

## PhoCaL - Supplementary Material

Pengyuan Wang<sup>\*1</sup>, HyunJun Jung<sup>\*1</sup>, Yitong Li<sup>1</sup>, Siyuan Shen<sup>1</sup>, Rahul Parthasarathy Srikanth<sup>1</sup>,  
Lorenzo Garattoni<sup>2</sup>, Sven Meier<sup>2</sup>, Nassir Navab<sup>1</sup>, Benjamin Busam<sup>1</sup>

<sup>\*</sup> Equal Contribution    <sup>1</sup> Technical University of Munich    <sup>2</sup> Toyota Motor Europe

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

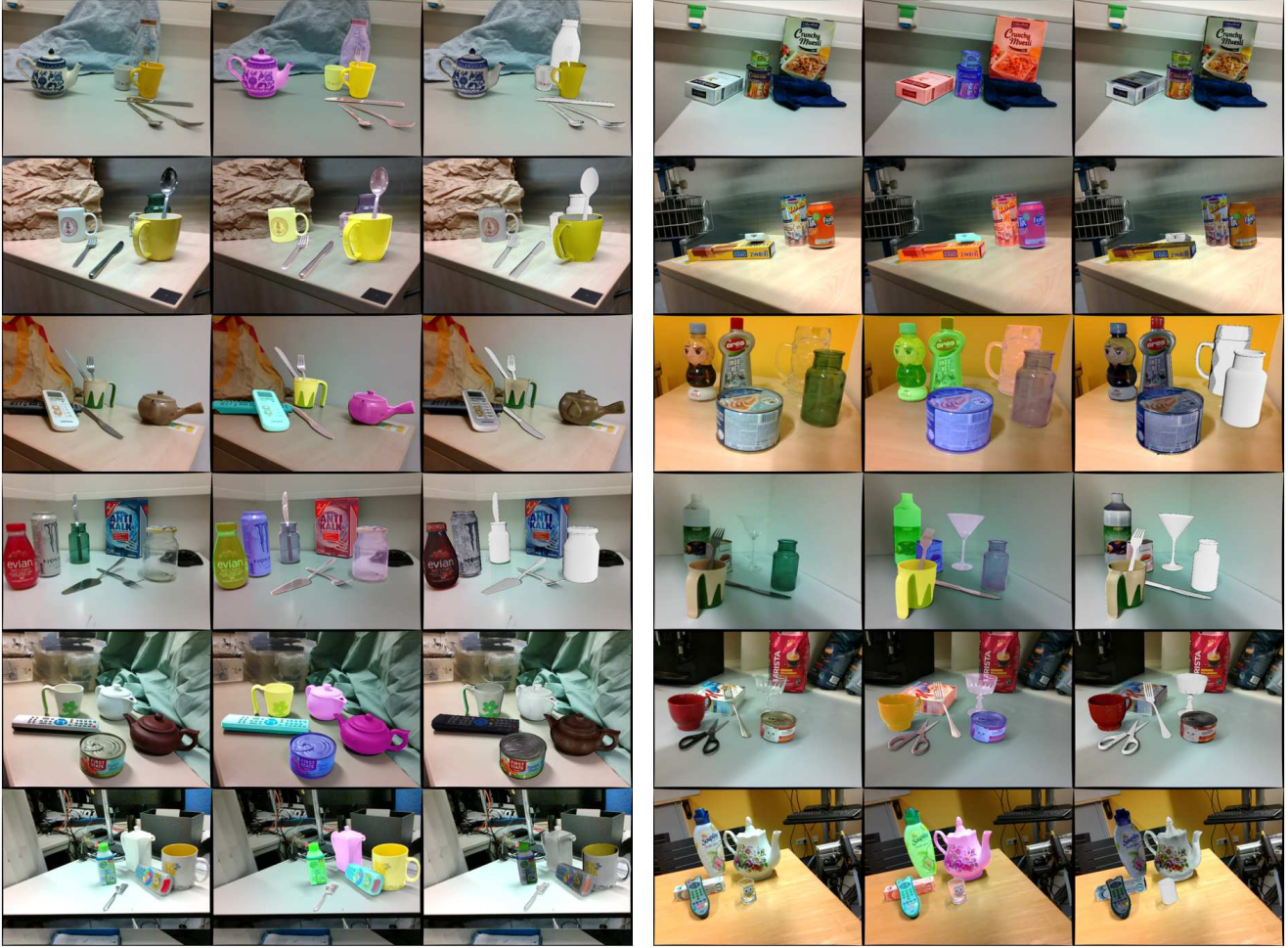


Figure 1. Example images from all scenes in PhoCaL dataset. The figure shows RGB, coloured masks and rendered models in scenes. Note that the ground truth annotations are accurate even for photometrically challenging objects.

### 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).

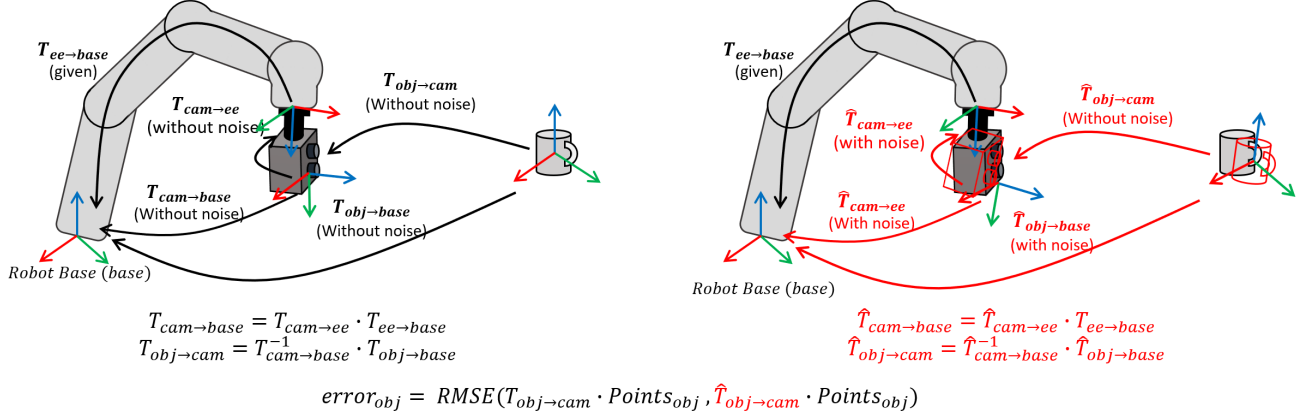


Figure 2. Pose graph for our simulated annotation evaluation setup. RMSE of pointwise error is calculated from object mesh points with pose from camera base with noise ( $\hat{T}_{obj \rightarrow cam}$ ) and without noise ( $T_{obj \rightarrow cam}$ ).

## 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 ( $T_{obj \rightarrow base}$ ). Then the recorded trajectory of each camera is repeated by applying hand-eye calibration matrix ( $T_{cam \rightarrow ee}$ ) on the end-effector pose ( $T_{ee \rightarrow base}$ ). Here, the absolute ground truth pose of the objects from each camera center ( $T_{obj \rightarrow cam}$ ) is obtained as follows (Fig. 2, left):

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

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

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

**Simulated Error on Object Pose Annotation** For each object in the scene, translation noise of 0.20 mm and rota-

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

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

**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  $RMSE_{RGBD} = 0.89$  mm and  $RMSE_{Polarization} = 0.83$  mm on  $\hat{T}_{cam \rightarrow ee}$  (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  $T_{ee \rightarrow base}$  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.

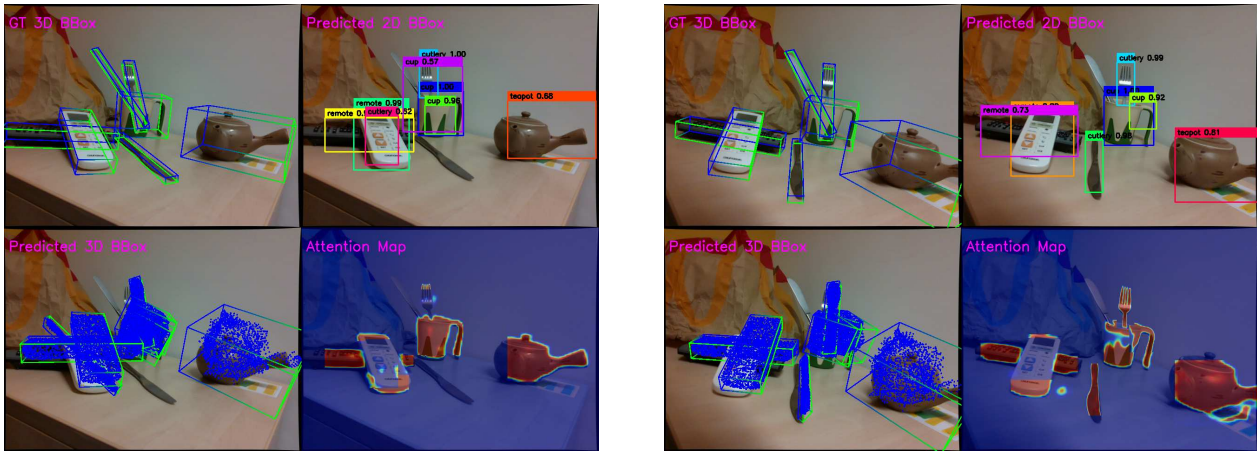


Figure 4. Two example of CPS result for test images in experiment 1, in each example the result contains the ground truth 3D bounding boxes, the predicted 2D bounding boxes, the predicted 3D bounding boxes, the attention maps

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 Recognition*, pages 11602–11610, 2020. 2