Supplemental Materials
SOCS: Semantically-aware Object Coordinate Space for Category-Level 6D Object Pose Estimation under Large Shape Variations

Boyan Wan* Yifei Shi* Kai Xu†
National University of Defense Technology
wanboyan@163.com {yifei.j.shi, kevin.kai.xu}@gmail.com

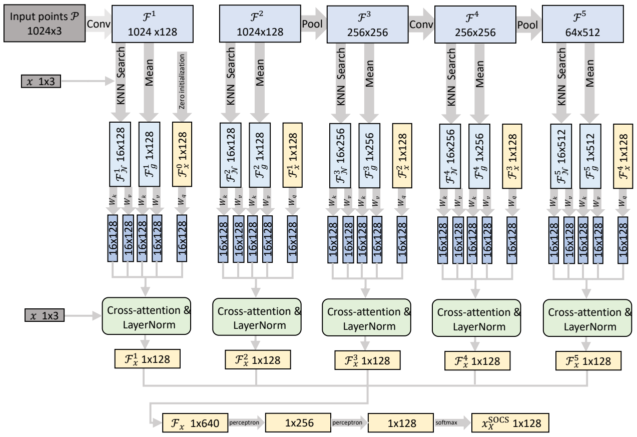

Figure 1: The detailed network architecture of the proposed multi-scale coordinate-based attention network. The numbers in the rectangles denote the feature length.

1. Method Details

Detailed network architecture. To facilitate the reproduction of our work, we show the detailed network architecture of the multi-scale coordinate-based attention network in Figure 1.

Training data augmentation. Since the object detection/segmentation by Mask-RCNN is imperfect and may cause both over-segmentation or under-segmentation during the inference, we adopt data augmentation to mimic the imperfections of segmentation during the network training. Specifically, we use the dilation and erosion operations described in [1] to generate the imperfect object mask based on the ground-truth mask in the NOCS-REAL275 dataset. We found this data augmentation strategy greatly increases our performance on pose estimation despite the inferior performance on object detection/segmentation.

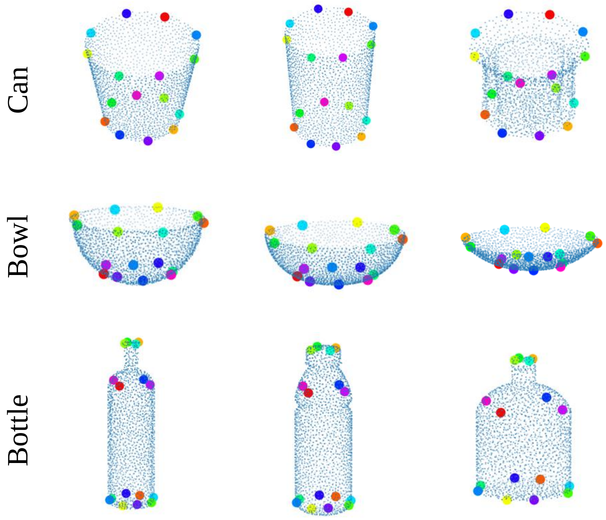

Figure 2: Keypoints detection on symmetric objects.

2. Handling Symmetric Objects

Both the NOCS-REAL275 and Modelnet40-partial datasets contain symmetric objects (i.e., objects with rotational symmetry) that might incur ambiguity on the pose estimation. In this section, we describe how to handle symmetry in terms of keypoints detection and pose estimation, respectively, to alleviate this problem.

Handling symmetry in keypoints detection. To better leverage the keypoints and determine the correspondences between symmetric object instances and the categorical mean shape, as shown in Figure 2, we detect the *symmetric keypoints* on symmetric object instances. To achieve this, during the keypoints detection with Skeleton Merger [3], we first generate $k/4$ keypoints on the object surface. Then,

*Joint first authors
†Corresponding author

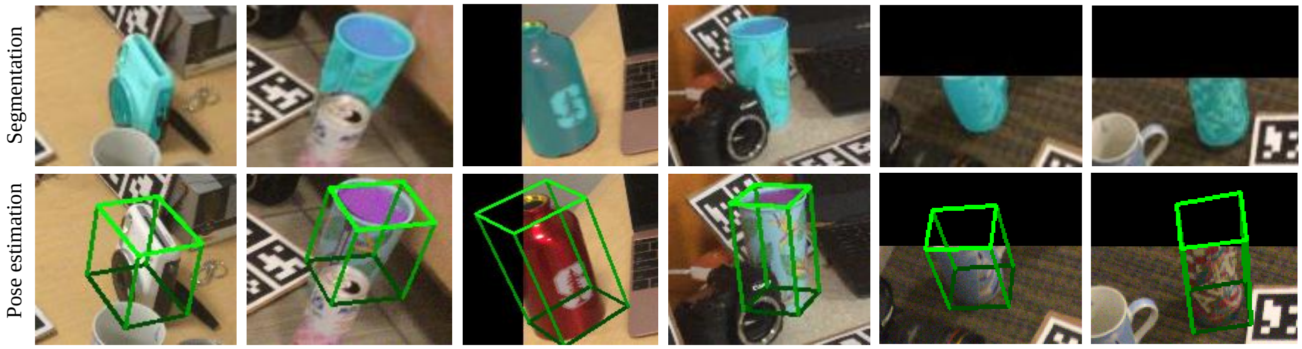

Figure 3: Examples of objects and the pose estimation of our method in the NOCS-REAL275 heavy occlusion subset.

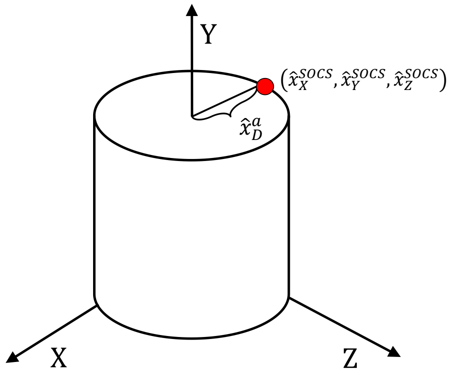

Figure 4: Illustration of handling symmetry in pose estimation.

we rotate the detected keypoints with $90°, 180°, 270°$ along the symmetric axis, respectively, to generate the rest $3/4 \cdot k$ keypoints. Last, all the generated keypoints are optimized by the encoder-decoder network in Skeleton Merger [3]. We found this symmetric keypoints detection trick is able to regularize the per-instance SOCS and thus lead to better pose estimation results. The quantitative comparisons of methods with (Full method) and without (w/o symmetric keypoints) this trick are reported in Table 1.

Handling symmetry in pose estimation. Handling symmetry is a crucial component in many pose estimation approaches. In our method, we tackle this problem by introducing a new output representation and loss function for the rotational symmetric categories. Suppose the object is symmetric w.r.t. the Y-axis in the canonical space (shown in Figure 4). Instead of estimating the absolute coordinate along each axis, for any point x, we estimate its distance to the Y-axis $x_{\mathrm{D}}^{\mathrm{SOCS}}$ as well as the coordinate on Y-axis $x_{\mathrm{Y}}^{\mathrm{SOCS}}$:

$$\begin{aligned} x_{\mathrm{D}}^{\mathrm{SOCS}} &= \mathrm{Softmax}(\mathrm{MLP}_{\mathrm{D}}(\mathcal{F}_x)), \\ x_{\mathrm{Y}}^{\mathrm{SOCS}} &= \mathrm{Softmax}(\mathrm{MLP}_{\mathrm{Y}}(\mathcal{F}_x)), \end{aligned} \tag{1}$$

Table 1: Ablation study of handling symmetry. Experiments are conducted on the categories of symmetric object in the NOCS-REAL275 dataset.

Method	IoU75↑	10°2cm↑
w/o symmetric keypoints	0.72	0.87
w/o symmetric output	0.71	0.80
Full method	0.74	0.89

where $\mathrm{MLP}_{\mathrm{D}}(\cdot)$ and $\mathrm{MLP}_{\mathrm{Y}}(\cdot)$ are the MLPs, $\mathcal{F}_x$ is the extracted feature at x. Therefore, the training loss of SOCS is:

$$\mathcal{L}_{\mathrm{SOCS}} = \sum_{x \in \mathcal{X}} [\mathcal{L}_{\mathrm{CE}}(x_{\mathrm{D}}^{\mathrm{SOCS}}, \hat{x}_{\mathrm{D}}^{\mathrm{SOCS}}) + \mathcal{L}_{\mathrm{CE}}(x_{\mathrm{Y}}^{\mathrm{SOCS}}, \hat{x}_{\mathrm{Y}}^{\mathrm{SOCS}})]. \tag{2}$$

The optimization function for pose and size estimation should be modified accordingly:

$$\min_i \sum_{x \in \mathcal{X}} \left\| \mathbf{T} \cdot \mathbf{S} \cdot \overline{\Phi}(x)_i - x \right\|^2 \tag{3}$$

where $\overline{\Phi}(x)$ denotes a point that lies on the orbit along the symmetry axis whose radius is the estimated $x_{\mathrm{D}}^{\mathrm{SOCS}}$ and the coordinate on Y-axis is $x_{\mathrm{Y}}^{\mathrm{SOCS}}$. Its exact location on the orbit is estimated along with the object pose and size in the optimization. The quantitative comparisons of methods with (Full method) and without (w/o symmetric output) this modified loss function are reported in Table 1.

3. More Results on NOCS-REAL275

We provide more results on the NOCS-REAL275 dataset in this section. There are several phenomena we can observe from the results. First, as mentioned in the paper, the NOCS-REAL275 subset containing objects with heavy occlusions is visualized in Figure 3. It is clear that the pose estimation on these objects could be challenging due to

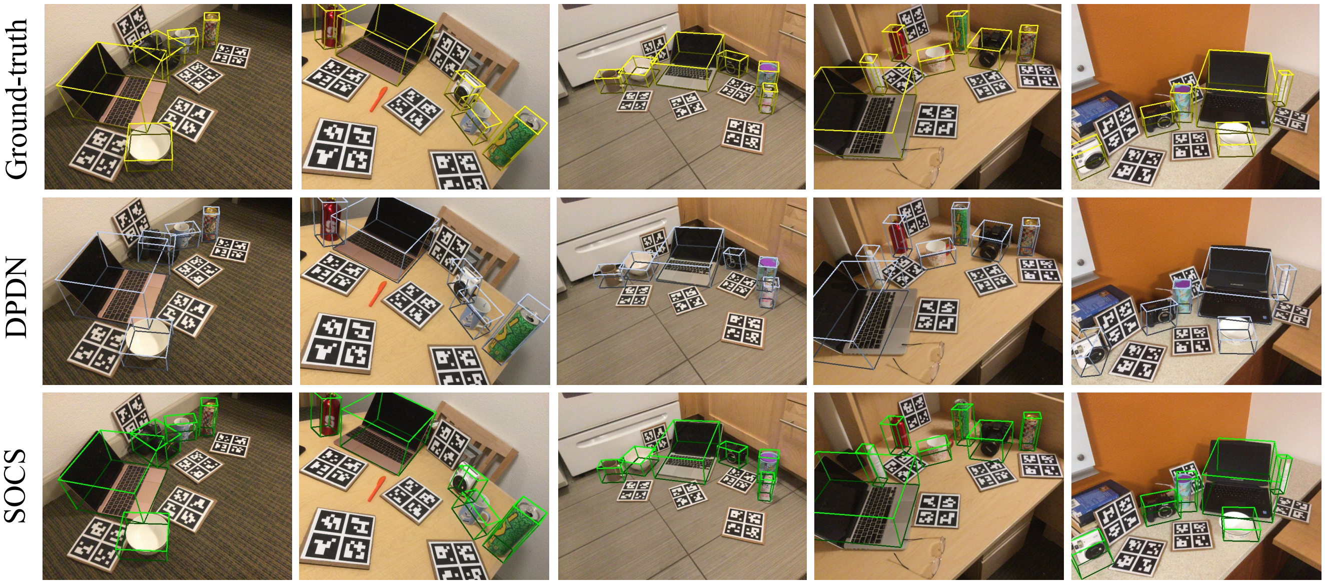

Figure 5: More visual comparisons to the state-of-the-art on the NOCS-REAL275 dataset.

the occlusion. Second, more qualitative comparisons are shown in Figure 5. We see that our method outperforms the NOCS-based baseline method DPDN [2] in most of the cases, demonstrating the advantages of the proposed SOCS. Third, we report the per-category quantitative performances in Table 2. The results show that our method produces accurate results in all metrics over all categories.

Table 2: Per-category Quantitative Performances on the NOCS-REAL275 dataset.

Category	IoU75↑	5°2cm↑	5°5cm↑	10°2cm↑
Bottle	0.52	0.51	0.53	0.83
Bowl	1.00	0.88	0.92	0.96
Camera	0.63	0.02	0.02	0.20
Can	0.70	0.67	0.75	0.88
Laptop	0.68	0.64	0.90	0.66
Mug	0.94	0.25	0.26	0.80
Average	0.75	0.49	0.56	0.72

4. More Results on Modelnet40-partial

The qualitative comparisons to the state-of-the-art on the Modelnet40-partial dataset are visualized in Figure 6. The Modelnet40-partial dataset contains categories with large shape variations, making the pose estimation on unseen instances difficult. The results show that our method outperforms the RBP-Pose [4], especially in categories with large shape variations, such as airplane and sofa. We also provide the per-category quantitative performances on the ModelNet40-partial dataset in Table 3.

Table 3: Per-category Quantitative Performances on the ModelNet40-partial dataset.

Category	Rotation		Translation	
	Mean(°) ↓	5° ↑	Mean(m)↓	5°0.05m ↑
Airplane	3.22	0.94	0.03	0.43
Car	41.29	0.26	0.03	0.15
Chair	10.37	0.65	0.04	0.23
Sofa	42.01	0.24	0.04	0.06
Bottle	15.76	0.87	0.03	0.45
Average	22.53	0.59	0.03	0.26

References

[1] 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. 1

[2] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Is-*

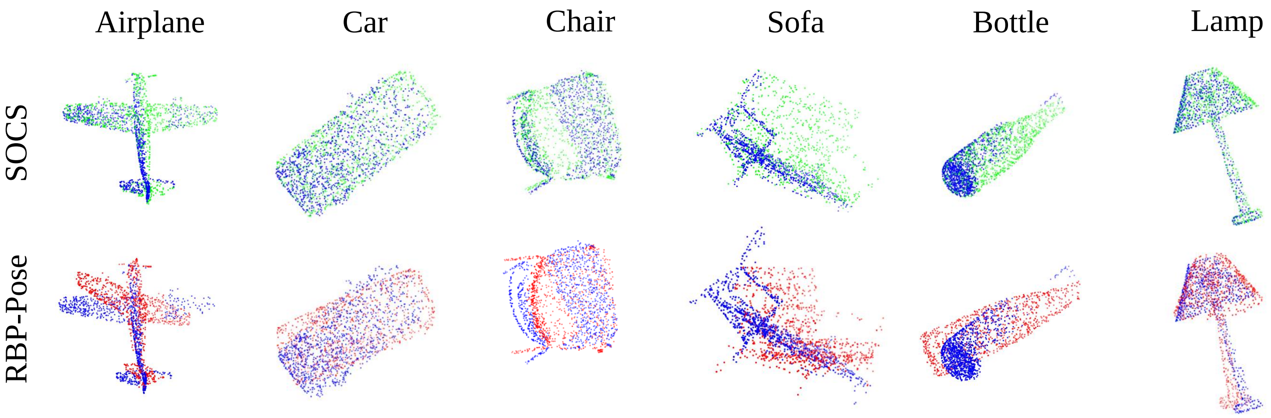

Figure 6: Visual comparisons to the state-of-the-art on the Modelnet40-partial dataset. Blue points are the input. Green and red points denote the transformed points estimated by our method and RBP-Pose [5], respectively.

rael, October 23–27, 2022, Proceedings, Part IX, pages 19–34. Springer, 2022. 3

[3] Ruoxi Shi, Zhengrong Xue, Yang You, and Cewu Lu. Skeleton merger: an unsupervised aligned keypoint detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 43–52, 2021. 1, 2

[4] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 655–672. Springer, 2022. 3

[5] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 655–672. Springer, 2022. 4