GS-Pose: Supplementary

Pengyuan Wang[1], Takuya Ikeda[2], Robert Lee[2], and Koichi Nishiwaki[2]

[1] Technical University of Munich Germany
[2] Woven by Toyota, Japan

1 ShapeNet Objects for Our Training

CPPF [5] leverages synthetic models from ShapeNet [1] for the network training, where a large amount of synthetic object models are collected per category. However, only a small amount of objects are needed for our pipeline and we randomly selected only 10 CAD models from the ShapeNet objects for the training. As is mentioned in NOCS [4], ShapeNet objects contain object meshes that do not look real or have topology problems which results in rendering failures, therefore we manually remove meshes with defects. Training objects of bottle, bowl, camera, can, laptop and mug categories are visualized in Fig. 1, and the trained networks are evaluated on NOCS and Wild6D [6] dataset. Even though part of the objects in certain categories such as cans and mugs are textured with single colors and have large domain gap in comparison with real textures, our network robustly captures the semantic features and can be applied on real data after training. The objects in the chair category for our training are visualized in Fig. 2 and the result is evaluated in the SUN RGB-D [3] dataset. We observe that the chair category has a large variety of shapes and the part of the model textures are single-colored. Despite the challenging task setup including heavy occlusions, our method shows great improvements in comparison with the CPPF baseline in the SUN RGB-D dataset.

2 Detailed Evaluations on NOCS dataset

Inputs Preprocessing The masks from the NOCS dataset contain inaccurate segmentation results and leads to point cloud outliers, therefore radius outlier removal is applied to filter the sparse background points from the partial inputs.

Evaluation Results per Category The evaluation of our method on NOCS dataset for each category is reported in Tab. 1. It is observed that bottle, camera and can categories have relatively low scores for $3D^*_{25}$ and $3D^*_{50}$. The scores for bottle and can are low due to missing bounding box predictions in the dataset. The $3D^*_{50}$ score for the camera is low because the testing cameras have large deformation in the camera lens, which is hard to be approximated by affine transformation of shape priors. For the 5°5cm, 10°5cm, 15°5cm scores, the camera category is low. The mug category has a lower score in 10°5cm because of the difficulty of rotation estimation based on mug handles. Despite this,

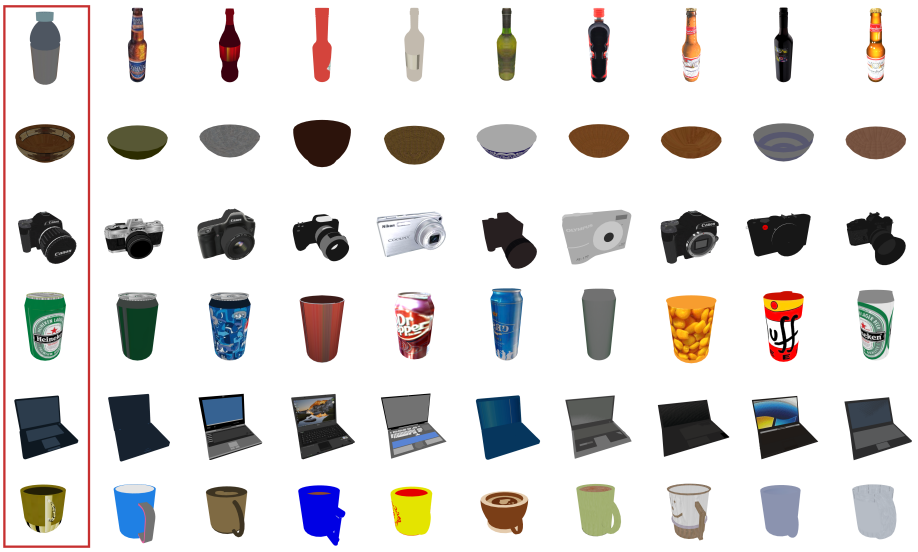


Fig. 1: Visualizations of ShapeNet synthetic models for the training of bottle, bowl, camera, can, laptop and mug categories. Afterwards we evaluate the trained models on NOCS and Wild6D datasets. The objects in the red rectangle are used for inference as shape priors.

Fig. 2: Visualizations of ShapeNet synthetic models for the training of the chair category. Afterwards we evaluate the trained models on SUN RGB-D datasets. The objects in the red rectangle are used for inference as shape priors.

our method outperforms CPPF and NOCS baseline by a large margin for the challenging mug category.

Rotation AP Comparison per Category The rotation average precision is plotted in Fig. 4 including NOCS, CPPF and our method. Our method greatly outperforms the CPPF and NOCS on the challenging mug and laptop category. The reason is that crucial semantic features such as mug handles help the network to predict accurate rotations, which tend to be neglected from geometric features. The performances of bottle and can category between our method and CPPF are similar, because of relative simple geometry of bottles and cans. For the bowl category, ours are comparable with NOCS and outperforms CPPF. The camera category is challenging and our method are slightly better than NOCS and CPPF which are almost zero. Overall our method surpasses the baselines and shows great improvements on challenging categories.

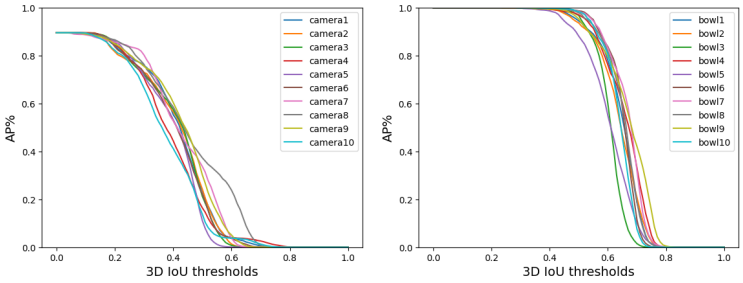


Fig. 3: Visualization of 3D IoUs for all the ten template instances in the camera and mug category.

CategoryMetric	$3D^*_{25}$ ↑	$3D^*_{50}$ ↑	5°5cm ↑	10°5cm ↑	15°5cm ↑
Bottle	53.0	46.9	67.4	92.4	97.4
Bowl	100.0	79.2	25.3	75.3	91.0
Camera	80.4	22.4	0.0	0.2	3.9
Can	65.8	46.9	66.6	83.7	91.1
Laptop	100.0	94.8	9.2	71.5	90.3
Mug	93.7	88.7	4.3	37.6	67.8
Average	82.1	63.2	28.8	60.1	73.6

Table 1: Evaluation results on NOCS REAL275 dataset for each category

	$3D^*_{25}$ ↑	$3D^*_{50}$ ↑	5°5cm ↑	10°5cm ↑	15°5cm ↑
Mask R-CNN	81.9	**64.3**	**31.2**	**61.6**	**74.2**
HQ-SAM	**82.1**	63.2	28.8	60.1	73.6

Table 2: Ablation on segmentation masks

Translation AP Comparison per Category As is visualized in Fig. 5. Our method outperforms the baselines for bowl, bottle, laptop, mug categories and has similar results for the camera and can categories. Overall our translation prediction is robust for the NOCS categories.

Template Sensitivity For ablation on template sensitivity, we iterate through all the ten templates shown in Fig. 1 from supplementary for each category and calculate averaged evaluation results for comparison. To visualize performances between different templates, the 3D IoUs for ten templates of camera and bowl categories are plotted in Fig. 3. Camera templates have large shape variations such as 1st camera and 7th camera, however there are only small differences in prediction results as shown in the left of Fig. 3. This is because testing cameras in the NOCS dataset also have various shapes. The challenge of in-category variation is reflected in a lower averaged metric of camera category.

Segmentation Masks To explore the influence of segmentation masks, we evaluate our model with original Mask RCNN results along with HQ-SAM [2] refined masks and update the result in Tab. 2. Experiment results show that the metrics such as $3D^*_{50}$ increase by 1.1% without HQ-SAM masks.

Category	Metric	$3D_{25}^{+} \uparrow$	$3D_{50}^{+} \uparrow$	5°5cm ↑	10°5cm ↑
Bottle		92.9	82.3	72.0	89.4
Bowl		99.2	91.4	26.9	67.0
Camera		47.3	13.4	0.0	0.1
Laptop		98.6	94.0	50.1	83.1
Mug		85.0	57.2	4.9	15.5
Average		84.6	67.7	30.8	51.0

Table 3: Evaluation results on Wild6D dataset for each category

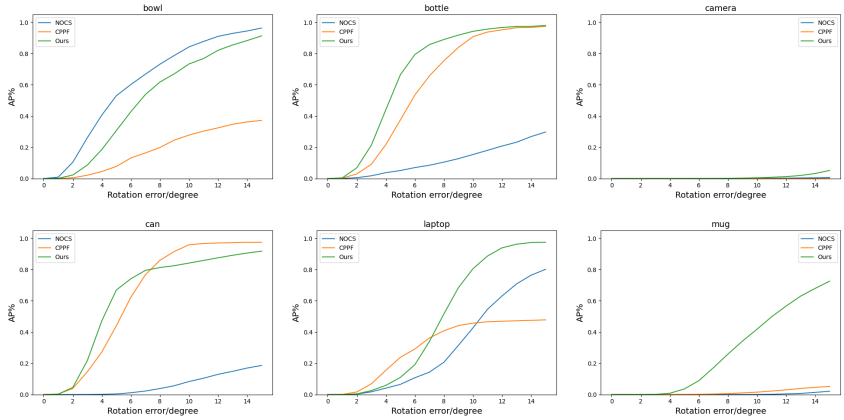


Fig. 4: Rotation average precision visualizations for each NOCS dataset category, comparing CPPF, NOCS and ours.

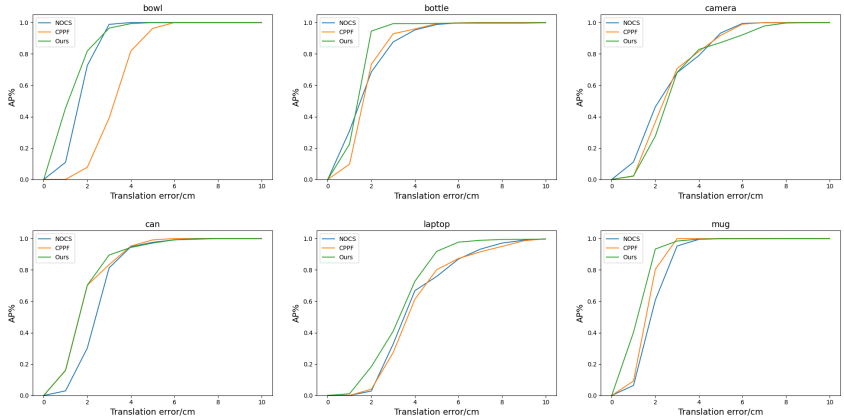


Fig. 5: Translation average precision visualizations for each NOCS dataset category, comparing CPPF, NOCS and ours.

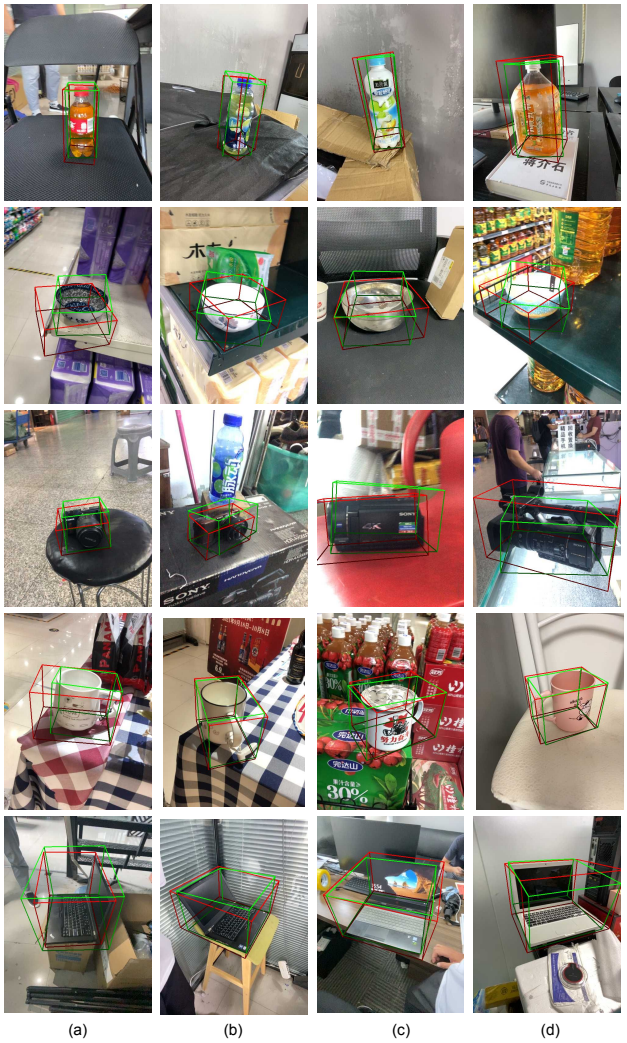


Fig. 6: Visualization of predicted 3D bounding boxes on Wild6D dataset. Green is predicted, red is the ground truth.

3 Detailed Evaluations on Wild6D Dataset

Since Wild6D dataset features a large variety of object shapes and textures in the object models and provides 162 objects among 486 sequences for the testing, the evaluation results reflect the generalization ability of the methods on both object geometries and appearances. The detailed evaluation results for the bottle, bowl, camera, laptop and mug categories are reported in Tab. 3. The evaluation result shows that the bottle category has the highest score as 72 % for the 5°5cm

(a) (b) (c) (d)

Fig. 7: Visualization of predicted 3D bounding boxes within challenging scenes on Wild6D dataset. Green is predicted, red is the ground truth.

metric. The mug and camera categories have low 5°5cm scores, which is the same as the NOCS dataset. Overall the average $3D_{25}$, $3D_{50}$, 5°5cm, 10°5cm scores are 84.6%, 67.7%, 30.8%, 51.0% on Wild6D dataset are comparable with NOCS dataset evaluations, which shows the robust generalization ability on diverse object shapes and textures of our method.

4 More Visualizations on Wild6D Dataset

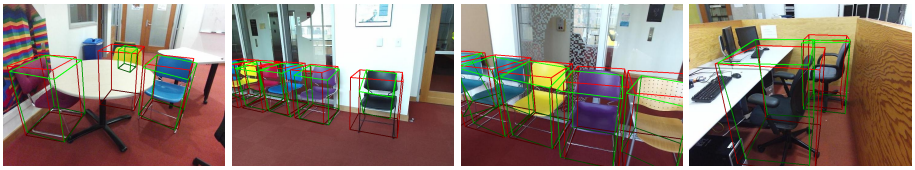
More evaluations of our method on Wild6D dataset are visualized in Fig. 6. The evaluation shows our result on objects with diverse textures and challenging materials such as bottle (b) and bowl (c) from Fig. 6. Our methods predicts object poses robustly under the overexposure of the camera in bowl (b) and (d) from Fig. 6. It is observed that the ground truth poses are inaccurate in certain frames, such as bowl (a) from Fig. 6. The camera category is challenging because of the large shape variations in the dataset, as shown in camera (c) and (d) from Fig. 6.

Challenging Scenes As is shown in Fig. 7, our method fails for certain frames. In Fig. 7 (a), there are cases where the ground truth does not perfectly match the object. Detection failures cause the prediction to fail, as shown in Fig. 7 (c). Transparent objects such as (b) and (d) from Fig. 7 lead to the depth measurements failures and inaccurate pose estimations of our method.

5 More Visualizations on SUN RGB-D Dataset

Evaluation results on SUN RGB-D dataset are visualized in Fig. 8, our method predicts the object pose accurately even under partial occlusions. However, heavy occlusions lead to inferior results.

Challenging Scenes The scenes are challenging for our method on certain frames and the results are visualized in Fig. 9. Objects that are only visible from the corner such as Fig. 9 (a) are hard for prediction. The chairs stacked together or heavily occluded such as (b) and (c) from Fig. 9 lead to heavy occlusions.



(a) (b) (c) (d)

Fig. 8: Visualization of predicted 3D bounding boxes on SUN RGB-D dataset. Green is predicted, red is the ground truth.

(a) (b) (c) (d)

Fig. 9: Visualization of predicted 3D bounding boxes within challenging scenes on SUN RGB-D dataset. Green is predicted, red is the ground truth.

Object pose annotation errors such as the chair scales in Fig. 9 (d) decrease our $3D^{+}_{10}$, $3D^{+}_{25}$ scores on certain frames.

References

1. Chang, A.X., Funkhouser, T.A., Guibas, L.J., Hanrahan, P., Huang, Q.X., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. ArXiv **abs/1512.03012** (2015), `https://api.semanticscholar.org/CorpusID:2554264`
2. Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y.W., Tang, C.K., Yu, F.: Segment anything in high quality. In: NeurIPS (2023)
3. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 567–576 (2015), `https://api.semanticscholar.org/CorpusID:6242669`
4. Wang, H., Sridhar, S., Huang, J., Valentin, J.P.C., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2637–2646 (2019), `https://api.semanticscholar.org/CorpusID:57761160`
5. You, Y., Shi, R., Wang, W., Lu, C.: Cppf: Towards robust category-level 9d pose estimation in the wild. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6856–6865 (2022), `https://api.semanticscholar.org/CorpusID:247291938`
6. Ze, Y., Wang, X.: Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. Advances in Neural Information Processing Systems **35**, 27469–27483 (2022)