Supplementary Material of RBP-Pose: Residual Bounding Box Projection for Category-level Pose Estimation

Ruida Zhang[1*], Yan Di[2*], Zhiqiang Lou[1], Fabian Manhardt[3], Federico Tombari[2,3], and Xiangyang Ji[1]

[1] Tsinghua University
[2] Technical University of Munich
[3] Google
{zhangrd21@mails. lzq20@mails. xyji@}tsinghua.edu.cn
shangbuhuan13@gmail.com, fabianmanhardt@google.com, tombari@in.tum.de

1 Details of Non-linear Shape Augmentation.

We propose two types of augmentation strategies for categories provided by the NOCS dataset [5]: axis-based non-linear scaling transformation ($A1$) for *camera, bottle, can, bowl, mug* and plane-based rotation transformation ($A2$) for *laptop*. Since $A1$ has been introduced in the main text, here we present the details of $A2$ transformation (Fig 1).

laptop is an articulated object consisting of two rigid planes, so we conduct shape augmentation by adjusting the angle between the upper and lower plane (Figure 1 (III)). Thereby, we rotate the upper plane by a certain angle along the fixed axis, while the lower plane remains static. In canonical space, the fixed axis is parallel to z-axis, $x - a_x = y - a_y = 0$, where a_x and a_y are constants determined by the object model. The transformation function $\mathcal{T}_{A2}(P)$ for point $P = \{P_x, P_y, P_z\}$ on the upper plane is defined as,

$$\begin{aligned}\mathcal{T}_{A2}(P) = \{&\gamma_1(a_x - \sin\theta(P_y - a_y) + \cos\theta(P_x - a_x)),\\ &\gamma_2(a_y + \sin\theta(P_x - a_x) + \cos\theta(P_y - a_y)), \gamma_3 P_z\}\end{aligned} \tag{1}$$

θ is the random variable which controls the rotation angle and $\gamma_1, \gamma_2, \gamma_3$ are the random variables controlling the size transformation. For implementation, we uniformly sample $\theta \sim \mathcal{U}(-15°, 15°)$ and $\gamma_1, \gamma_2, \gamma_3 \sim \mathcal{U}(0.8, 1.2)$.

After shape augmentation, we augment the input point cloud, the ground truth normalized shape model and NOCS coordinates correspondingly. Specifically, we first transform the points to the canonical coordinate space and then conduct deformation. The augmented point P_{aug} of point P is

$$P_{aug} = R(\mathcal{T}(R^{-1}(P - t))) + t \tag{2}$$

[*] Authors with equal contributions.
Codes are released at https://github.com/lolrudy/RBP_Pose.

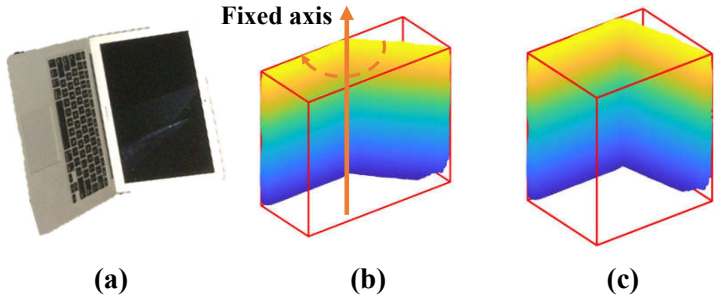

Fig. 1. Demonstration of non-linear data augmentation. The figure shows non-linear data augmentation for *laptop*. For the instances **(a)**, we augment the object shape in our non-linear manner from **(b)** to **(c)**, by rotating the upper plane along the fixed axis.

where R, t is the ground truth pose and $\mathcal{T}$ is the deformation function.

After conducting the shape augmentation, we re-normalize the object model. The augmented model point M_{aug} of point M is,

$$M_{aug} = \mathcal{T}(M)/L_{aug} \tag{3}$$

where L_{aug} is the diagonal length of the object's bounding box after augmentation. The corresponding NOCS coordinates are transformed similarly.

2 Details of Other Data Augmentation

Besides the non-linear shape augmentation, we add random Gaussian noise to the input point cloud, and employ random rotational and translational perturbations. We set the probabilities to conduct each kind of augmentation to be 0.3.

To add **random Gaussian noise**, we sample point-wise random noise from the Gaussian distribution $N(0mm, 2mm)$ and add it to the original coordinates.

In **random rotational and translational perturbations**, we apply random rotation R_{aug} and translation t_{aug} to the ground truth pose R_{gt}, t_{gt}. For a point P, its corresponding point after augmentation is,

$$\mathcal{T}_{rt}(P) = R_{aug}(R_{gt}^{T}(P - t_{gt})) + t_{aug} \tag{4}$$

The random rotation R_{aug} is computed by sampling three Euler angles from the uniform distribution $U(-15°, 15°)$ respectively and then converting them into the rotation matrix. The random translation t_{aug} is sampled from the uniform distribution $U(-50mm, 50mm)$.

3 Supplementary Results on NOCS

We provide per-category results of RBP-Pose on REAL275 and CAMERA25 in Tab. 1, 2 respectively.

category	IoU_{25}	IoU_{50}	IoU_{75}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
bottle	57.7	57.6	47.5	38.7	43.5	76.4	89.8
bowl	100	100	99.2	75.4	81.7	92.1	100
camera	90.9	86.4	60.3	1.3	1.5	21.4	27.1
can	71.4	71.4	63.7	53.5	67.1	78.8	96.3
laptop	86.0	85.2	44.8	41.3	75.2	44.3	94.5
mug	99.4	98.0	91.3	18.9	19.4	65.7	67.6
average	84.2	83.1	67.8	38.2	48.1	63.1	79.2

Table 1. Per-category results of RBP-Pose on REAL275 dataset.

category	IoU_{25}	IoU_{50}	IoU_{75}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
bottle	93.9	93.8	90.3	81.2	96.7	82.0	97.8
bowl	96.9	96.8	96.7	97.9	98.3	99.1	99.6
camera	94.6	84.9	74.0	50.5	55.7	64.4	72.5
can	92.5	92.4	92.2	98.7	99.2	99.0	99.5
laptop	98.1	97.0	88.9	63.5	78.5	72.6	91.7
mug	94.1	93.7	91.6	49.3	49.4	75.8	76.0
average	95.0	93.1	89.0	73.5	79.6	82.1	89.5

Table 2. Per-category results of RBP-Pose on CAMERA25 dataset.

4 More Qualitative Results

We show more qualitative results on REAL275 in Fig. 2, 3, comparing our method and SGPA [1].

5 Failure Cases and Limitations

Dependency on the Detection Results. We employ an off-the-shelf detector MaskRCNN [2] to segment out the objects of interest and then back-project their corresponding depth values to generate the input object point clouds. We only use the real dataset to train the detector like the previous methods [1,4,3]. As a result, our pose estimation accuracy depends on the detection result. If the object is not detected or classified incorrectly, RBP-Pose may fail to output reasonable results. Moreover, if the object is only partially detected, our pose estimation tends to predict a smaller size since the input point cloud is incomplete. Under the 2D IoU accuracy threshold of 50%, the recall rates of the detection results of object *bottle* and *can* in the NOCS dataset are 57.7% and 71.4% respectively, which leads to significant limitations on the pose accuracy under all metrics, especially for the 3D IoU metric. In a nutshell, we employ an off-the-shelf detector to enable fair comparison with previous methods [1,4,3], but adopting a better detector might improve our accuracy.

Modality of Input Data. The input of RBP-Pose is the point cloud generated by backprojecting the segmented depth map. The advantages are two-fold

in comparison with RGB-D methods [1,4]. First, this strategy supports point-cloud based data augmentation, which enables effective training on only a small amount of data and avoids over-fitting. Second, without the feature extraction on RGB image, RBP-Pose achieves the fastest inference speed in all category-level pose estimation methods. However, this strategy also suffers from two shortcomings in comparison with RGB-D methods [1,4]. First, the translation is sometimes hard to predict for certain categories, *e.g. laptop*. Since RBP-Pose predicts the residual translation between the real translation and the mean coordinate of the input point cloud, which is hard to predict if the object center is outside the object and the mean coordinate is far from the real object center (*e.g. laptop*). Consequently, for object *laptop*, our accuracy under IoU_{75} is 44.8, while RGB-D based method SGPA [1] is about 59.0. Second, we uniformly sample the input point cloud on the segmented depth map. The sampling process introduces randomness and undermines robustness. In conclusion, our point cloud based pose estimator boosts the inference speed, but incorporating RGB information may improve the accuracy.

References

1. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2773–2782 (2021) 3, 4, 5, 6
2. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 3
3. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y.: Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. arXiv preprint arXiv:2103.06526 (2021) 3
4. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: European Conference on Computer Vision. pp. 530–546. Springer (2020) 3, 4
5. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019) 1

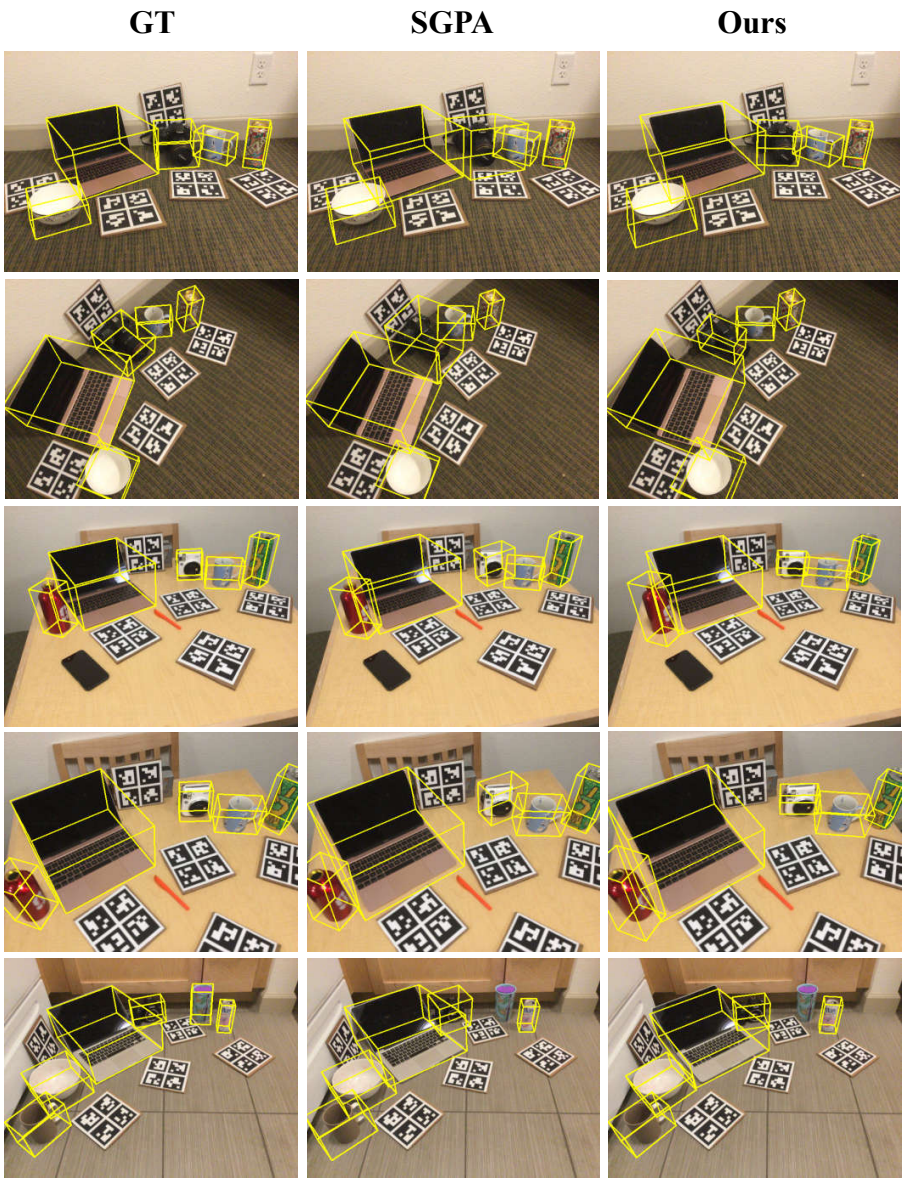

Fig. 2. Qualitative comparison of our method and SGPA [1] on REAL275.

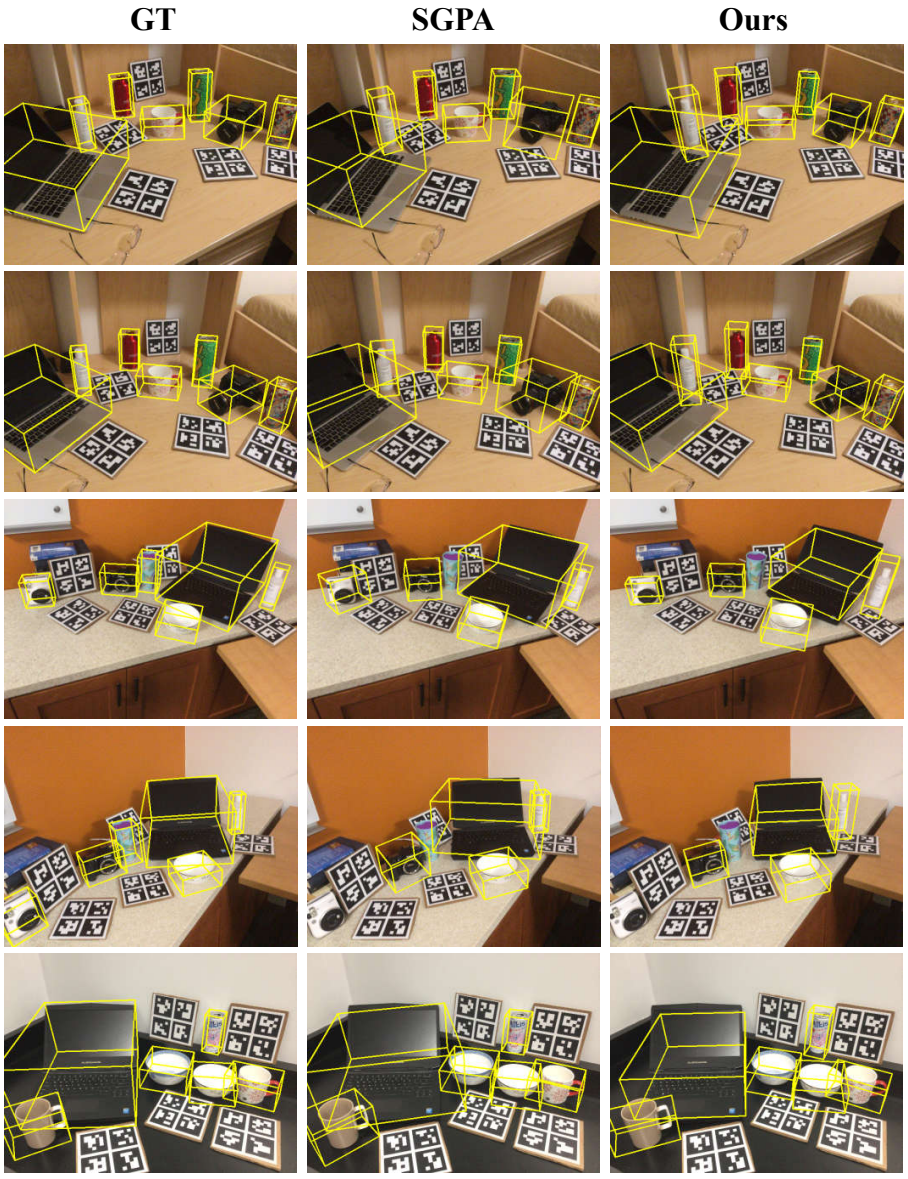

Fig. 3. Qualitative comparison of our method and SGPA [1] on REAL275.