PhoCaL: A Multi-Modal Dataset for Category-Level Object Pose Estimation with Photometrically Challenging Objects

Pengyuan Wang[*1], HyunJun Jung[*1], Yitong Li[1], Siyuan Shen[1], Rahul Parthasarathy Srikanth[1],
Lorenzo Garattoni[2], Sven Meier[2], Nassir Navab[1], Benjamin Busam[1]
[*] Equal Contribution   [1] Technical University of Munich   [2] Toyota Motor Europe

pengyuan.wang@tum.de   hyunjun.jung@tum.de   b.busam@tum.de

Figure 1. PhoCaL comprises 60 high quality 3D models of household object in 8 categories with different photometric complexity. The selected objects include challenging texture-less, occluded, symmetric, reflective and transparent objects. Our robotic-induced pose annotation pipeline provides highly accurate 6D pose labels even for objects that are hard to capture by modern RGBD sensors. The figure shows RGB, 3D bounding boxes and rendered Normalized Object Coordinate Space (NOCS) map for 4 example scenes.

Abstract

Object pose estimation is crucial for robotic applications and augmented reality. Beyond instance level 6D object pose estimation methods, estimating category-level pose and shape has become a promising trend. As such, a new research field needs to be supported by well-designed datasets. To provide a benchmark with high-quality ground truth annotations to the community, we introduce a multimodal dataset for category-level object pose estimation with photometrically challenging objects termed PhoCaL. PhoCaL comprises 60 high quality 3D models of household objects over 8 categories including highly reflective, transparent and symmetric objects. We developed a novel robot-supported multi-modal (RGB, depth, polarisation) data acquisition and annotation process. It ensures sub-millimeter accuracy of the pose for opaque textured, shiny and transparent objects, no motion blur and perfect camera synchronisation.

To set a benchmark for our dataset, state-of-the-art RGB-D and monocular RGB methods are evaluated on the challenging scenes of PhoCaL.

1. Introduction

Vision systems interacting with their environment need to estimate the position and orientation of objects in space, which highlights why 6D object pose estimation is an important task for robotic applications. Even though there have been great advances in the field [6, 42], instance-level 6D pose methods require pre-scanned object models and support limited number of objects. Category-level object pose estimation [40] scales better to the needs of real operating environments. However, photometrically challenging objects such as shiny, e.g. metallic, and transparent, e.g. glass, objects are very common in our daily life and little work has been done to estimate their 6D poses within practical accuracy on a category-level. The difficulty arises from two aspects: first, it is difficult to annotate 6D pose ground truth for photometrically challenging objects since no texture can be used to determine key points; second, commonly used depth sensors fail to return the correct depth information, as structured light and stereo method often fail to correctly interpret reflection and refraction artefacts. As a consequence, RGB-D methods [25, 40] do not work reliably with photometrically challenging objects. We intro-

Figure 2. Our dataset comprises 60 household objects among 8 object categories. The training and test split is depicted here.

duce PhoCaL, a class-level dataset of photometrically challenging objects with high-quality ground-truth annotations. The dataset provides multi-modal data such as RGB, depth and polarization which enables investigation into object's surface reflectance properties.

We obtain highly accurate ground truth poses with a novel method using a collaborative robot arm in gravity compensated mode and a calibrated mechanical tip. In order to annotate the 6D pose of transparent and non-textured objects, a specially designed tip is mounted on the robot arm. With the calibrated tip, the positions of pre-defined points on the object surface are acquired on the real object and matched to a scan thereof. Using this method, the object pose can be determined with an order of magnitude more accuracy than previous methods. For transparent and textureless objects, topographic key points are used instead of textural ones. The points gathered in this way are then matched to the object model in a final ICP [2] step to yield an accurate fit.

The camera to robot end-effector transformation is needed to obtain the object poses in camera coordinates. Typically, hand-eye calibration approaches solve this by visually estimating the marker position and optimizing for the transformation between camera and end-effector. To minimize the error propagation and obtain highly accurate ground truth labels, we instead used the end-effector tip of the arm in gravity-compensated mode to measure the position of 12 points on a ChArUco [1] board. This allows us to use the robot's accurate position system to obtain both object poses and camera poses for image sequences.

Beyond photometrically challenging categories and high-quality annotations, multi-modal input is another highlight of PhoCaL. As the active depth sensors fail on metallic and transparent surfaces, we include an additional passive sensor modality in the form of a polarization camera. It provides valuable information on object surfaces [22]. In our

setup, we designed and 3D printed a rig that holds multiple cameras, each mounted on it and carefully calibrated. During recording, a pre-defined trajectory is repeated by the robot arm. The robot arm stops when capturing images from all cameras, which avoids motion blur and diminished effects from imperfect synchronization.

In summary, our main contributions are:

1. We propose **PhoCaL**, a **multi-modal** (RGBD + RGBP) **dataset for category-level object pose estimation**. The dataset comprises 60 high-quality 3D models of household objects including symmetric, transparent and reflective objects in 8 categories with 24 sequences featuring occlusion, partial visibility and clutter.

2. We introduce a new and **highly accurate pose annotation method using a robotic manipulator** that allows for sub-millimeter precision 6D pose annotations of photometrically challenging objects even with reflective or transparent surfaces.

2. Related Work & Current Challenges

Standardized datasets are used in the field of object pose and shape estimation to quantify and compare contributions and advances in the field. These datasets generally fall in two domains: instance-level datasets, where the 3D model of the object is known a priori; and category-level datasets, where the exact CAD model is unknown. Tab. 1 provides an overview of related datasets in both domains.

2.1. Instance-level 6D Object Pose Dataset

One of the earliest, most widely used publicly available datasets for instance level pose estimation is LineMOD [19] and its occlusion extension LM-Occlusion [5]. Their data

Dataset	RGB	Depth	Polarisation	Real	Multi-View	Robotic GT	Occlusion	Symmetry	Transparent	Reflective	Categories	Objects	Sequences	License
FAT [38]	✓	✓			✓		✓	✓			–	21	> 1k	CC BY-NC-SA 4.0
BlenderProc [12]	✓	✓			✓		✓	✓			–	–	> 1k	GNU GPL 3.0
LabelFusion [31]	✓	✓		✓			✓				–	12	138	BSD 3-Clause
Toyota Light [21]	✓	✓		✓				✓			–	21	21	MIT
YCB [8, 41]	✓	✓		✓			✓	✓			–	21	92	MIT
Linemod [5, 19]	✓	✓		✓			✓	✓			–	15	15	CC BY 4.0
GraspNet-1Billion [15]	✓	✓		✓			✓	✓			–	88	190	CC BY-NC-SA 4.0
T-LESS [20]	✓	✓		✓			✓	✓			–	30	20	CC BY 4.0
HomebrewedDB [23]	✓	✓		✓			✓	✓			–	33	13	CC0 1.0 Universal
ITODD [14]		✓		✓	✓		✓	✓		(✓)	–	28	800	CC BY-NC-SA 4.0
StereoOBJ-1M [26]	✓			✓	✓		✓	✓	✓	✓	–	18	183	Not (yet) released
kPAM [30]	✓	✓		✓			✓	✓			2	91	362	MIT
CAMERA25 [40]	✓	✓		(✓)			✓	✓			6	42	30	MIT
REAL275 [40]	✓	✓		✓				✓			6	42	13	MIT
TOD [27]	✓	✓		✓	✓			✓	✓		3	20	10	CC BY 4.0
Ours (PhoCaL)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	8	60	24	CC BY 4.0

Table 1. Overview of pose estimation datasets. The upper part shows instance-level datasets while the lower part includes category-level setups. PhoCaL is the only dataset that includes both photometrically challenging objects with high quality (robotic) pose annotations and all three modalities, RGB, depth, and polarisation.

was acquired using a PrimeSense RGB-D Carmine sensor and a marker board was used to keep track of the relative sensor pose. While undoubtedly pioneering this field, the 3D model quality is now outdated and the leader boards on these datasets have become saturated. HomebrewedDB [23] accounts for the latter shortcoming by providing high quality 3D models scanned with a structured light sensor. Including three models from LineMOD, they add 30 more toy, household and industrial objects. Different illumination conditions and occlusions make the scenes more challenging. Other datasets also include household objects [13, 21, 34, 37] or focus on industrial parts [14, 20] with low texture for which it is also possible to manually design or retrieve accurate CAD models [20]. The BOP 6D pose benchmark [21] includes a summary of these datasets with standardized metrics in a common format.

While the datasets mentioned so far provide individual frames, the YCB-Video dataset [41] also includes video sequences of 21 household objects. While YCB uses LabelFusion [31] for semi-manual frame annotation and pose propagation through the sequence, Garon et al. [16] leverage tiny markers on the object to estimate the poses in their videos directly at the cost of synthetic data cleaning afterwards. The advent of photo realistic rendering further enables a branch of works that leverages training on purely synthetic data [12, 38]. Although this circumvents the cumbersome pose labelling process, it introduces a domain gap between synthetic data for evaluations and real-world appearances faced in the final applications.

2.2. Category-level Object Poses and Datasets

In real-world applications, a 3D model is not always available, but pose information is still required. Detection of such objects under these conditions has classically been tackled using 3D geometric primitives [3, 4, 9].

While these methods consider outdoor scenes for which kitti [18] provides 3D bounding box annotations, they lack object shape comparison and the information is often too inaccurate for robotic grasping tasks. The pioneering work of NOCS [40] was the first category-level method that could detect object pose and shape in indoor environments. Further investigations consider correspondence-free methods [10] where a deep generative model learns a canonical shape space from RGBD and a method to estimate pose and shape for fully unseen objects is also proposed [32], albeit this method requires a reference image for latent code generation. CPS [28] demonstrates how to estimate pose and metric shapes at category level, using only a monocular view. The extension CPS++ [29] further utilizes synthetic data and a domain transfer approach using self-supervised refinement with a differentiable renderer from RGBD data without annotations. SGPA [11] explores shape priors to estimate the object pose. DualPoseNet [25] leverages spherical fusion for better encoding of the object information.
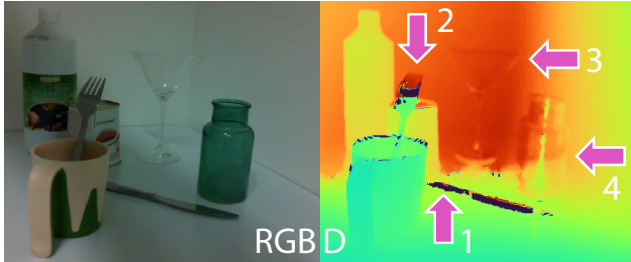
Figure 3. Limitations of RGBD sensors. The depth for photometrically challenging objects is difficult to measure with a commodity depth sensor. The intel RealSense D515 LiDAR ToF sensor used here is affected by reflections that lead to invalid (1) incorrect (2) distance estimates. Moreover, the glassware becomes invisible to the sensor (3) and causes noise (4).
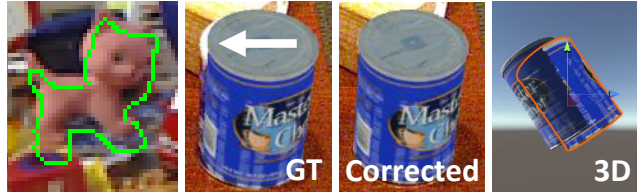

Figure 4. Annotation quality for poses in datasets Linemod [19] (projected green silhouette, left) and YCB [8] (rendered overlay, right) together with its correction [7] (right).

We leverage the standard RGBD method NOCS and the strong state-of-the-art RGB method CPS to set the baselines on our new dataset. While task-specific datasets for general object detection exist for robot grasping [15, 30], methods for category-level pose estimation are typically tested on NOCS [40] data. The NOCS objects comprise various categories, but do not contain photometric challenges often present in everyday objects such as reflectance and transparency.

2.3. Photometric Challenges and Multimodalities

While texture-less objects [20] were initially challenging for pose estimation, transparency presents an even bigger hurdle. While the problem is not new, previous methods have addressed this using RGB stereo without a 3D model to identify grasping points only [36]. Rotational object symmetry can be leveraged by contour fitting for transparent object reconstruction [33] using template matching. ClearGrasp [35] proposes a method for geometry estimation of transparent objects based on RGBD. However, this method passes over the transparent regions from the depth map and predicts depth from RGB in these areas instead. Liu et al. [27] investigate instance- and category-level pose estimation from stereo imagery. Since their depth sensing fails on transparent objects, they use an opaque object twin as proxy to establish ground truth depth. More recently StereOBJ-1M proposed [26] a large dataset including transparent and translucent objects with specular reflections and symmetry. However, at the time of this writing it is not yet available for download.

For 2D object detection, information from multiple orthogonal sensor modalities such as polarisation (RGBP) can help for transparent object segmentation [22]. This modality can provide information in regions were depth sensors fail. Their inherent connection with surface normals [43] can also make them attractive for pose estimation of photometrically challenging objects.

2.4. Ground Truth Pose Annotation

Manual annotation of 6D pose is difficult and extremely time-consuming. Therefore, most datasets rely on semi-manual processes for ground truth annotation. The data from a depth sensor, if available, is often used to register the 3D model and manual adjustments are applied to visually refine the pose for this one frame. Relative camera motion is typically calculated using visual markers [19, 23] to propagate the pose information through a sequence of images. The use of depth sensors for ICP-based alignment of pose labels reduces labour and improves fully-manual annotation quality. However, depth maps from RGBD sensors are erroneous or invalid for photometrically challenging objects with high reflectance and translucent or transparent surfaces [26]. An examples is shown in Fig. 3.

Ensuring high quality of pose labels over a series of images is difficult and errors accumulate as the examples in Fig. 4 show. This equally affects depth-based refinement strategies of 6D pose pipelines [21, 24]. We propose a mechanical measurement process using a robotic manipulator to circumvent this issue and allow for high precision labels that omits the error propagation of relative camera pose retrieval from images.

3. Dataset Acquisition Pipeline

Our dataset features multiple object classes including photometrically challenging classes such as objects with reflective surfaces or transparent material. It also provides multi-modal sensor data with highly accurate 6D pose annotation. This section describes our dataset acquisition pipeline as shown in Fig. 5.

3.1. Objects Model Acquisition

To represent a cross section of common household objects, we selected eight common categories for our category-level 6D object pose dataset: bottle, box, can, cup, remote, teapot, cutlery, glassware. All object models are scanned using an EinScan-SP 3D Scanner (SHINING 3D Tech. Co., Ltd., Hangzhou, China). The scanner is a structured light stereo system with a single shot accuracy of $\leq$ 0.05 mm in a scanning volume of $1200 \times 1200 \times 1200$ mm^3.
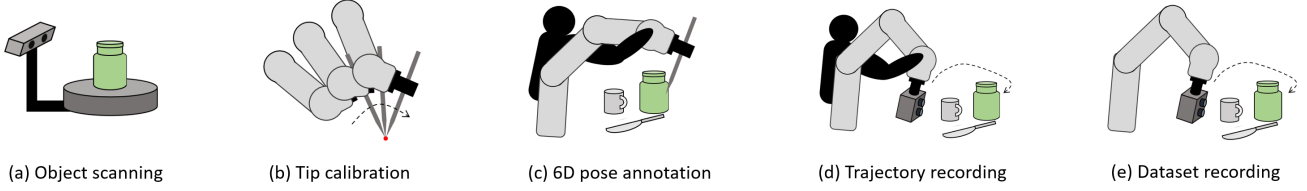
Figure 5. Overview of dataset acquisition pipeline. (a): 3D models are extracted with a structured light scanner. (b): Pivot calibration calibrates a tipping tool to robot coordinates. (c): 6D poses are annotated using the tool and manual movements of the robot. (d): The camera trajectory is saved. (e): Dataset is recorded automatically following the planned trajectory.
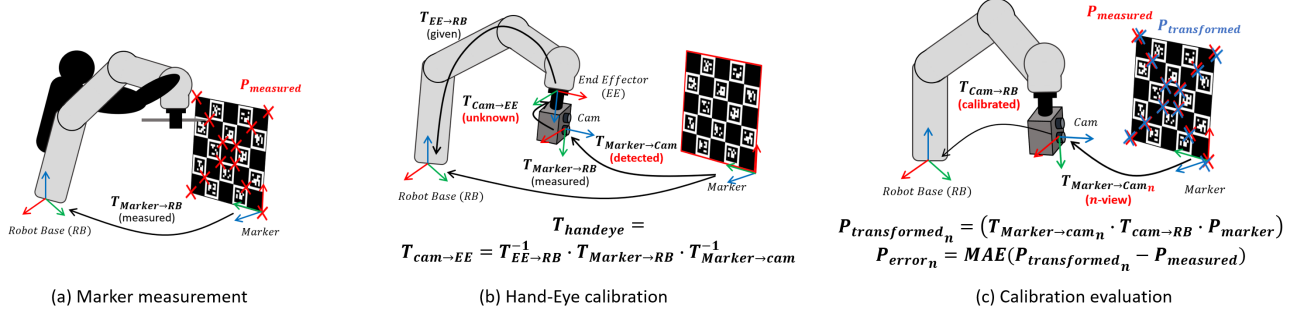

Figure 6. Overview of hand-eye-calibration and its evaluation. (a): shows the marker-to-robot calibration. (b): illustrates camera-to-robot hand-eye calibration. (c) depicts our accuracy evaluation.

The models from the first six categories are provided as textured obj files. Since the cutlery and glassware objects are photometrically challenging with their highly reflective and transparent surfaces, we apply a self-vanishing 3D scanning spray (AESUB Blue, Aesub, Recklinghausen, Germany) to make the objects temporarily opaque for scanning. We scan the object and provide an obj file without texture. The spray sublimes after approx. 4 h.

3.2. Scene Acquistion Setup

For each scene, 5-8 objects are placed on the table with the random background. We use a KUKA LBR iiwa 7 R800 (KUKA Roboter GmbH, Augsburg, Germany) 7 DoF robotic arm that guarantees a positional reproducibility of ± 0.1 mm. The vision system comprises a Phoenix 5.0 MP Polarization camera (IMX264MZR/MYR) with Sony IMX264MYR CMOS (Color) Polarsens (i.e. PHX050S1-QC) (LUCID Vision Labs, Inc., Richmond B.C., Canada) with a Universe Compact lens with C-Mount 5MP 2/3" 6mm f/2.0 (Universe, New York, USA). As depth camera, the Time-of-Flight (ToF) sensor Intel®RealSense™LiDAR L515 is used, which acquires depth images at a resolution of 1024x768 pixels in an operating range between 25 cm and 9 m with a field-of-view of 70°x 55°and an accuracy of 5 ± 2.5 mm at 1 m distance up to 14 ± 15.5 mm at 9 m distance.

3.3. Tip Calibration

We use a rigid, pointy metallic tip to obtain the coordinate position of selected points on the object. Tip calibration is therefore essential to ensure the accuracy of the system. The rig attached to the robot's end-effector consists of custom 3D printed mount which holds the tool-tip rigidly. The pivot calibration is performed as shown in Fig. 8 (left), where the tip point is placed in a fixed position, while only the robot end-effector position is changed. We collect data from N such tip positions with corresponding end-effector poses, $_iT_e^b$, which contain rotation $_iR_e^b$ and translation $_it_e^b$, the final translation t_t^e of the end-effector is calculated as follows:

$$t_t^e = \begin{bmatrix} _1R_e^b - _2R_e^b \\ _2R_e^b - _3R_e^b \\ \vdots \\ _nR_e^b - _1R_e^b \end{bmatrix}^\dagger \cdot \begin{bmatrix} _1t_e^b - _2t_e^b \\ _2t_e^b - _3t_e^b \\ \vdots \\ _nt_e^b - _1t_e^b \end{bmatrix} \quad (1)$$

where † denotes the pseudo-inverse. We evaluate the tip calibration by calculating the variance of each tip location at the pivot point. The variance of the tip location in our setup is $\varepsilon = 0.057$ mm.

3.4. 6D Pose Annotation

Annotating the precise 6D pose of the objects is a challenging task as mentioned in Sec. 2.4. Here, we utilize
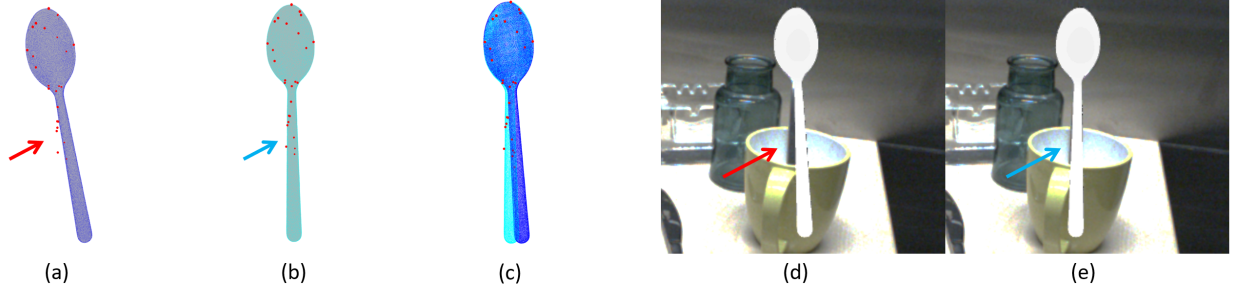
Figure 7. Example of annotation quality before and after ICP based refinement on the textureless object. (a) Initial pose of mesh overlaid with measured surface points (red dots) shows error in initial pose (red arrow). (b) After the ICP, refined pose matches with the surface points properly (blue arrow). (c) Shows improvement in 6D pose annotation. Rendering of the mesh with initial pose (d) and refined pose (e) shows a significant difference in quality.
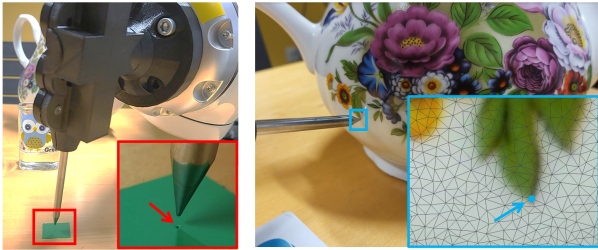

Figure 8. Tip calibration (left) with its pivot point (red). Tip measuring points of object surface (right) and its correspondence on the object's model mesh (blue).

the robot accuracy and its reproducible encoders to annotate the object pose. Our annotation steps are as follows: first, we attach the tool tip on the robot's end-effector and measure several keypoints along with 20-30 surface points of the given object by hand guiding the end-effector while the robot is in gravity compensation mode (Fig 5 (c), Fig 8 (right)). Then, corresponding keypoints are manually picked on the object model's mesh to obtain the initial pose of the respective objects (Fig 8 (right) (blue)). Finally, ICP is applied to align the dense mesh points of the object and the measured sparse surface points as the refinement step for the initial object poses.

To evaluate the refinement performance, 25 points on a specific area of the object surface are picked and uniformly distributed noise of ± 0.2mm is added to simulate the measurement noise. We then apply a small perturbation of random translation errors of range ± 2mm in x,y,z and a rotation error about a random axis with an angle of up to 4 degrees to the object pose to simulate the error introduced by the point correspondences. Thereafter, we apply ICP between the picked surface points and the perturbed mesh to refine the pose. We test this pipeline with 3 selected objects with 5 different random perturbation before applying ICP to recover the initial pose. The pose error is measured in translation and rotational distance [17] after the refinement

and it gives an average RMSE of 0.20 mm in translation and $0.38°$ in rotation.

It is observed that ICP improves the annotation in real life scenario particularly on textureless objects, where it is difficult to find exact correspondence from the mesh. An extreme example of annotated poses before and after ICP on the textureless objects is shown in Fig. 7.

3.5. Hand-Eye Calibration

Traditional hand-eye calibration, such as [39] requires detection of the marker from the camera in various positions to obtain an accurate calibration result. The transformation from camera to end-effector is difficult to estimate as the marker transformation to robot base is unknown and both have to be jointly estimated. In our case, however, the marker position can be accurately measured with the robot tip. Considering this fact, we measure 12 selected points on the marker board and calculate $T_{Marker \rightarrow RB}$ (Fig. 6 (a)) to link the end-effector pose to the camera frame. From $T_{Marker \rightarrow RB}$, the $T_{handeye}$ is calculated as shown in Fig. 6 (b).

The overall accuracy of the entire procedure is measured as shown in Fig. 6 (c). $T_{Marker \rightarrow cam}$ is formed by applying $T_{handeye}$ and multiply $T_{marker \rightarrow cam}$ of n different views to transform the 12 points from the marker board to the robot base ($P_{transformed_n}$). RMSE is calculated by comparing the result to $P_{measured}$. We evaluate our hand-eye calibrations in one of our scenes on both RGBD and Polarization camera with the mentioned approach with $n = 10$ and obtained $\text{RMSE}_{RGBD} = 0.89$ mm and $\text{RMSE}_{Polarization} = 0.83$ mm across all the view points. This calibration is performed procedure for all cameras before recording each scene as shown in Fig. 9.

3.6. Synchronized Robot Pose Capture with Images

RGBD and polarization cameras are used for the data acquisition. A specially designed and 3D printed rig is used to mount both cameras tightly on the end-effector. The tra-
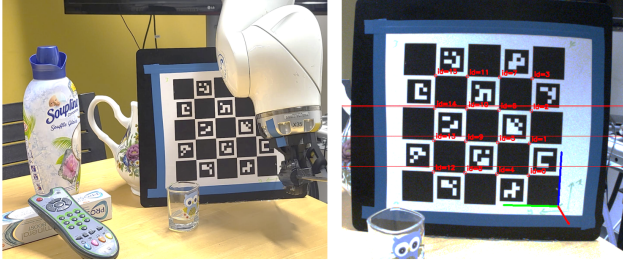
Figure 9. Measuring the marker points for the calibration on the scene (left) and detected marker from one of the cameras (right)

jectory of all joints of the robot is recorded by manually moving the end-effector while the robot arm is in gravity compensated mode. Thereafter, we record the images of the scene by replaying the joint trajectory while stopping the robot every 5-7 joint positions to capture the images and the robot pose (approx 10-15 fps). This ensures no motion blur and camera synchronization artefacts are recorded while reproducing the original hand-held camera trajectory.

3.7. Evaluation of Overall Annotation Quality

We evaluate overall annotation quality of our dataset by running simulated data acquisition with two measured error statistics : ICP error (Sec 3.4) and hand-eye calibration error (Sec 3.5). For both RGBD and Polarization camera, setup from one of the scenes is used including the objects and the trajectories. The acquisition is simulated twice, with and without the aforementioned error. In the end, RMSE error is calculated pointwise in mm between the acquisitions. We averaged the error per object and per each frame in the trajectories.

RMSE error for RGBD camera is 0.84 mm and for polarization camera is 0.76 mm. Detailed description of this procedure is attached in the supplementary material. The annotation quality in comparison with other dataset acquisition principles is shown in Tab. 2.

Dataset	RGBD Dataset	TOD [27]	StereOBJ [26]	Ours
3D Labeling	Depth Map	Multi-View		Robot
Point RMSE	$\geq$ 17mm	3.4mm	2.3mm	0.80mm

Table 2. Comparison of pose annotation quality for different dataset setups. The error for RGBD is exemplified with the standard deviation of the Microsoft Azure Kinect [26].

4. Benchmarks and Experiments

Both monocular (CPS) and RGB-D based (NOCS) category-level methods are considered for the baseline evaluation on the PhoCaL dataset. For the evaluation of NOCS, the normal object coordinate space maps are rendered for each training image and will be published together with the dataset. With the predicted normalized object shape from NOCS map, the depth information is used to lift 2D detection to 3D space using ICP. Considering the artifacts in the depth data from metallic and transparent objects in the dataset, along with the occlusion, the test sequences are very challenging for RGBD methods.

Similiar to NOCS, CPS first detects 2D bounding boxes. Then lifting modules for each class transform 2D image features to 6D pose and scales. Simultaneously the method also estimates the point cloud shape for the respective object class. CPS is trained on approximately 1000 object instance models for each category to learn a deep point cloud encoding of each class. The 2D detection and lifting modules are trained together for 100k steps with a learning rate of 1e-4, decaying to 1e-5 at 60k steps.

4.1. Evaluation Pipeline

Our dataset consists of 24 image sequences in total with training and testing split in each sequence. In our evaluation pipeline, the training split of the first 12 sequences are used to train the network. To have an evaluation on both the known and novel objects in each category, two experiments are designed. To evaluate on seen objects firstly, the network is trained on the training split of the first 12 sequences and tested on the testing split of the same sequences. To further evaluate the generalization ability of NOCS and CPS to novel objects in the same category, the same training split of the first 12 sequences is used, but we evaluate the result on the testing split of the latter 12 sequences, where objects are mostly unseen. With this way, generalization ability of the methods to novel objects in the category is emphasized, which is a common issue in real operating environments. The evaluation metric is the intersection over union (IoU) result with a threshold of 25% and 50%.

4.2. Evaluation Result

The 3D IoU at 25% and 50% evaluations of NOCS for the first experiment setup is shown in Tab. 3. The mean average precision (mAP) for 3D IoU at 25% is 43.34%. It is observed in the experiment that even if the segmentation and normalized object coordinate map predictions are accurate, the lifting from NOCS map to 6D space is sensitive to artifacts in depth maps. Since the objects are highly occluded in the PhoCaL dataset, and depth measurements are inaccurate because of cutlery and glassware categories, the method does not have a good performance on the dataset which indicates the drawbacks of RGBD methods in these photometrically challenging cases. The average precision of each category with respect to 3D IoU threshold is plotted in Fig. 10a. Note that the results of cutlery and glassware categories are among the worst three categories.

For comparison, the result of CPS is also listed in Tab. 3. As can be seen from the table, CPS has a higher preci-

3D$_{25}$ / 3D$_{50}$	Bottle	Box	Can	Cup	Remote	Teapot	Cutlery	Glassware	Mean
NOCS [40]	**91.17** / 0.65	16.10 / 0.01	**85.44 / 23.01**	51.83 / 1.48	**93.26 / 86.05**	0.00 / 0.00	4.89 / 0.01	4.00 / 0.06	43.34 / 13.91
CPS [28]	80.08 / **40.30**	**31.68 / 28.18**	68.96 / 6.69	**81.60 / 70.24**	86.30 / 37.08	**67.43 / 4.31**	**44.00 / 24.95**	**30.33 / 17.74**	**61.30 / 28.69**

Table 3. Class-wise evaluation of 3D IoU for NOCS [40] and CPS [28] on test split of known objects.

3D$_{25}$ / 3D$_{50}$	Bottle	Box	Can	Cup	Remote	Teapot	Cutlery	Glassware	Mean
Experiment 1	**91.17** / 0.65	16.10 / **0.01**	**85.44 / 23.01**	51.83 / **1.48**	**93.26 / 86.05**	0.00 / 0.00	4.89 / **0.01**	**4.00 / 0.06**	**43.3 / 13.91**
Experiment 2	13.70 / **1.28**	**27.74** / 0.00	48.17 / 0.00	**61.77** / 0.00	8.35 / 0.00	**4.90** / 0.00	**16.10** / 0.00	0.83 / 0.00	22.70 / 0.17

Table 4. Class-wise evaluation of 3D IoU for NOCS [40] on seen (Experiment 1) and mostly unseen (Experiment 2) objects.

(a) NOCS result in the first experiment   (b) CPS result in the first experiment   (c) NOCS result in the second experiment
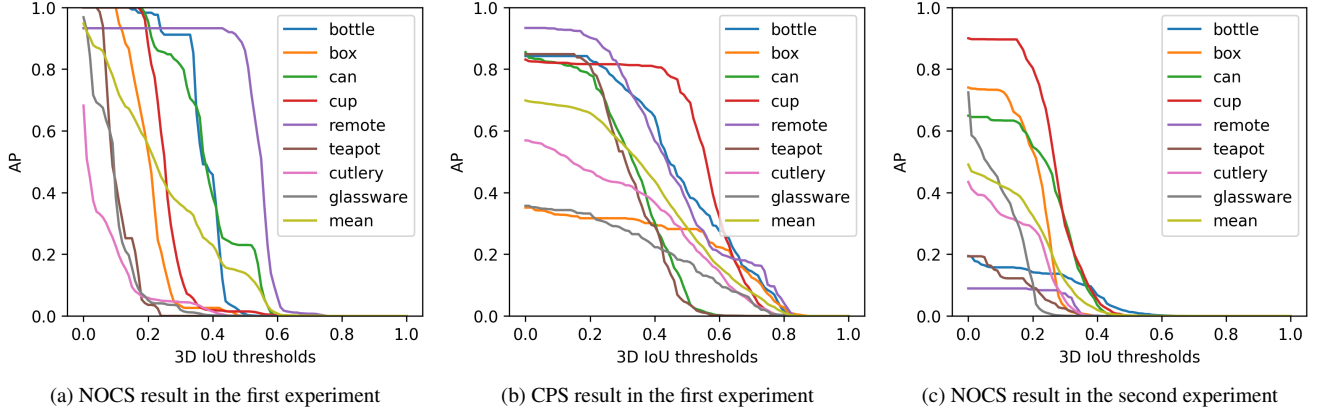
Figure 10. Plots of average precision (AP) with respect to 3D IoU thresholds for each category.

sion for cutlery and glassware categories. Monocular methods are not affected by artifacts in depth images, which explains the result from the experiment. CPS has a higher mAP of 61.30%, which means RGB has an advantage in dealing with photometrically challenging objects. The detailed APs for each category are plotted in Fig. 10b.

In addition, the NOCS evaluation on both experiments are compared in table 4. The evaluation result for the second experiment has a lower mAP for 3D IoU at 25% and 50% as expected, as most of the test objects are novel in the second experiment. Fig. 10c plots NOCS APs in the second experiment. In comparison to NOCS, the CPS result drops significantly in the second experiment and the 3D IoU at 25% is 4.3%. The result shows that pretraining with a large amount of synthetics images is necessary for monocular methods, to learn the correct lifting from 2D detection to 3D space without the help of depth images.

4.3. Limitations

Even though the proposed pipeline for annotating the 6D pose ground truth is accurate, annotating the objects with deformable surface, such as empty boxes, poses a challenge during the surface measurement step in the workflow due to its light deformation which could deteriorate the quality of both initial pose and ICP based refinement. Moreover, the limited workspace of the robot constrains the view an-

gles in the image sequences which is an issue the PhoCaL shares with other robotic acquisition setups. The hand eye calibration of the camera plays a key role for the annotation quality. If the camera resolution is low, a good calibration result requires significantly more input images from different angles.

5. Conclusion

In this paper we introduce the PhoCaL dataset, which contains photometrically challenging categories. High-quality 6D pose annotations are provided for all categories and multiple camera modalities, namely RGBD and RGBP. With our manipulator-driven annotation pipeline, we reach pose accuracy levels that are one order of magnitude more precise than previous vision-sensor-only pipelines even for photometrically complex objects. Moreover, baselines are provided for future works on category-level 6D pose on our dataset by evaluating both monocular and RGB-D methods. The evaluation shows the difficulty level of the dataset in particular for objects that include reflective and transparent surfaces. PhoCaL therefore constitutes a challenging dataset with accurate ground truth that can pave the way for future pose pipelines that are applicable to more realistic scenarios with everyday objects.

References

[1] 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

[2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 2

[3] Tolga Birdal, Benjamin Busam, Nassir Navab, Slobodan Ilic, and Peter Sturm. A minimalist approach to type-agnostic detection of quadrics in point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3530–3540, 2018. 3

[4] Tolga Birdal, Benjamin Busam, Nassir Navab, Slobodan Ilic, and Peter Sturm. Generic primitive detection in point clouds using novel minimal quadric fits. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1333–1347, 2019. 3

[5] 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

[6] Yannick Bukschat and Marcus Vetter. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach. *arXiv preprint arXiv:2011.04307*, 2020. 1

[7] 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. 4

[8] 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. 3, 4

[9] Peter Carr, Yaser Sheikh, and Iain Matthews. Monocular object detection using 3d geometric primitives. In *Proceedings of the European Conference on Computer Vision*, pages 864–878. Springer, 2012. 3

[10] 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 Conference on Computer Vision and Pattern Recognition*, pages 11973–11982, 2020. 3

[11] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2773–2782, 2021. 3

[12] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019. 3

[13] 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

[14] 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*, Oct 2017. 3

[15] 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 Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2020. 3, 4

[16] 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*, pages 582–597, 2018. 3

[17] 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. 6

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 3

[19] 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. 2, 3, 4

[20] 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. 3, 4

[21] 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*, pages 19–34, 2018. 3, 4

[22] 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, 4

[23] 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. 3, 4

[24] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1529, 2017. 4

[25] 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 International Conference on Computer Vision*, pages 3560–3569, 2021. 1, 3

[26] 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. 3, 4, 7

[27] 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. 3, 4, 7

[28] 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. 3, 8

[29] 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. 3

[30] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. *arXiv preprint arXiv:1903.06684*, 2019. 3, 4

[31] 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. 3

[32] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10710–10719, 2020. 3

[33] Cody J Phillips, Matthieu Lecce, and Kostas Daniilidis. Seeing glassware: from edge detection to pose estimation and shape recovery. In *Robotics: Science and Systems*, volume 3, 2016. 4

[34] 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

[35] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *IEEE International Conference on Robotics and Automation*, pages 3634–3642. IEEE, 2020. 4

[36] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008. 4

[37] 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

[38] 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. 3

[39] 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. 6

[40] 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, 3, 4, 8

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

[42] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1941–1950, 2019. 1

[43] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 3d human shape reconstruction from a polarization image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 351–368. Springer, 2020. 4

PhoCaL - Supplementary Material

Pengyuan Wang[*1], HyunJun Jung[*1], Yitong Li[1], Siyuan Shen[1], Rahul Parthasarathy Srikanth[1],
Lorenzo Garattoni[2], Sven Meier[2], Nassir Navab[1], Benjamin Busam[1]
* Equal Contribution   [1] Technical University of Munich   [2] Toyota Motor Europe
pengyuan.wang@tum.de   hyunjun.jung@tum.de   b.busam@tum.de

Figure 1. Example images from all scenes in PhoCaL dataset. The figure shows RGB, coloured masks and rendered models in scenes. Note that the ground truth annotations are accurate even for photometrically challenging objects.

1. Scene Example Visualization

To make the dataset challenging and similar to real environments, different backgrounds are chosen with occlusion between objects. The detailed views of setups and backgrounds are visualized in Fig. 1. Our dataset is composed of 12 scenes with two different trajectories per each scene (i.e. a total of 24 trajectories).
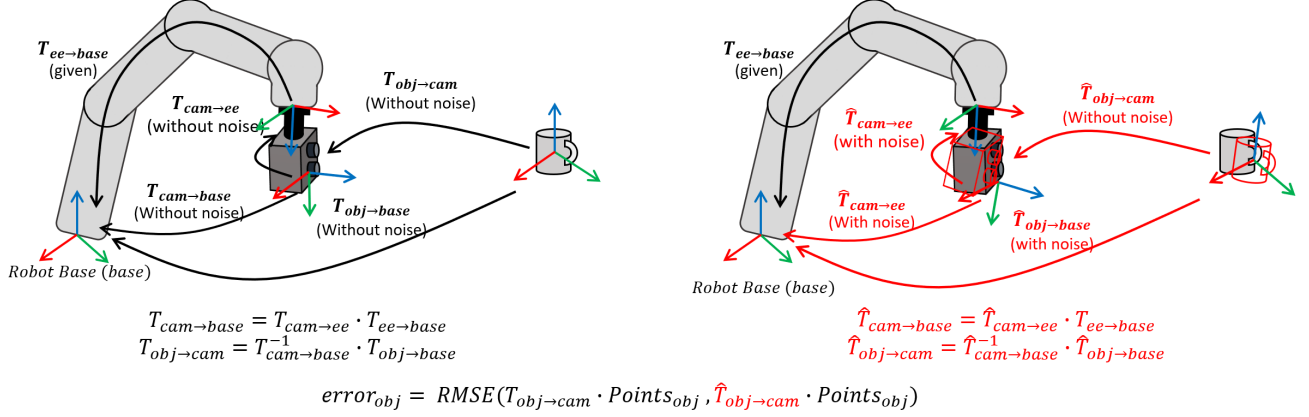
1

$$T_{cam\to base} = T_{cam\to ee} \cdot T_{ee\to base}$$
$$T_{obj\to cam} = T_{cam\to base}^{-1} \cdot T_{obj\to base}$$

$$\hat{T}_{cam\to base} = \hat{T}_{cam\to ee} \cdot T_{ee\to base}$$
$$\hat{T}_{obj\to cam} = \hat{T}_{cam\to base}^{-1} \cdot \hat{T}_{obj\to base}$$

$$error_{obj} = RMSE(T_{obj\to cam} \cdot Points_{obj}, \hat{T}_{obj\to cam} \cdot Points_{obj})$$

Figure 2. Pose graph for our simulated annotation evaluation setup. RMSE of pointwise error is calculated from object mesh points with pose from camera base with noise ($\hat{T}_{obj->cam}$) and without noise ($T_{obj->cam}$).

2. Details of Annotation Quality Evaluation

To evaluate overall annotation quality of our dataset, we run simulated data acquisition with pre-calculated error statistics on object pose annotation step (Sec 3.4 in main paper) and hand-eye-calibration step (Sec 3.5 in main paper), then compare with ground truth similar to [1, 2]. However, as our error statistics per step are obtained in 3D (translation) and 6D (translation + rotation), instead of a projection error in pixel as in [1, 2], the acquisition is simulated by directly applying the error statistics on the steps. In this section, we describe the details of the simulated evaluation pipeline.

Simulated Scene Setup To simulate the dataset acquisition in a realistic way, we chose scene 9 (Fig. 1 6th column, 2nd row) to evaluate our hand-eye calibration accuracy. All the objects are synthetically placed with their annotated pose from the robot base ($T_{obj\to base}$). Then the recorded trajectory of each camera is repeated by applying hand-eye calibration matrix ($T_{cam\to ee}$) on the end-effector pose ($T_{ee\to base}$). Here, the absolute ground truth pose of the objects from each camera center ($T_{obj\to cam}$) is obtained as follows (Fig. 2, left):

$$T_{gt} = T_{obj\to cam} = T_{cam\to ee}^{-1} \cdot T_{ee\to base}^{-1} \cdot T_{obj\to base} \quad (1)$$

The simulated annotated pose from the camera ($\hat{T}_{obj\to cam}$) is obtained by applying noise on both $T_{obj\to base}$ and $T_{cam\to ee}$, where we denote as $\hat{T}_{obj\to base}, \hat{T}_{cam\to ee}$ (Fig. 2, right):

$$T_{annotated} = \hat{T}_{obj\to cam} = \hat{T}_{cam\to ee}^{-1} \cdot T_{ee\to base}^{-1} \cdot \hat{T}_{obj\to base} \quad (2)$$

Simulated Error on Object Pose Annotation For each object in the scene, translation noise of $0.20\,\mathrm{mm}$ and rota-

tion noise of $0.38°$ (Sec 3.4 in the main paper) is added on the $T_{obj\to base}$. To add randomness, we first generate two 3D unit vectors with random orientation per object, where the first vector is multiplied by $0.20\,\mathrm{mm}$ for the translation error t_{error} and the second vector is utilized as axis in axis-angle representation with an angle of $0.38°$, for the rotation error R_{error}.

$$\hat{T}_{obj\to base} = [R_{error}|t_{error}] \cdot T_{obj\to base} \quad (3)$$

Simulated Error on Hand-Eye Calibration To add noise on the hand-eye calibration matrix, a small perturbation is applied on each camera's hand-eye calibration matrix. We apply random perturbation multiple times on the matrix, and choose the perturbation which gives error range of $RMSE_{RGBD} = 0.89\,\mathrm{mm}$ and $RMSE_{Polarization} = 0.83\,\mathrm{mm}$ on $\hat{T}_{cam\to ee}$ (Sec 3.5 in the main paper) as the simulated error on hand-eye calibration.

Simulated Error on Object Pose from Camera Two real trajectories of the end-effector poses $T_{ee\to base}$ are used in the test. For each camera, we run the two sequences and calculate the pointwise error from each object's mesh obtained from T_{gt} and $T_{annotated}$ (equation 1 and 2). The RMSE error is calculated through the frames and averaged for each object. In the test, the final RMSE error is 0.84 mm for RGBD camera, and 0.76 mm for the polarization camera.

3. 3D Object Models in the Dataset

PhoCal comprises high quality 3D models for 60 objects in 8 categories. Textured models are available in 6 categories for bottles, boxes, cans, cups, remotes, and teapots. Photometrically very challenging objects without texture are given in 2 categories, namely cutlery (which are highly

Figure 3. Illustration of high-quality 3D object models from all categories used in the dataset. On the left side are bottles, boxes, cans and cups. On the right side are remotes, teapots, cutlery and glassware.
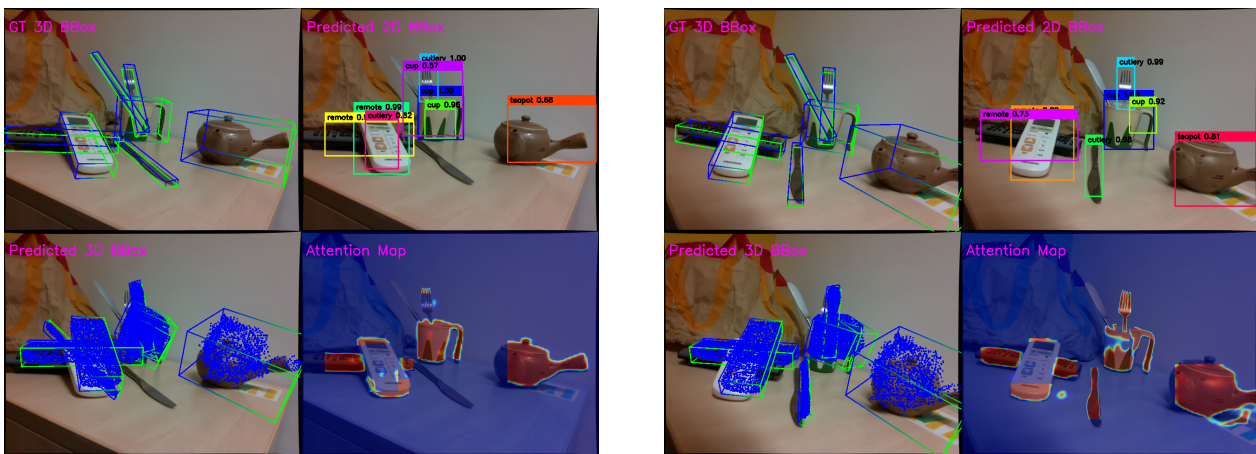

Figure 4. Two example of CPS result for test images in experiment 1, in each example the result contains the ground truth 3D bounding boxes, the predicted 2D bounding boxes, the predicted 3D bounding boxes, the attention maps

reflective) and glassware (which are transparent). The high quality models are visualized in Fig. 3.

4. Visualization of CPS Result

The testing results of CPS in experiment 1 is visualized in Fig. 4, where ground truth 3D bounding boxes, predicted 2D bounding boxes, predicted 3D bounding boxes, and attention maps are plotted. More visualizations are included in the supplementary video.

References

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

[2] 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 Recogni- tion*, pages 11602–11610, 2020. 2