HouseCat6D - A Large-Scale Multi-Modal Category Level 6D Object Perception Dataset with Household Objects in Realistic Scenarios

HyunJun Jung[1,*], Shun-Cheng Wu[1,*], Patrick Ruhkamp[1,*], Guangyao Zhai[1,2,*,†], Hannah Schieber[1,3,*], Giulia Rizzoli[4], Pengyuan Wang[1], Hongcheng Zhao[1], Lorenzo Garattoni[5], Sven Meier[5], Daniel Roth[1], Nassir Navab[1], Benjamin Busam[1,2,6]

[1] Technical University of Munich [2] Munich Center for Machine Learning
[3] FAU Erlangen-Nürnberg [4] University of Padova [5] Toyota Motor Europe [6] 3dwe.ai

https://sites.google.com/view/housecat6d

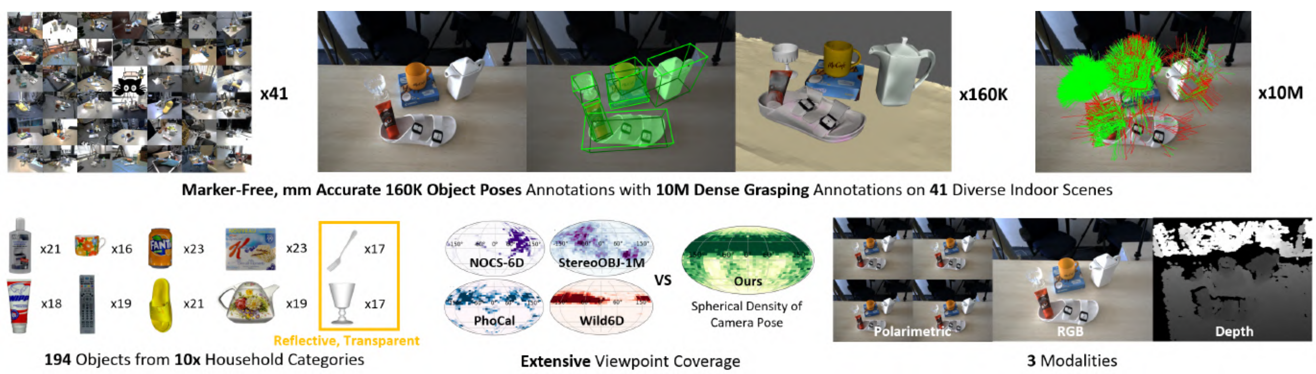


Figure 1. HouseCat6D is a multi-modal category level 6D object pose and grasping dataset with highly diverse household object categories of different photometric complexity and a high number of varying scenes covering large viewpoint distributions. It comprises room-scale high-quality camera trajectories and object poses without markers in realistic scenarios including occlusions as well as dense grasping pose annotation. Data includes synchronized RGB, depth from active stereo, and polarimetric RGB+P images in scenes comprising objects without texture, strong reflections, or translucency.

Abstract

*Estimating 6D object poses is a major challenge in 3D computer vision. Building on successful instance-level approaches, research is shifting towards category-level pose estimation for practical applications. Current category-level datasets, however, fall short in annotation quality and pose variety. Addressing this, we introduce **HouseCat6D**, a new category-level 6D pose dataset. It features **1**) multi-modality with Polarimetric RGB and Depth (RGBD+P), **2**) encompasses 194 diverse objects across 10 household categories, including two photometrically challenging ones, and **3**) provides high-quality pose annotations with an error range of only 1.35 mm to 1.74 mm. The dataset also includes **4**) 41 large-scale scenes with comprehensive viewpoint and occlusion coverage, **5**) a checkerboard-free environment, and **6**) dense 6D parallel-jaw robotic grasp annotations. Additionally, we present benchmark results for leading category-level pose estimation networks.*

* Equal contributions.
† Corresponding Author (e-mail: guangyao.zhai@tum.de).

1. Introduction

6D pose estimation is one of the cornerstones in many computer vision tasks, especially for interactions like robotic manipulation [60, 66–68] or augmented reality [20]. Many methods have been proposed to solve this task from various perspectives and achieve outstanding results on public benchmarks [5, 25, 34, 63]. Most of the methods focus on instance-level where each network is trained and tested on a single object instance [42, 58]. However, generalization and applicability are limited, as the object mesh is required, and an individual network needs to be trained for each instance. Recent methods focus on category-level pose estimation [9, 11, 37, 43, 44, 59] by training on multiple objects within one category. They can later gener-

Table 1. **Dataset Overview.** HouseCat6D represents a large-scale and highly accurate category-level 6D pose dataset that combines the advantages of various established datasets (e.g. extensive pose coverage, highly accurate GT, occlusions cases, and grasping annotation).

Dataset	RGB	Depth	Polarisation	Real	Multi-View	mm-accurate GT	Occlusion	Symmetry	Transparent	Reflective	Grasping	Pose Density	Pose Variation	Workspace	Categories	Objects	Scenes
FAT [54]	✓	✓			✓	✓	✓	✓							–	21	> 1k
BlenderProc [13]	✓	✓			✓	✓	✓	✓							–	–	> 1k
LabelFusion [46]	✓	✓		✓			✓								–	12	138
Linemod [3, 25]	✓	✓		✓			✓	✓							–	15	15
IC-BIN [15]	✓	✓		✓			✓	✓							–	2	55
Toyota Light [27]	✓	✓		✓				✓							–	21	21
TUD Light [28]	✓	✓		✓											–	3	9
YCB [5, 63]	✓	✓		✓			✓	✓							–	21	92
T-LESS [26]	✓	✓		✓			✓	✓							–	30	20
HomebrewedDB [34]	✓	✓		✓			✓	✓							–	33	13
ITODD [17]		✓		✓	✓	✓	✓	✓		(✓)					–	28	800
GraspNet-1Billion [21]	✓	✓		✓	✓		✓	✓			✓	+++	+++	+	–	88	190
HOPE [56]	✓	✓		✓	✓		✓	(✓)				++++	++++	+++	–	28	50
StereoOBJ-1M [40]	✓			✓	✓	✓	✓	✓	✓	✓		++++	++++	+++	–	18	183
kPAM [45]	✓	✓		✓			✓	✓							2	91	362
TOD [39]	✓	✓		✓	✓	✓		✓	✓						3	20	10
Wild6D [22]	✓	✓		✓	✓			✓	(✓)			++	++	++++	5	162	486
NOCS [59]	✓	✓			✓							+	+	++	6	42	18
PhoCaL [61]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		+++	+++	+	8	60	24
HouseCat6D (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	+++++	+++++	+++++	**10**	**194**	41

alize to unseen objects from the same category. However, a significant limitation blocking further progress is the lack of datasets for training and evaluation that fulfill all criteria like large-scale, accurate, and realistic. Existing category-level datasets only comply partly, e.g., high quantity and low quality [59], or high quality but insufficient quantity [61].

To this end, we propose a new category-level dataset HouseCat6D. It consists of high-quality ground-truth annotations on diverse objects acquired by multiple sensor modalities with extensive viewpoint coverage. Our dataset includes 194 objects from 10 different categories, including photometrically challenging classes such as glass and cutlery (Fig. 1), occlusion cases, and 3 sensor modalities, i.e., RGB, depth, and polarimetric images, with a total of 23.5k frames and approx. 160k annotated object poses. We additionally provide 10M grasp pose annotations to a subset of the dataset, endowing it with the capacity to serve robotic manipulation tasks, e.g., category-level robotic grasping [62]. Our dataset recording relies on an accurate external infrared tracking system and additional subsequent post-processing through sparse bundle adjustment to avoid errors induced by timestamp offsets and motion blur of the freely moving camera rig [50, 51]. Specifically, we conduct three calibrations, *i.e.* pivot calibration, timestamp calibration, and hand-eye-calibration. For the timestamp calibration, we adopt existing methods [18, 30] adjusted to our setup with an ICP-based refinement. For the hand-eye-calibration, we improve the calibration from recent work [61] by aggregating multiple measurements of a ChArUcO [2] calibration board (Sec. 3.3). Compared to the recent PhoCaL dataset [61] that relies on a robotic end-effector to estimate poses and thus has limited viewpoint coverage and backgrounds, our method provides accurate object pose annotation and wide viewpoint coverage while providing pose annotations of similar quality. We use active stereo as depth maps, which is more reliable on different surface materials [32, 61]. In addition to the typical RGB and depth, we provide polarimetric images with four different filter angles. Recent investigations have shown that this modality is especially suitable for tasks such as depth and surface normal estimation [33, 35, 57], and 6D pose estimation [23], especially for photometrically challenging objects or surfaces. In summary, our main contributions are:

1. We propose HouseCat6D a **large-scale multi-modal category-level object pose dataset** with RGBD + RGBP data, comprising 194 high-quality 3D models of household objects including transparent and reflective objects in 41 scenes with broad viewpoint coverage, challenging occlusions and no markers.
2. We develop a novel **pipeline for annotation, recording, and post-processing** to achieve comparable accuracy to

robotic GT, but with a mobile handheld multi-camera rig. We detail all acquisition and calibration steps and make the **high-quality 6D object pose annotations together with 6D grasp labels** accessible to the community.

3. We provide and discuss the **benchmark evaluation results** on HouseCat6D for SOTA category-level baselines to show challenges and foster novel research in the field.

2. Related Work

The recent state-of-the-art methods are mostly data-driven approaches. A common need for these methods is a dataset for training and evaluation. In this section, we give an overview of existing datasets and provide a summary of mentioned datasets in Tab. 1.

2.1. Instance-level 6D Object Pose Dataset

Early-stage datasets provide nontemporal consistent images. LineMOD [25] and LM-Occlusion [3] are arguably the most used datasets. They use an RGBD camera to annotate the pose of the objects. The camera pose is estimated with checkerboards, which constantly appear in all images. Although these two datasets were heavily used, the quality of object meshes and annotations varies [4]. Other datasets were proposed to overcome these issues. Such as HomebrewedDB [34] and others [16, 52]. However, those datasets still rely on checkerboard-based camera localization, or human-powered annotation [49], or a rotating table [17] to provide tolerable annotations. Consecutive datasets focus on providing sequential images with camera and object pose annotations. This allowed to investigate pose tracking approaches with temporal constraints [4, 24, 36]. One very popular such benchmark is YCB [63]. The annotation is achieved by leveraging an RGBD camera and Structure from Motion (SfM) [46]. Although this makes large-scale annotation possible, the annotation quality is bound to the quality of the depth camera used [32, 61]. In comparison, the Laval 6DOF dataset [24] marker-based tracking results in high-quality annotations and checkerboard-free images. However, marker-induced depth artifacts need a depth map post-correction. On the other hand, StereoOBJ-1M [40] uses SfM with checkerboards in a more precise way to ensure quality and quantity. However, this also introduces checkerboards in every image. GraspNet-1Billion [21] provides parallel-jaw grasping labels besides object pose annotations, making it more feasible for downstream robotic bin-picking. However, the dataset has limited viewpoint changes and only simple backgrounds. In contrast, our dataset captures multiple household scenarios with adequate viewpoint coverage.

2.2. Category-level 6D Object Pose Dataset

Category-level pose estimation has been proposed to address generalizability in 6D pose estimation over multiple objects of the same category. The task is to generalize pose estimation per class and not for individual instances, which is challenging due to high intra-class variance. Many recent methods have been proposed [7, 8, 14, 31, 37, 43, 44, 53] to solve this problem due to its realistic setup. Only a few datasets exist, which we will briefly review here.

The NOCS dataset [59] is the first category-level 6D pose dataset. It contains six categories and two sub-datasets, namely CAMERA25 and REAL275. In REAL275, the poses are aligned using checkerboards. For CAMERA25, ShapeNetCore [6] objects are placed in table-top scenarios. A dataset focusing more on the robotic field is kPam, which uses keypoints. Manuelli *et al*. [45] capture kPam using a similar approach as Marion et al. [46]. They perform 3D reconstruction before manually labeling the keypoints on the 3D reconstruction. The dataset results in 117 training sequences and 245 testing sequences. While NOCS and kPAM contain solid objects, TOD [39] and PhoCaL [61] specifically focus on either translucent or transparent and reflective objects. TOD [39] is captured with a robotic arm and annotated keypoints and focuses on stereo images. Wang *et al*. [61] introduce a category-level dataset including polarimetric images besides RGBD only. For annotation, a robotic arm is used to tip individual objects with a calibrated pointer. Annotations are refined via ICP. Instead of using a robotic arm, Wild6D [22] is annotated via tracking. Every 50th keyframe is annotated and then registered via TEASER++ [65] and colored ICP [48]. The training dataset is label-free, and only the test dataset contains annotations. For recording, multiple iPhones are used to capture RGB images, depth, and point cloud. Fu *et al*. [22] introduce Wild6D, an unlabeled RGBD video dataset with diverse scenes. They also investigate the use of additional synthetic labels and annotate a fraction of the real videos for evaluation. Other class-based datasets exist. Objectron [1] focuses on scale and provides over 14k scenes. However, it only provides annotated 3D bounding boxes and does not give detailed shape information for the objects.

3. Dataset

Our dataset aims to provide large scale with extensive view coverage and high-quality pose annotations without a checkerboard. It is composed of 34 training scenes (20k frames), five test scenes (3k frames), and two validation scenes (1.4k frames). The scenes comprise objects from 10 household categories, including photometrically challenging objects like glass and cutlery, with occlusions. With a total of 194 objects, each category contains 19 objects on average. Our dataset also features multiple modalities,

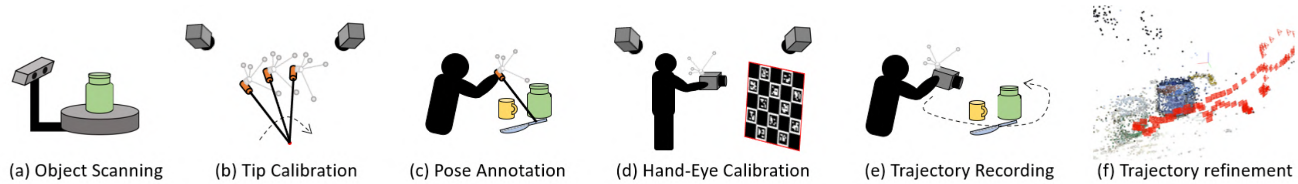


Figure 2. **Dataset Acquisition Pipeline.** (a): Pre-scanning 3D models. (b): Pivot calibration to calibrate measurement tip from the tracking body. (c): Pose annotation of objects using measurement tip. (d): Hand-Eye-Calibration to calibrate camera center of tracking body. (e): Camera trajectory recording (f): Post-processing step to reduce synchronization-induced trajectory error.

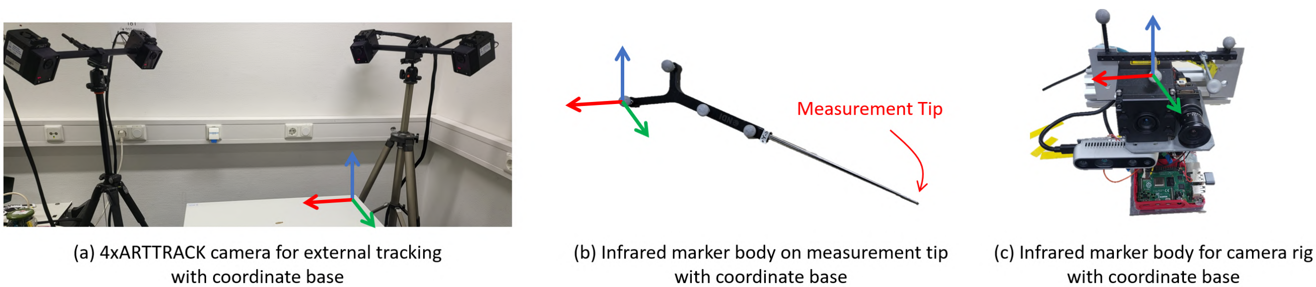


Figure 3. **Tracking System.** ARTTRACK2 tracking system and sets of infrared marker bodies we used for our setup. Once at least four infrared spheres are detected from at least two cameras, the tracking system provides the pose of the marker body as transformation from tracker system base to marker body base.

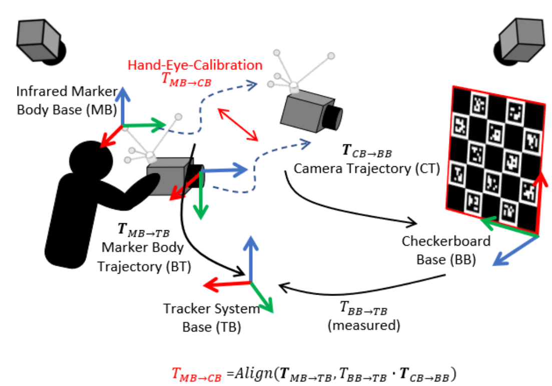


Figure 4. **Hand-Eye-Calibration.** Instead of single image based using closed form solution with measured checkerboard pose like in [61], our newly proposed approach takes more image captures into account. This makes it more robust against checkerboard detection errors in a wider range of camera poses.

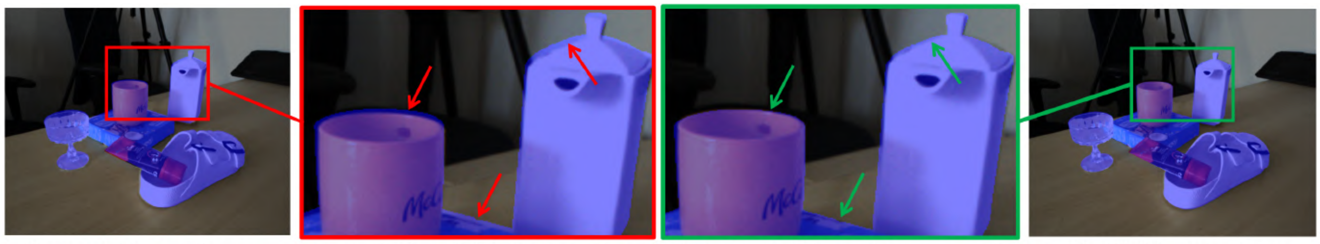
namely RGB images, polarimetric images, and depth maps. This section details our dataset. The acquisition setup is described in Fig. 2.

3.1. Objects & Hardware

Here, we briefly describe the hardware setup we use for the dataset acquisition. More detailed information, such as product names and their specs, is provided in the supplementary material. For our dataset, we choose 10 household categories to represent typical household scenarios: bottle, box, can, cup, cutlery, glass, remote, shoe, teapot, and tube. All objects are scanned with a structured light stereo-based 3D scanner to ensure the quality of the reconstructed meshes. For the photometrically challenging categories, we use self-vanishing 3D scanning spray to enable scanning. For tracking the annotation tool and camera rig, we utilize an external tracker system composed of 4 infrared cameras to ensure tracking quality without using a checkerboard. Fig. 3 shows our camera setup and used tracking bodies for the annotation pipeline. We evaluate the accuracy of the tracking with translation and rotational error [24] using a robotic setup (details in supplementary material). The average error is 0.67 mm / 0.12° in the static case and 0.92 mm / 0.16° in the dynamic tracking scenario. Our dataset comprises two main modalities: Polarimetric RGB image and active stereo depth. We use a dedicated sensor for each of the modalities. For polarimetric images, we use a polarimetric camera, which produces four polarized RGB images for every shot. To measure depth, we decided on an active stereo depth sensor over Time-of-Flight sensors as active stereo depth provides, in general, more robust depth on photometrically challenging material [32]. To synchronize the two cameras, an external hardware trigger is used to trigger both cameras simultaneously.

3.2. Object Pose Annotation

Annotating the 6D pose of the object is, without a doubt, the most crucial part of a 6D pose dataset. In our dataset, we adopt the highly accurate object pose annotation pipeline



(a) Rendered object mask on camera trajectory without COLMAP refinement step (b) Rendered object mask on camera trajectory with COLMAP refinement step

Figure 5. **Post-Processing via Bundle Adjustment.** Example of COLMAP [50, 51] refinement on selected frame with large camera displacement. Even though our timestamp synchronization step reduces the effect of motion induced pose offset, subtle errors still remain ((a), marked red). In comparison, the post processing step significantly reduces the given offset error ((b), marked green).

from [61] but replace the robotic end-effector pose with an IR tracking body. This ensures reliable tracking quality while covering a more extensive working volume. The annotation step follows tool tip calibration, 3D points measurement of the objects, and point correspondence with ICP-based refinement. In this subsection, we describe the details of each step.

Tip Calibration. The poses of the object meshes are annotated by measuring the 3D point using the tooltip. Thus, calibrating the location of the tip from the tracking body is essential to ensure the accuracy of the annotation. We use an NDI Active 4-Marker Planar Rigid Body (Northern Digital, Ontario Canada) as the measurement tip (Fig. 3 (b)). The tip is calibrated by fixing the tip while pivoting the tracking body and finding the optimal location of the point to minimize the variance of the fixed point (pivot calibration). The most common way to evaluate the quality of the pivot calibration is by measuring the variance of the fixed pivot point. We carefully calibrated with 18 points, with the final variance of the tip location of $\varepsilon = 0.040$ mm.

Pose Annotation. After the tip is calibrated from the tracking body, it can measure accurate 3D points in space in the world coordinates of the tracker system. We measure points for the initial point correspondence and ICP refinement as in [61] while covering around three times more point measurements with various surfaces of the object, thanks to the enlarged working space without the constraint given by using a robot arm [61]. We evaluate the quality of the pose annotation step by simulating the pose annotation pipeline on randomly selected three objects with the addition of pivot calibration error (Sec. 3.2) and static tracking error (Sec. 3.1), which gives an average RMSE of 0.32 mm in translation and $0.43°$ in rotation.

3.3. Camera Trajectory Annotation

Another critical aspect of 6D pose annotation is accurate camera trajectories. The object poses are annotated from the center of the tracker system, not from the the individual camera. Thus, the camera pose from the tracker system base has to be applied to obtain the 6D pose of the object from the camera center. In this section, we describe the detailed steps of camera trajectory annotation precisely.

Hand-Eye-Calibration In our scenario, Hand-Eye-Calibration obtains the transformation between the tracker marker body and the center of the camera image sensor. The most common way to perform Hand-Eye-Calibration [55] is detecting the checkerboard multiple times via camera while tracking the camera body from an external source and optimizing both checkerboard base from the tracker system base $T_{BB \to TB}$ and camera base from the marker body base $T_{MB \to CB}$ (hand-eye-calibration). In comparison, [61] proposes a way to use the measurement tip to measure $T_{BB \to TB}$ and form a close form solution with a single checker board detection to obtain $T_{MB \to CB}$. However, we found that the accuracy of the closed-form solution is often unreliable. To solve this, we propose a new hand-eye calibration, which takes into account multiple image captures (Fig. 4). We capture multiple static images from different locations to form two trajectories - one from the camera and one from the tracking body and extract a fixed offset matrix by applying Horn's alignment method [29], which is the hand-eye-calibration matrix. Once the calibration $T_{MB \to CB}$ is obtained, we align the two trajectories and compare the errors to the calibration accuracy. The RMSE for this calibration is measured as 0.27 mm for the translation and $0.42°$ for the rotation.

Camera to Tracker Time Synchronization Another important aspect of using the external IR tracking system is the timestamp calibration between the tracking system and the camera image acquisition time. It can cause severe offset on poses depending on the movement of the camera. A common practice to synchronize the timestamp difference is to measure the trajectory of the camera from two modalities via image and tracking system, along with their timestamp, and maximize the similarity between the trajectories. This brings the best timestamp offset [18, 30]. In our case, we

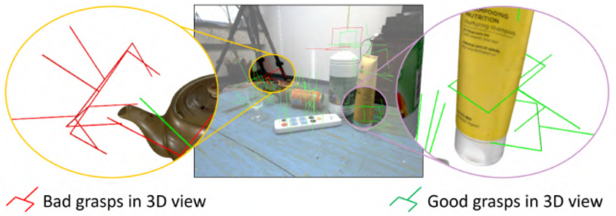


Figure 6. **Grasp Annotations.** After inspection of grasps, we annotate successful grasps (here coloured in green) and failed grasps (here red). We downsample the amount of annotated grasps for better visualization.

Table 2. **Accuracy Comparison Against Existing Datasets.** While RGBD-based datasets are limited by sensor standard deviation [40], multi-view setups [39, 40] offer improvements. Our dataset annotation quality, though not as high as robotic acquisitions [61], surpasses checkerboard-based datasets [40], and excels in terms of viewpoint coverage and annotation accuracy.

Dataset	RGBD based	TOD [39]	StereOBJ [40]	PhoCal [61]	Ours
3D Labeling	Depth Map	Multi-View	Multi-View	Robot	IR tracker
Point RMSE [mm]	≥ 17	3.4	2.3	0.80	$1.35 \leq \epsilon \leq 1.73$

use ICP-based trajectory alignment to find the best timestamp offset instead of using a similarity measure. We empirically find it is more robust to noise and able to synchronize two trajectories with arbitrary frequency without any interpolation to match the frequency. For the camera timestamp, we use the hardware trigger timestamp. We evaluate the synchronization by simulating signals with measured noise. One with the tracking system error (Sec. 3.1) and one with a detection-based error (Sec. 3.3). The simulated error is measured as 0.03 sec.

Pose Refinement Although time synchronization improves the quality of the camera pose, the motion-induced pose error cannot be obliterated as the time synchronization is imperfect due to noise in the checkerboard detection during calibration [30] as well as the difference in individual camera image acquisition time due to its hardware condition. This effect can be observed when camera motion involves large displacement between consecutive frames (Fig. 5, (a)). To tackle this, we use the RGB input to minimize the reprojection error with multi-view images. We use structure from motion [50, 51] with given initial poses and carefully selected fixed frames. The initial poses are used for initial feature matching and structure reconstruction. The fixed frames are excluded in the later bundle adjustment stage. These frames are manually picked upon careful inspection of the frame with the largest IoU between rendered object masks on the RGB image given the pose annotation. We show the improvement in Fig. 5, (b).

3.4. Grasp Annotation

To facilitate downstream robotic manipulation tasks, *e.g.*, robotic pick and place, we endow HouseCat6D with feasible 6D grasping poses for every object under each frame for a subset of collected sequences, following the well-established pipeline introduced in [19, 21]. Taking the annotation process for one object as an example: Firstly, we use antipodal sampling with an inspection of Isaac Gym [41] to distinguish the successful grasps $\mathbf{G}^{s}_{obj}$ from failed ones $\mathbf{G}^{uns}_{obj}$ to generate grasp candidates $\mathbf{G}_{obj} := \{\mathbf{G}^{s}_{obj}, \mathbf{G}^{uns}_{obj}\}$ for the object mesh. Then we use the obtained object pose under the tracker base $T_{obj \rightarrow TB}$ to transform $\mathbf{G}_{obj}$ to the tracker frame where we also reconstruct and amend the background. Finally, we perform collision checking and prune grasping labels in the whole environment and reproject remaining grasps $\mathbf{G}_{TB}$ to each camera frame in the whole sequence according to the transformation from the tracker base to each camera base $T_{TB \rightarrow CB}$. An annotated example is shown in Fig. 6. More details about the annotation pipeline and parameters can be found in the supplementary material.

3.5. Pose Annotation Quality Evaluation

For the evaluation of the annotation quality of object poses, we report the point-wise RMSE between objects and the camera center with and without the consideration of three systematic errors: tracking system error (Sec. 3.1), pose annotation error (Sec. 3.2) and hand-eye-calibration error (Sec. 3.3). As the accuracy gain from the structure-from-motion cannot be directly quantified, we report the RMSE with upper- and lower bounds. In the upper bound, we report the number with the object annotation error and the static tracking error, assuming no synchronization error. In the lower bound, we include all three mentioned systematic errors, including dynamic tracking error as the tracking system error. We report our annotation quality compared to recent datasets in Tab. 2. Our method achieves a low RMSE of 1.35 mm to 1.73 mm.

3.6. Scene Statistics

HouseCat6D features 41 large-scale scenes with 194 objects in 10 categories with grasping labels for 16 scenes. It comprises 34 training scenes with 124 objects, 5 test scenes with 50 objects, and 2 validation scenes with 20 objects for object pose estimation tasks. For the 34 training scenes, a total of 20k frames are recorded. Each training scene contains, on average, 6 objects of different categories. The 5 test scenes and 2 validation scenes consist of 3k and 1.4k frames. They are composed of 10 unseen objects per scene with different categories. Compared to other category-level datasets, HouseCat6D covers the most diverse number of instances and categories. For robotic grasping, we provide 14 training, 1 validation, and 1 test scene. Nonetheless, all 16 scenes can serve to train a real grasping pipeline, as the test would be performed in a real-world setup where success rate serves as the main grasping metric.

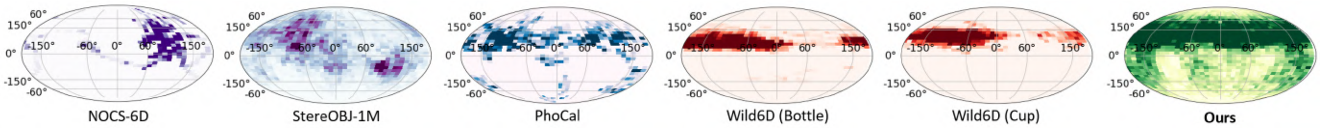


Figure 7. **Pose Distribution.** The pose distribution for category-level datasets NOCS [59] (Test), StereOBJ-1M [40] (Val), PhoCal [61] (Train), two categories of Wild6D [22] and Ours is plotted as the Mollweide projection of the spherical histogram, to exemplify the density and pose variation. Ours shows larges diversity of poses around objects, also for the lower hemisphere, and denser overall distribution.

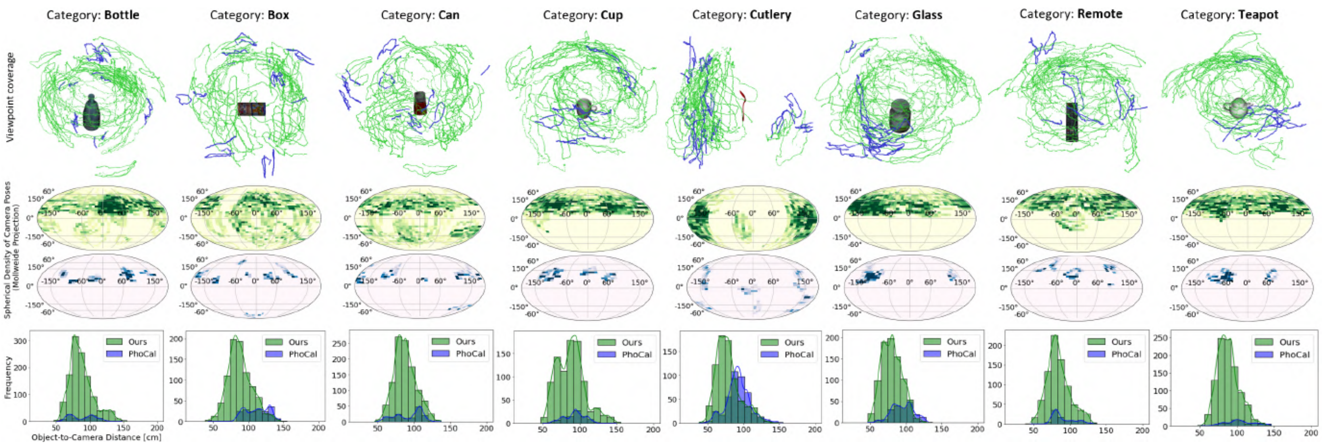


Figure 8. **Pose Distribution per Category.** We compare the pose distribution of HouseCat6D (green) against PhoCal [61] dataset (blue), for the categories included in both of them. The trajectory visualization (top) verifies the much larger and better distributed pose coverage of our HouseCat6D dataset. Compared is further the rotational pose coverage as a spherical histogram plotted as Mollweide projection (center) and the object-to-camera distance as a histogram with relative frequency against the distance in cm (bottom).

Table 3. **Quantitative Benchmark Comparisons.** Class-wise evaluation of 3D IoU (at 25%, 50%) for NOCS [59], FS-Net [8], VI-Net [38] and GPV-Pose [14] on the test split of HouseCat6D. Best results on the full set are reported in bold.

Approach	$3D_{25}$ / $3D_{50}$	Bottle	Box	Can	Cup	Remote	Teapot	Cutlery	Glass	Tube	Shoe
NOCS [59]	50.0 / 21.2	41.9 / 5.0	43.3 / 6.5	81.9 / 62.4	68.8 / 2.0	81.8 / 59.8	24.3 / 0.1	14.7 / 6.0	95.4 / 49.6	21.0 / 4.6	26.4 / 16.5
FS-Net [8]	74.9 / 48.0	65.3 / 45.0	31.7 / 1.2	98.3 / 73.8	96.4 / 68.1	65.6 / 46.8	69.9 / 59.8	71.0 / 51.6	99.4 / 32.4	79.7 / 46.0	71.4 / 55.4
GPV-Pose [14]	74.9 / 50.7	66.8 / 45.6	31.4 / 1.1	98.6 / 75.2	**96.7** / 69.0	**65.7 / 46.9**	**75.4 / 61.6**	70.9 / 52.0	**99.6 / 62.7**	76.9 / 42.4	67.4 / 50.2
VI-Net [38]	**80.7 / 56.4**	**90.6 / 79.6**	**44.8 / 12.7**	**99.0 / 67.0**	**96.7 / 72.1**	54.9 / 17.1	52.6 / 47.3	**89.2 / 76.4**	99.1 / 93.7	**94.9 / 36.0**	**85.2 / 62.4**

Table 4. **Sensor Depth Issues and Pose Coverage Influence.** Class-wise evaluation of 3D IoU (at 25%, 50%). NOCS* denotes using ground truth NOCS maps and sensor depth for lifting. VI-Net* [38] denotes a reduced training set. VI-Net is the best baseline.

Approach	$3D_{25}$ / $3D_{50}$	Bottle	Box	Can	Cup	Remote	Teapot	Cutlery	Glass	Tube	Shoe
VI-Net [38]	80.7 / 56.4	90.6 / 79.6	44.8 / 12.7	99.0 / 67.0	96.7 / 72.1	54.9 / 17.1	52.6 / 47.3	89.2 / 76.4	**99.1 / 93.7**	**94.9 / 36.0**	85.2 / 62.4
NOCS* [59]	**96.7 / 93.6**	**99.8/ 98.2**	**98.3 / 95.8**	**100.0 / 99.0**	**100.0 / 99.0**	**100.0 / 97.3**	**99.9 / 86.1**	**100.0 / 99.9**	80.0 / 68.0	89.1/29.1	**99.7 / 83.9**
NOCS [59]	50.0 / 21.2	41.9 / 5.0	43.3 / 6.5	81.9 / 62.4	68.8 / 2.0	81.8 / 59.8	24.3 / 0.1	14.7 / 6.0	95.4 / 49.6	21.0 / 4.6	26.4 / 16.5
VI-Net [38]	**80.7 / 56.4**	**90.6 / 79.6**	**44.8 / 12.7**	**99.0 / 67.0**	**96.7 / 72.1**	**54.9 / 17.1**	**52.6 / 47.3**	**89.2 / 76.4**	**99.1 / 93.7**	**94.9 / 36.0**	**85.2 / 62.4**
VI-Net* [38]	68.4 / 31.3	91.1 / 67.7	44.1 / 10.2	97.7 / 62.9	91.8 / 39.0	43.0 / 15.6	22.8 / 5.8	80.7 / 41.4	93.7 / 49.2	82.2 / 6.3	36.5 / 14.5

3.7. Viewpoint Coverage

Established datasets in 6D pose estimation lack well-distributed and dense camera pose coverage around the object. They usually focus on the upper hemisphere, even for large-scale dataset variants like SterOBJ-1M [40]. In contrast, HouseCat6D provides very dense and well-distributed poses (cf. Fig. 7). In terms of category level, we compare our trajectories for mutual classes against the recent PhoCal [61] dataset, which provides very accurate annotations but is limited in the range of motion by the robotic arm used for acquisition (cf. Fig. 8).

4. Benchmark and Experiments

Object Pose Estimation In 6D pose estimation, RGB-D input is often used. RGB data aids in classifying objects amidst high intra-class variability. Initially, RGB images identify objects, followed by depth maps for shape and boundary information. NOCS [59] generates 2D

NOCS maps, integrating depth data and ICP for 3D prediction. GPV-Pose [14] segments objects in RGB, then back-projects the depth map for 3D pose prediction through geometry-guided voting. FS-Net [8] derives 3D point clouds from depth images post RGB detection, extracting features via a residual network for size and translation estimation. VI-Net [38] simplifies this task, decoupling the rotation into viewpoint and in-plane rotations, learned separately. Our experiment employs Di et al.'s implementation of FS-Net.

We report IoU results with 25% and 50% thresholds in Tab. 3 (cf. supp. mat. for additional metrics). The geometry-guided methods GPV-Pose and FS-Net outperform the 2D lifting approach, with VI-Net achieving the best results at both thresholds. GPV-Pose and FS-Net benefit from precise 2D detection training in HouseCat6D, aiding in detailed object localization. In contrast, NOCS offers a single-stage approach, lifting results from 2D to 3D. Compared to the NOCS [59], our dataset features cluttered scenes leading to occlusions and closely situated objects (cf. Fig. 9). While current methods often overlook occlusions, our initial evaluations suggest significant potential for improvement. The supplementary material details occlusion ratios per category and comparisons with the NOCS dataset. Although VI-Net handles clutter and occlusions better than others, there remains considerable scope for enhancement.

Further, an experiment using NOCS with ground truth predictions but sensor depth alignment reveals inaccuracies in 3D lifting, impacting results as reported in Tab. 4 (top). This is reflected in a significant decline to 22.6% at the mean IoU at 75% as also shown in Fig. 9. The categories glass and tube suffer especially from the sensor depth. For these, the trained VI-Net even outperforms the ground truth to sensor depth lifting approach. To demonstrate the importance of extensive pose coverage, we trained VI-Net* on a data subset similar to the pose coverage of PhoCaL (cf. Tab. 4 (bottom). Results indicate that our scene coverage notably enhances prediction accuracy.

Grasp Pose Estimation KGN [10] processes an RGB-D image to estimate gripper keypoints and employs PnP to align 3D keypoints in the gripper frame with 2D camera frame keypoints, solving for 6D grasp poses. We have retrained both the complete KGN model and a simplified version without keypoint offset refinement on HouseCat6D data.

Following the metrics from the original study [10], we report grasp coverage rate (GCR) and object success rate (OSR) in Tab. 5a (cf. [12, 47, 64]). Real-world experiments were conducted using a 7-DoF Franka robot, with grasp success rates detailed in Tab. 5b, in line with the approach in [68] of calculating successful trial percentages over 15 runs per object. Notably, Tab. 5b demonstrates the method's

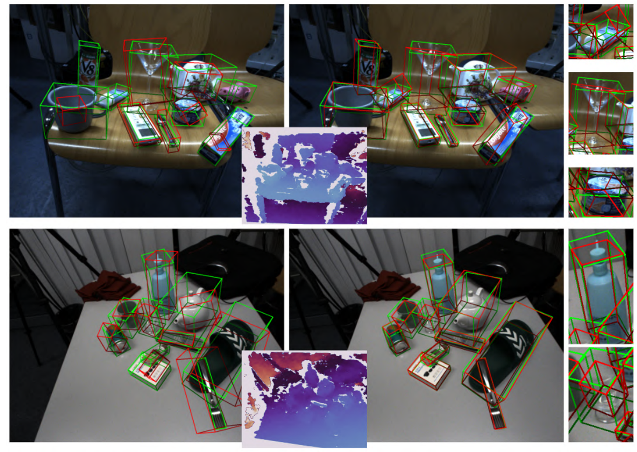


Figure 9. **Sensor Depth Issues.** Comparison of NOCS prediction (left) and using the NOCS ground truth map but sensor depth (small, center). Even with perfect NOCS maps, the lifting from 2D to 3D suffers under the noisy sensor depth map (right).

Table 5. **Grasping Results.** KGN [10] *without / with* keypoint refinement, respectively.

(a) Grasp coverage rate and object success rate on the test set.

Metric	Bottle	Can	Cup	Glass	Remote	Tube
GCR (%)	17.4 / 24.1	35.4 / 58.3	32.6 / 34.1	16.3 / 40.8	64.3 / 64.5	47.3 / 61.1
OSR (%)	97.8 / 99.8	75.2 / 87.6	100 / 100	100 / 100	92.1 / 94.8	100 / 100

(b) Real-world grasp success rate.

Metric	Box	Cup	Glass	Remote	Unknown
GSR (%)	80.0 / 80.0	66.7 / 66.7	66.7 / 73.3	53.3 / 33.3	53.3 / 60.0

generalization to unseen objects across different categories. A surprising finding is KGN's efficacy in grasping transparent objects (Glass) post-training on HouseCat6D, both in tests and real-world settings. This contrasts with the original method's limited performance on photometrically challenging objects due to its synthetic training.

5. Conclusion

HouseCat6D is introduced as a comprehensive 6D pose dataset, acquired using a multi-modal camera rig and an external tracking system, offering highly accurate pose annotations. This dataset addresses the limitations of existing datasets by featuring realistic, marker-free scenes with well-distributed object poses. It includes photometrically challenging objects lacking texture and those made of translucent materials, alongside precise robotic grasping annotations. HouseCat6D, with its quality and breadth, advances research in categorical pose estimation, setting a new standard for applications in everyday household environments and other areas.

References

[1] 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. 3

[2] Gwon Hwan An, Siyeong Lee, Min-Woo Seo, Kugjin Yun, Won-Sik Cheong, and Suk-Ju Kang. Charuco board-based omnidirectional camera calibration method. *Electronics*, 7 (12):421, 2018. 2

[3] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Proceedings of the European Conference on Computer Vision*, pages 536–551. Springer, 2014. 2, 3

[4] Benjamin Busam, Hyun Jun Jung, and Nassir Navab. I like to move it: 6d pose estimation as an action decision process. *arXiv preprint arXiv:2009.12678*, 2020. 3

[5] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv preprint arXiv:1502.03143*, 2015. 1, 2

[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 preprint arXiv:1512.03012*, 2015. 3

[7] 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. 3

[8] 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 (CVPR)*, pages 1581–1590, 2021. 3, 7, 8

[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. 1

[10] Yiye Chen, Yunzhi Lin, Ruinian Xu, and Patricio A Vela. Keypoint-graspnet: Keypoint-based 6-dof grasp generation from the monocular rgb-d input. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7988–7995. IEEE, 2023. 8

[11] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *CVPR*, 2024. 1

[12] Fu-Jen Chu, Ruinian Xu, and Patricio A Vela. Real-world multiobject, multigrasp detection. *IEEE Robotics and Automation Letters*, 3(4):3355–3362, 2018. 8

[13] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Dmitry Olefir, Tomas Hodan, Youssef Zidan, Mohamad Elbadrawy, Markus Knauer, Harinandan Katam, and Ahsan Lodhi. Blenderproc: Reducing the reality gap with photorealistic rendering. In *International Conference on Robotics: Sciene and Systems, RSS 2020*, 2020. 2

[14] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 3, 7, 8

[15] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3583–3592, 2016. 2

[16] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3583–3592, 2016. 3

[17] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd - a dataset for 3d object recognition in industry. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017. 2, 3

[18] Ulrich Eck, Frieder Pankratz, Christian Sandor, Gudrun Klinker, and Hamid Laga. Precise haptic device co-location for visuo-haptic augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 0:15, 2015. 2, 5

[19] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021. 6

[20] Marco Esposito, Benjamin Busam, Christoph Hennersperger, Julia Rackerseder, Nassir Navab, and Benjamin Frisch. Multimodal us–gamma imaging using collaborative robotics for cancer staging biopsies. *International journal of computer assisted radiology and surgery*, 11(9):1561–1571, 2016. 1

[21] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020. 2, 3, 6

[22] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *arXiv:2206.15436*, 2022. 2, 3, 7

[23] Daoyi Gao, Yitong Li, Patrick Ruhkamp, Iuliia Skobleva, Magdalena Wysock, HyunJun Jung, Pengyuan Wang, Arturo Guridi, and Benjamin Busam. Polarimetric pose prediction, 2021. 2

[24] Mathieu Garon, Denis Laurendeau, and Jean-François Lalonde. A framework for evaluating 6-DOF object trackers.

In *Proceedings of the European Conference on Computer Vision*, 2018. 3, 4

[25] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 858–865. IEEE, 2011. 1, 2, 3

[26] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 2

[27] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018. 2

[28] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 577–594. Springer, 2020. 2

[29] Berthold Horn, Hugh Hilden, and Shahriar Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America A*, 5:1127–1135, 1988. 5

[30] Manuel Huber, Michael Schlegel, and Gudrun Klinker. Temporal calibration in multisensor tracking setups. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pages 195–196. IEEE, 2009. 2, 5, 6

[31] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10632–10640, 2022. 3

[32] HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, et al. On the importance of accurate geometry data for dense 3d vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–791, 2023. 2, 3, 4

[33] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8602–8611, 2020. 2

[34] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 1, 2, 3

[35] Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild, 2021. 2

[36] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep iterative matching for 6d pose estimation. *International Journal of Computer Vision*, 128(3):657–678, 2019. 3

[37] 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. 1, 3

[38] Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14001–14011, 2023. 7, 8

[39] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11602–11610, 2020. 2, 3, 6

[40] Xingyu Liu, Shun Iwase, and Kris M Kitani. Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10870–10879, 2021. 2, 3, 6, 7

[41] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021. 6

[42] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6841–6850, 2019. 1

[43] Fabian Manhardt, Manuel Nickel, Sven Meier, Luca Minciullo, and Nassir Navab. Cps: Class-level 6d pose and shape estimation from monocular images. *arXiv preprint arXiv:2003.05848*, 2020. 1, 3

[44] 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. 1, 3

[45] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019. 2, 3

[46] Pat Marion, Peter R Florence, Lucas Manuelli, and Russ Tedrake. Label fusion: A pipeline for generating ground truth labels for real rgbd data of cluttered scenes. In *IEEE International Conference on Robotics and Automation*, pages 3235–3242. IEEE, 2018. 2, 3

[47] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2901–2910, 2019. 8

[48] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 143–152, 2017. 3

[49] Colin Rennie, Rahul Shome, Kostas E Bekris, and Alberto F De Souza. A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place. *IEEE Robotics and Automation Letters*, 1(2):1179–1185, 2016. 3

[50] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 5, 6

[51] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 5, 6

[52] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3D object detection and pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 462–477. Springer, 2014. 3

[53] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision*, pages 530–546. Springer, 2020. 3

[54] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2038–2041, 2018. 2

[55] Roger Y Tsai, Reimar K Lenz, et al. A new technique for fully autonomous and efficient 3 d robotics hand/eye calibration. *IEEE Transactions on robotics and automation*, 5(3): 345–358, 1989. 5

[56] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13081–13088. IEEE, 2022. 2

[57] Yannick Verdié, Jifei Song, Barnabé Mas, Benjamin Busam, Ales Leonardis, and Steven McDonagh. Cromo: Cross-modal learning for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3937–3947, 2022. 2

[58] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 1

[59] 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 Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 1, 2, 3, 7, 8

[60] Pengyuan Wang, Fabian Manhardt, Luca Minciullo, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Demograsp: Few-shot learning for robotic grasping with human demonstration. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5733–5740. IEEE, 2021. 1

[61] 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. In *CVPR*, 2022. 2, 3, 4, 5, 6, 7

[62] Hongtao Wen, Jianhang Yan, Wanli Peng, and Yi Sun. Transgrasp: Grasp pose estimation of a category of objects by transferring grasps from only one labeled instance. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 445–461. Springer, 2022. 2

[63] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter . Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems*, 2018. 1, 2, 3

[64] Ruinian Xu, Fu-Jen Chu, and Patricio A Vela. Gknet: grasp keypoint network for grasp candidates detection. *The International Journal of Robotics Research*, 41(4):361–389, 2022. 8

[65] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2020. 3

[66] Guangyao Zhai, Yu Zheng, Ziwei Xu, Xin Kong, Yong Liu, Benjamin Busam, Yi Ren, Nassir Navab, and Zhengyou Zhang. Da^2 dataset: Toward dexterity-aware dual-arm grasping. *IEEE Robotics and Automation Letters*, 7(4):8941–8948, 2022. 1

[67] Guangyao Zhai, Xiaoni Cai, Dianye Huang, Yan Di, Fabian Manhardt, Federico Tombari, Nassir Navab, and Benjamin Busam. Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs. *arXiv preprint arXiv:2309.12188*, 2023.

[68] Guangyao Zhai, Dianye Huang, Shun-Cheng Wu, HyunJun Jung, Yan Di, Fabian Manhardt, Federico Tombari, Nassir Navab, and Benjamin Busam. Monograspnet: 6-dof grasping with a single rgb image. In *IEEE International Conference on Robotics and Automation*. IEEE, 2023. 1, 8