sensor	mesh*	RGBD	9	2k
Wild6D [23]	✓	limited	sensor	pcl	RGBD	5	1.8k
Ours	✓	full	sensor	pcl	RGBD	44	8.5k

Table 1. **Comparison of WildRGB-D dataset with other 3D object dataset.** Some datasets don't provide video and we mark in "none". Some only cover partial angles, which is marked in "limited". Asterisk(*) means partial annotations. Depth Src. means where the depth information comes from, including CAD models, COLMAP, scanner devices and depth sensor in iPhones. pcl is the abbreviation of point cloud.

categories, which indicates our large-scale category-level dataset can serve as a training source for generalizable camera pose estimation.

(iii) Object surface reconstruction. We conduct object surface reconstruction in our dataset with RGB or RGB-D videos and object masks through Instant-NGP [49] and Neusfacto [85]. Results show that our depth information endow reconstruction with more accurate precision and SDF-based algorithm [85] performs better in this setting.

(iv) Object 6D pose estimation. We exploit the self-supervised algorithm in category-level object 6D pose estimation [87] with large-scale RGB-D images in our dataset and then evaluate the pre-trained model on the Wild6D [23] test set. We show our dataset can facilitate 6D pose estimation even without training labels, and we also study its generalization ability to the out-of-distribution test set.

2. Related Work

3D Object Datasets One representative kind of 3D object dataset is the 3D synthetic dataset, like ShapeNet [5] and ModelNet40 [76], which consist of category-level objects. 3D-FUTURE [22] and ABO [14] datasets are typical of higher quality mesh with textures. [34] and [64] introduce real-world category-specific object datasets that mainly focus on birds and chairs respectively. DTU [1] and BlendedMVS [80] are datasets designed for multi-view stereo and lack category-level classification. Objectron [2] provides rich annotations but only partial videos are fully 360 degree covered. CO3D [54] is a large-scale category-level dataset that annotates camera poses and depths with COLMAP [57], which doesn't provide depths in real-world scale, and MVImgNet [84] is also a dataset similar to CO3D. Pascal3D [77] contains real-

world 3D objects with pose annotations and CAD models in limited categories. Datasets collected with specialized hardware (scanner, dome, etc.) like GSO [18] and OmniObject3D [75] have more accurate 3D geometry models and rendered depths from them. However, they don't have RGBD wild object videos collected and lack real captured depths as well as background depths. In the aspect of RGBD object datasets, Wild6D [23] features RGBD image sequences and 6D pose annotations while lacking full 360 coverage and category types. Choi et al. [13] proposes RGBD object-centric datasets in 44 categories, but with limited camera annotations. As a comparison, our proposed WildRGB-D dataset contains almost 8500 recorded objects and nearly 20000 RGBD videos recorded all 360 degrees in 46 common categories from well-known 2D datasets, all with real-world scale camera poses and object mask annotations as well as aggregated point clouds. We present the detailed comparison in Tab. 1.

Neural Radiance Field and Object Surface Reconstruction Neural Radiance Field (NeRF) [47] is a kind of scene representation based on MLPs. It takes in sampled points along each ray and outputs the density and color of each point, which are then aggregated by volume rendering to synthesize views. [3, 4, 48, 66] propose new changes to the original NeRF to improve the visual quality and [7, 21, 39, 49, 60] advance the NeRF efficiency. In order to generalize the NeRF representations to other scenes, [6, 28, 43, 70, 83] learn latent 3D representations and priors from a bunch of existed scenes to help synthesize views across different scenes. Derived from original NeRF, [15, 50, 68, 72, 74, 81, 85] leverage Sign Distance Function (SDF) and represent the 3D scene by implicit surface, which has a more clear object boundary definition. Recently, 3D Gaussian Splating [31] has become a competitive alternative to NeRF. WildRGB-D dataset comprises various category-level objects and scenes on a large scale, which is suitable for novel view synthesis benchmarks and helps boost more mature reconstruction algorithms and generalizable 3D scene representations.

Camera Pose Estimation Given dense image views, mature algorithms of SfM [57] and SLAM [19] can estimate camera poses well by computing visual matches [44], verifying through RANSAC [20] and optimizing via bundle adjustment [63]. However, in a sparse camera view setting, camera pose estimation remains a challenging task. Some approaches [61, 71] leverage RNN or adopt auto-regression [79] targeting at SLAM applications. For category-agnostic sparse view camera pose estimation, [46, 55] adopt a direct regression approach. [29] estimates 6D pose upon training on synthetic dataset. Energy-based method [86] estimates distributions over relative rotations
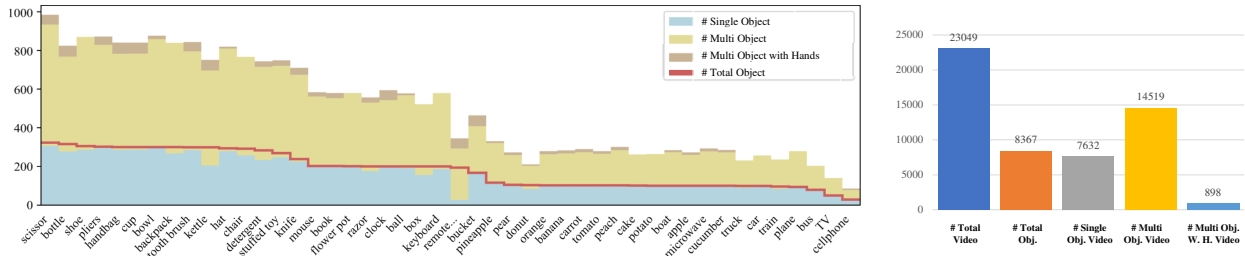
Figure 3. **Statistics of WildRGB-D Dataset** list the total and per-category number of objects and different types of videos.
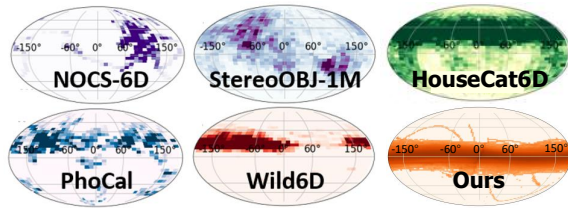

Figure 4. **Distribution visualization** of different kinds of Object 6D pose dataset and WildRGB-D dataset. We observe obvious disparity between Wild6D and Our dataset. Visualizations of [23, 30, 42, 67, 69] are from [30].

and [35] incorporates multi-view context to estimate camera 6D pose. Bundle adjustment gets learned after predictions in [59] to refine the estimated poses. In WildRGB-D dataset, with full 360-degree multi-view videos, the sparse view camera pose estimation setting is easily accessible, enabling our dataset to serve as a large-scale training database for these algorithms.

Object 6D Pose Estimation   In the setting of category-level 6D pose estimation, algorithms predict object poses in the same category and meet with various intra-class shapes. [67] predicts 6D pose using Umeyama algorithm [65] with NOCS map estimation and [8, 23, 62] follow up to learn more accurate NOCS representations. Other algorithms learn to estimate 6D pose through direct regression [9, 36], keypoint location estimations [38] and so on. Apart from supervised learning, self-supervision emerges due to the high cost of annotations. One approach [10, 24, 27, 82] is to adapt sim-to-real upon the pre-trained model on synthetic data. Another one [23, 45, 51] resorts to semi-supervised training. [87] proposes cycles across 2D-3D space learned correspondence, which enables training using only in-the-wild RGBD images without any annotations and is compatible with our dataset. With large-scale category-level RGBD wild object images for self-supervised learning, our dataset has the potential to boost future developments in this field.

3. The WildRGB-D Dataset

3.1. Data Collections, Processing, and Annotation

Data Collections   In order to collect RGBD video on a large scale expediently and economically, we record with the help of an iPhone front camera using Record3D [1] App and rotate the camera around the object so that full 360-degree views of objects are captured with RGB images and the corresponding depth images. Camera rotating speed is controlled equally by our collection setup to ensure less blur in videos. We select 46 common categories from well-known 2D datasets [17, 37]. We record three videos for every selected object, which are composed of single-object video, multi-object video, and hand-object video. Every recorded video has been checked and some are left behind due to poor SLAM camera pose estimation. Details of WildRGB-D Dataset are listed in Fig. 3.

Generating Camera Poses and 3D Point Cloud   Our WildRGB-D dataset has 3D annotations including camera poses in real-world scale, scene point clouds, and central object masks. In order to attain real-world scale camera poses, instead of relying on COLMAP [57] to first generate camera poses and then the depths using the poses, we generate more accurate camera poses with the mature RGBD Simultaneous Localization and Mapping [19, 58] (SLAM) algorithm, which leverages our captured depths. Additionally, it has the capability of exploiting the RGB images and depth information from the depth sensor of mobile phones to reconstruct the 3D camera poses in real-world scale, which is different from COLMAP depths, which are not in real-world scale. It enables us to simply project the depth images and gain aggregated 3D point clouds (see Figure 5). Then we manually check the quality of the aggregated 3D point cloud and exclude videos in which SLAM fails to get accurate camera poses. To increase the probability of getting correct SLAM results for each video, we adopt two kinds of SLAM algorithms including BAD SLAM [58] and SLAM implementation from Open3D [89], which increase our successful rate to over 90%.

---

[1]https://record3d.app/

Figure 5. **Point cloud reconstruction of objects in WildRGB-D Dataset.** We reconstruct the aggregated point cloud of the scene by leveraging existed 3D annotations of camera poses and depth images.

Method	PSNR↑/SD	SSIM↑/SD	LPIPS↓/SD	MAE↓/SD
NeRF [47]	23.03/1.50	0.690/0.072	0.390/0.075	0.306/0.109
NeRF (w mask)	34.65/4.44	0.943/0.077	0.031/0.032	0.029/0.019
Mip-NeRF 360 [4]	**23.84**/1.60	**0.762**/0.063	0.280/0.067	**0.185**/0.068
Mip-NeRF 360 (w mask)	35.60/4.51	**0.949**/0.077	0.024/0.025	**0.020**/0.015
Instant-NGP [49]	23.67/2.07	0.745/0.063	**0.257**/0.070	0.366/0.105
Instant-NGP (w mask)	**35.65**/5.20	0.946/0.077	**0.021**/0.031	0.068/0.074

Table 2. **Single-scene NVS results.** Average of four metrics w and w/o masks across all training dataset are reported with their standard deviation (SD).

Generating Central Object Masks We perform central object mask segmentation through a series of methods. Instead of the classic PointRend [32] algorithm, we leverage the novel segmentation tool Segment-Anything (SAM) [33]. We attain the prompt for SAM using Grounding-DINO [41], which generates a bounding box for SAM according to the category text prompt. After attaining the mask segmentation of the first frame in the video, XMem [11] is applied to track the mask in the video. The masking pipeline is largely integrated into [12, 78].

3.2. Statistics and Distribution

In WildRGB-D dataset collections, we recorded 8500 objects and 3 videos for each one. After excluding those SLAM-failed videos, we have 8367 objects in 23049 videos in our dataset (maintaining rates are 99.3%/91.0%). The selected videos contain 33.1% single object videos, 63.0% multi-object videos, and 3.9% hand-object videos. Details of WildRGB-D dataset are listed in Fig. 3.

4. Experiments

4.1. Novel View Synthesis

In this section, we conduct multiple experiments towards methods concerning novel view synthesis (NVS) in the following three scenarios: 1) **Single-Scene NVS**, where we train NeRF-based methods [4, 47, 49] on a single scene with only RGB image sequence. 2) **Cross-Scene NVS**, where we learn category-level scene representations to generalize

Method	Level	PSNR↑/SD	SSIM↑/SD	LPIPS↓/SD	MAE↓/SD
Pixel-NeRF [83]		20.28/0.65	0.645/0.043	0.495/0.074	**0.355**/0.120
MVSNeRF [6]	Easy	19.95/1.00	0.663/0.036	**0.351**/0.066	0.370/0.100
IBRNet [70]		**20.93**/0.98	**0.711**/0.031	0.395/0.153	-
Pixel-NeRF [83]		18.76/0.50	0.572/0.064	0.534/0.047	**0.299**/0.057
MVSNeRF [6]	Middle	18.75/0.74	0.601/0.069	0.363/0.036	0.345/0.102
IBRNet [70]		**19.77**/1.01	**0.663**/0.071	**0.362**/0.063	-
Pixel-NeRF [83]		17.23/0.66	0.521/0.035	0.624/0.054	**0.383**/0.121
MVSNeRF [6]	Hard	17.13/0.89	0.564/0.043	**0.425**/0.045	0.502/0.260
IBRNet [70]		**17.92**/1.12	**0.614**/0.056	0.439/0.069	-

Table 3. **Cross-scene NVS results.** Average of four metrics across all categories in training dataset are reported. We report metrics of three difficulty level respectively. Entries marked in - are not provided.

Method	PSNR↑/SD	SSIM↑/SD	LPIPS↓/SD	MAE↓/SD
Instant-NGP [49]	23.67/2.07	0.745/0.063	0.257/0.070	0.366/0.105
Instant-NGP (depth sup.)	**24.60**/2.13	**0.759**/0.062	**0.239**/0.066	**0.108**/0.057
Pixel-NeRF [83]	18.53/1.21	0.568/0.067	0.556/0.073	0.336/0.099
Pixel-NeRF (depth sup.)	**19.10**/1.21	**0.605**/0.060	**0.499**/0.064	**0.147**/0.087
MVSNeRF [6]	18.43/1.30	0.600/0.065	0.381/0.054	0.400/0.182
MVSNeRF (depth sup.)	**18.44**/1.29	0.600/0.065	0.381/0.054	**0.397**/0.186

Table 4. **Depth Supervised NVS and depth estimation results.** Average of four metrics w and w/o depth supervision across all training dataset are reported with their standard deviation (SD).

into other scenes with Generalizable NeRFs [6, 70, 83]. 3) **Depth Supervised NVS**, where we conduct NVS experiments with depth image priors in our dataset to study the potential that depth information will endow to NVS tasks.

Single-Scene NVS We select ten scenes from each category and uniformly sample images as validation split. We choose NeRF [47], Mip-NeRF 360 [4] and Instant-NGP [49] for evaluations. Results are shown in Tab. 2. We report the average PSNR, SSIM [73], LPIPS [88] and rendering depths Mean Average Error (MAE) compared with our sensor-collected depths across all categories. We also report metrics only related to the NVS quality of central objects using object masks. Results show that Mip-NeRF 360 and Instant-NGP outperform original NeRF in terms of visual quality metrics. NeRF-based methods perform better

Figure 6. **Novel view synthesis visualization** of different kinds of NeRF methods: NeRF [47], Mip-NeRF 360 [4] and Instant-NGP [49].

Figure 7. **Relpose++ [35] pair-wise evaluation visualization.** We show every image pair with its relative rotation predicted by Rel-pose.

when we only concern with the recovery of central objects under object masks. What's more, Mip-NeRF 360 performs best in learning single-scene geometry. Visualization can be found in Figure 6. In brief, our dataset offers extensive categories and scenes for in-depth NVS experiments.

Cross-Scene NVS Apart from single-scene optimizations, we also evaluate Generalizable NeRFs: Pixel-NeRF [83], MVSNeRF [6] and IBRNet [70] in the cross-scene setting. For each category in our dataset, we select the same test scenes as single-scene NVS experiments and train in the remaining scenes of the same category to learn per-category latent representations. We divide the 46 categories into three difficulty levels and report the average metrics of each level. For evaluation, we use three source views to synthesize novel views. We report PSNR, SSIM and LPIPS to measure visual quality and depth MAE to measure the learned geometry quality. From Tab. 3, we observe that IBRNet outperforms in all three difficulty levels in terms of visual quality. Additionally, learned geometry quality isn't highly correlated with the rendering visual quality in novel views. To sum up, our dataset provides great potential in learning category-level cross-scene NVS methods.

Depth Supervised NVS We also study the influences that depth supervision brings. We choose Instant-NGP [49] in single-scene NVS methods and both Pixel-NeRF [83] and MVSNeRF [6] in cross-scene NVS methods. Our experiment results in Tab. 4 prove that depth supervision is beneficial for these methods to learn better representations. In our experiment setting, we add L1 depth loss to every algorithm and choose the best-performance depth loss weight
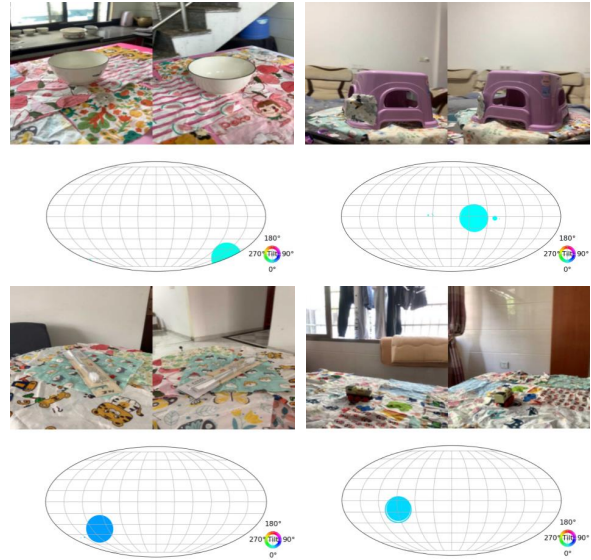
for them. Compared with conducting NVS tasks without depths, the performances of both Instant-NGP and Pixel-NeRF get boosted when training with depth loss. As for experiments of MVSNeRF, since we've already added depth information in the original training as the guidance in building rays, it turns out that the improvements when training with extra added depth loss are limited. In a nutshell, with depth priors, both single-scene NVS and cross-scene NVS methods learn better generalization capabilities, boosting NVS more accurate and generalizable.

4.2. Camera Pose Estimation

In this section, we benchmark two data-driven methods Rel-Pose [86] and RelPose++ [35] for inference of the relative camera poses from multi-view images in a sparse setting. Leveraging the given annotations of camera poses and large-scale category-level video in our dataset, we aim to learn generalizable viewpoint inference capability from training-seen categories to unseen ones. Since WildRGB-D dataset has a full and dense 360-degree camera trajectory, we can provide both a large-scale database and various view settings to assist training. In our experiments, we divided totally 46 categories into training and testing categories. We also hold some videos in training categories for evaluation. We adopt evaluation settings described in [35, 86] and report results in Tab. 5 and Tab. 6. We observe that these two methods can generalize well to other scenarios both in known categories and unseen categories since the relative rotation estimation errors are in a reasonable range (also see Figure 7 for visualization). However,
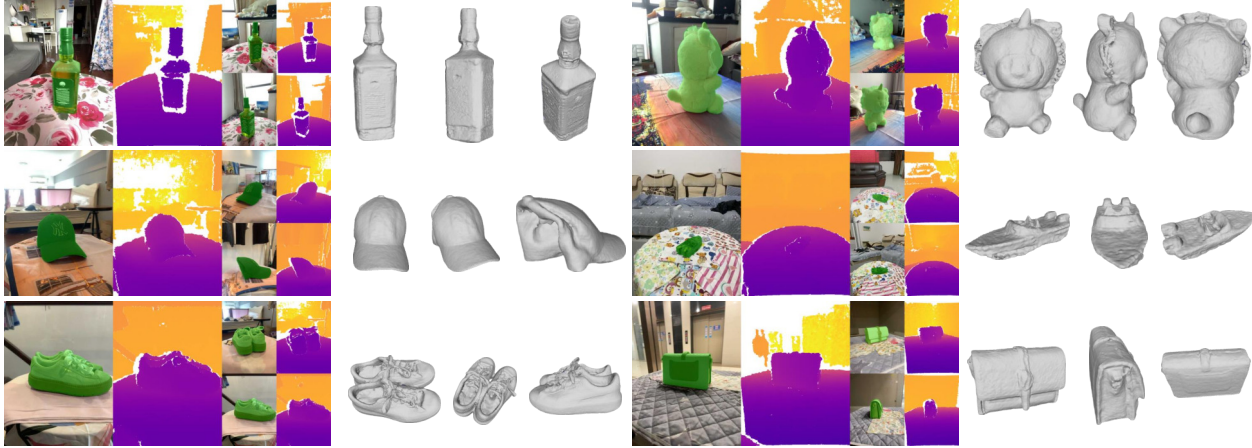
Figure 8. **Visualization of RGBD reconstruction surface** from Neusfacto [85]. Original RGBD image samples are listed on the left and multi-view reconstructed surface is on the right for each example.

the error of translation prediction is comparatively large in RelPose++, which still poses challenges in this field. To sum up, WildRGB-D dataset can serve as large-scale training sources for generalizable camera pose estimation algorithms to achieve remarkable results.

4.3. RGBD Object Surface Reconstruction

In our experiment setting of object surface reconstruction, algorithms need to utilize RGBD image sequence and central object masks to reconstruct the surface mesh of the central object. Reconstruction without depths is also evaluated for comparison. For evaluation of reconstruction quality, we calculate the Chamfer Distance between the reconstruction mesh and aggregated object point cloud which is derived from object-masked depth images. Ten single object scenes are selected in each category for evaluations of Instant-NGP [49] and Neusfacto [85]. From the results in Tab. 7, we observe that reconstruction is better with depth priors. Additionally, the performance of Neusfacto with RGBD is superior to Instant-NGP, which shows that depths help the sdf-based method Neusfacto learn the correct object boundary and boost the performance more compared with Instant-NGP. The average deviation is high due to the varied reconstruction qualities across different categories in the dataset. Visualization of Neusfacto RGBD reconstructions are shown in Figure 8. In brief, our dataset provides an RGBD object reconstruction evaluation track, boosting the development of more mature algorithms in this field.

4.4. RGBD 6D Object Pose Estimation

We explore our dataset in self-supervised training for 6D pose estimation. We adopt the algorithm proposed in [87], which leverages category shape prior and learns by matching the correspondence between images and shapes. In our experiment, we evaluate the trained model on Wild6D [23]
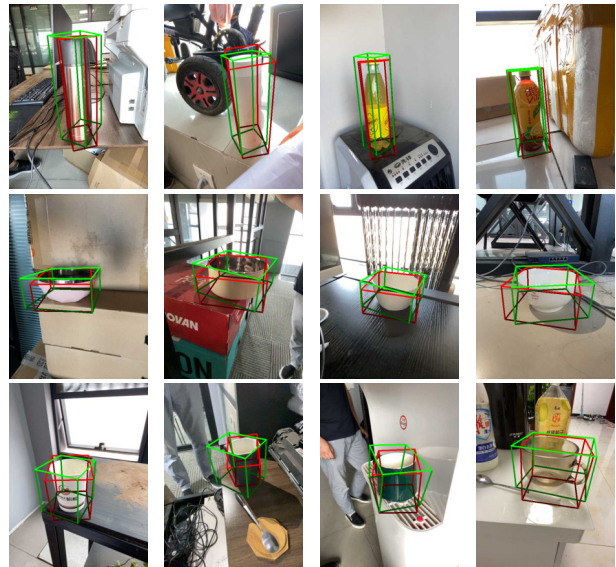

Figure 9. **Object 6D pose estimation visualization.** We visualize the predicted category-level 6D pose on three categories in Wild6D [23] test set (bottle, bowl, and mug) using models that **only** perform self-supervised training on the corresponding category of WildRGB-D Dataset. The ground truth bounding boxes are colored in green, and the predicted bounding boxes are in red.

test set. Three different settings concerning training sets are adopted: 1) Wild6D training set; 2) WildRGB-D dataset; 3) Wild6D training set + WildRGB-D dataset. Common categories in these two datasets are selected for self-supervised training and evaluation. Results in Tab. 8 show that in the case of an out-of-distribution setting, where we trained only on our dataset and evaluated on a different dataset, some of the metrics are decreased in our experiments. This is mainly due to the different distribution of camera rotations in these

Eval. Type	Categories	Metrics	#Frames 3	5	10	20
MST	seen	<15 deg.	57.4	55.1	51.4	47.4
		<30 deg.	82.1	79.8	77.2	74.0
	unseen	<15 deg.	38.4	37.7	36.6	35.2
		<30 deg.	62.5	61.8	60.4	59.0
Coord.Asc.	seen	<15 deg.	69.3	69.3	69.8	69.3
		<30 deg.	85.3	85.3	85.4	85.3
	unseen	<15 deg.	46.0	46.2	46.9	46.5
		<30 deg.	66.4	67.0	67.2	67.1
Sequential	seen	<15 deg.	51.9	45.1	36.0	26.9
		<30 deg.	78.3	72.5	61.9	49.3
	unseen	<15 deg.	34.9	31.0	25.1	18.3
		<30 deg.	59.1	54.4	46.4	37.1

Table 5. **RelPose [86] camera evaluation results.** We follow three evaluation types (MST, Coord.Asc., Sequential) propsed in [86] and report the average percent of rotation prediction errors in degrees both in training-seen categories and unseen ones.

Eval. Type	Categories	Metrics	#Frames 2	3	5	8
Pairwise	seen	<15 deg.	69.6	68.3	67.3	66.6
		<30 deg.	86.5	87.4	87.2	86.8
	unseen	<15 deg.	53.4	52.5	52.5	52.3
		<30 deg.	74.1	74.5	75.3	75.4
Coord.Asc.	seen	<15 deg.	70.4	71.5	71.9	71.6
		<30 deg.	86.7	87.9	88.5	88.7
	unseen	<15 deg.	52.9	55.4	54.9	54.8
		<30 deg.	73.9	75.7	76.3	76.8
Cam.Center	seen	<0.2 SS	100.0	29.8	12.6	5.9
		<0.3 SS	100.0	43.8	23.6	13.3
	unseen	<0.2 SS	100.0	30.5	12.4	5.6
		<0.3 SS	100.0	44.3	23.2	12.5
Cam.Trans.	seen	<0.2 SS	22.3	11.9	4.8	2.2
		<0.3 SS	30.1	20.4	10.8	6.2
	unseen	<0.2 SS	21.8	12.0	5.7	2.7
		<0.3 SS	29.8	20.9	12.0	6.4

Table 6. **RelPose++ [35] camera evaluation results.** We follow four evaluation types (Pairwise, Coord.Asc., Cam.Center, Cam.Trans) proposed in [35] and report the average percent of rotation prediction errors in degrees both in training-seen categories and unseen ones. Notes: SS means scene scale defined in [35].

RGB		RGBD	
Instant-NGP [49]	Neusfacto [85]	Instant-NGP [49]	Neusfacto [85]
45.91/64.01	88.92/89.94	28.46/29.28	**25.83**/34.07

Table 7. **RGBD object surface reconstruction results.** Average of chamfer distance with standard deviation across selected categories are reported (Average/SD).

two datasets (visualized in Fig. 4), where Wild6D doesn't cover full 360 degrees and WildRGB-D dataset covers a larger pitch angle range in object 6D poses. However, we

Category	Datasets	$IOU_{0.25}$	$IOU_{0.5}$	5 deg. 2cm	5 deg. 5cm	10 deg. 2cm	10 deg. 5cm
Bottle	Wild6D	93.2	85.2	71.3	**79.4**	79.8	90.9
	ROW	**93.3**	70.9	34.1	48.8	47.9	78.8
	Wild6D+ROW	**93.3**	**85.8**	**71.9**	78.6	**81.7**	**91.7**
Bowl	Wild6D	98.3	90.4	**66.1**	**70.0**	86.8	**94.6**
	ROW	98.3	**91.8**	33.8	35.7	86.3	93.5
	Wild6D+ROW	**98.4**	**91.8**	40.3	42.0	**87.5**	93.7
Mug	Wild6D	89.0	59.2	0.0	0.0	0.0	**0.1**
	ROW	89.1	**61.9**	0.0	0.0	0.0	**0.1**
	Wild6D+ROW	**89.3**	50.2	0.0	0.0	0.0	0.0

Table 8. **Self-Supervised 6D Pose Estimation results.** The evaluation results on Wild6D test dataset under three different settings in bottle, bowl and mug categories.

still witness some improvements in evaluations. Training with WildRGB-D dataset benefits in IOU evaluations and joint-dataset training improves rotation+translation evaluation in particular categories. Visualization of 6D pose estimation in Wild6D test set using models that only train in our dataset can be found in Figure 9, To summarize, our dataset provides large-scale category-level RGBD images sequences, serving as ample unsupervised training data, which has the potential to boost more accurate 6D pose estimation in the future.

5. Conclusion

The object-centric datasets in the computer vision community have mostly focused on RGB videos, while practical applications often involve depth as inputs or for better annotations. We collect the largest object-centric RGB-D video dataset WildRGB-D , where all videos are captured in cluttered scenes. It is composed of category-level RGB-D object videos taken using iPhones around the objects in 360 degrees, which contains around 8500 recorded objects and nearly 20000 RGB-D videos across 46 common object categories with three setups covering most scenarios. The dataset is well annotated with object masks, real-world scale camera poses, and reconstructed aggregated point clouds from RGBD videos. We set up four evaluation tracks with WildRGB-D, showing that the large-scale capture of RGB-D objects provides a large potential to advance 3D object learning. The current dataset does not come with annotations of the object 6D pose, which requires further crowd-sourcing effort. It will be one of our future efforts to collect this annotation for supervised training methods as well as evaluation. We are committed to releasing our dataset and evaluation code.

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*, 120:153–168, 2016. 3

[2] 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*, pages 7822–7831, 2021. 2, 3

[3] 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 *ICCV*, 2021. 3

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 3, 5, 6

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 3

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3, 5, 6

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 3

[8] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020. 4

[9] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021. 4

[10] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 139–156. Springer, 2020. 4

[11] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 2, 5

[12] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything, 2023. 2, 5

[13] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A large dataset of object scans. *arXiv preprint arXiv:1602.02481*, 2016. 3

[14] 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*, pages 21126–21136, 2022. 2, 3

[15] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6260–6269, 2022. 3

[16] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects, 2022. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[18] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 3

[19] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006. 2, 3, 4

[20] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[21] 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*, pages 5501–5510, 2022. 3

[22] 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*, 129:3313–3337, 2021. 2, 3

[23] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset, 2022. 2, 3, 4, 7

[24] Ge Gao, Mikko Lauri, Yulong Wang, Xiaolin Hu, Jianwei Zhang, and Simone Frintrop. 6d object pose regression via supervised learning on point clouds. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3643–3649. IEEE, 2020. 4

[25] 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*, 2022. 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[27] Yisheng He, Haoqiang Fan, Haibin Huang, Qifeng Chen, and Jian Sun. Towards self-supervised category-level object pose and size estimation. *arXiv preprint arXiv:2203.02884*, 2022. 4

[28] 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

[29] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. *arXiv preprint arXiv:2212.04492*, 2022. 3

[30] HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, Daniel Roth, Nassir Navab, and Benjamin Busam. Housecat6d – a large-scale multi-modal category level 6d object pose dataset with household objects in realistic scenarios, 2023. 4

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 3

[32] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 5

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 5

[34] Joseph J Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing ikea objects: Fine pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2992–2999, 2013. 3

[35] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 2, 4, 6, 8

[36] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021. 4

[37] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[38] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Single-stage keypoint-based category-level object pose estimation from an rgb image. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1547–1553. IEEE, 2022. 4

[39] Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, and Shenlong Wang. Neurmips: Neural mixture of planar experts for view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15702–15712, 2022. 3

[40] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 5

[42] Xingyu Liu, Shun Iwase, and Kris M. Kitani. Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation, 2022. 4

[43] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 3

[44] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 3

[45] Fabian Manhardt, Gu Wang, Benjamin Busam, Manuel Nickel, Sven Meier, Luca Minciullo, Xiangyang Ji, and Nassir Navab. Cps++: Improving class-level 6d pose and shape estimation from monocular images with self-supervised learning. *arXiv preprint arXiv:2003.05848*, 2020. 4

[46] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks. In *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18*, pages 675–687. Springer, 2017. 3

[47] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 5, 6

[48] 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*, pages 16190–16199, 2022. 3

[49] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 2022. 3, 5, 6, 7, 8

[50] 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, pages 5589–5599, 2021. 3

[51] Wanli Peng, Jianhang Yan, Hongtao Wen, and Yi Sun. Self-supervised category-level 6d object pose estimation with deep implicit shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2082–2090, 2022. 4

[52] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017. 2

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[54] 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*, pages 10901–10911, 2021. 2, 3

[55] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 3

[56] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[57] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3, 4

[58] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019. 2, 4

[59] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21349–21359, 2023. 4

[60] 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*, pages 5459–5469, 2022. 3

[61] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3

[62] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 4

[63] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu,*

Greece, September 21–22, 1999 Proceedings, pages 298–372. Springer, 2000. 3

[64] Shubham Tulsiani, Abhishek Kar, Joao Carreira, and Jitendra Malik. Learning category-specific deformable 3d models for object reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):719–731, 2016. 3

[65] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 4

[66] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 3

[67] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 4

[68] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[69] Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects, 2022. 4

[70] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 3, 5, 6

[71] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2043–2050. IEEE, 2017. 3

[72] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction, 2023. 3

[73] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[74] 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 preprint arXiv:2208.12697*, 2022. 3

[75] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 803–814, 2023. 2, 3

[76] 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 conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2, 3

[77] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 3

[78] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 2, 5

[79] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1281–1292, 2020. 3

[80] 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*, pages 1790–1799, 2020. 3

[81] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 3

[82] Yang You, Ruoxi Shi, Weiming Wang, and Cewu Lu. Cppf: Towards robust category-level 9d pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6866–6875, 2022. 4

[83] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3, 5, 6

[84] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images, 2023. 3

[85] Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. Sdfstudio: A unified framework for surface reconstruction, 2022. 3, 7, 8

[86] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pages 592–611. Springer, 2022. 2, 3, 6, 8

[87] Kaifeng Zhang, Yang Fu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Self-supervised geometric correspondence for category-level 6d object pose estimation in the wild. *arXiv preprint arXiv:2210.07199*, 2022. 3, 4, 7

[88] 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 conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[89] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 4