Supplementary Materials of Generative Category-level Object Pose Estimation via Diffusion Models

Anonymous Author(s)
Affiliation
Address
`email`

Contents

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

1 Implementation Details

Architecture of the Score Network The detailed architecture of the score network $\mathbf{\Phi}_\theta$ is illustrated in Figure 1. We utilize PointNet++ [1] to extract the global geometry feature $\mathcal{F}_O$ of the partially observed point cloud O^*. And the sampled pose $\mathbf{p}$ and timestep t features are embedded as $\mathcal{F}_{\mathbf{p}}$ and $\mathcal{F}_t$, respectively, using a Multi-Layer Perceptron (MLP). Then $\mathcal{F}_O$, $\mathcal{F}_{\mathbf{p}}$ and $\mathcal{F}_t$ are concatenated to obtain the global feature $\mathcal{F}$, and three parallel branches are employed to predict the scores of R_x, R_y, and T individually, where $[R_x|R_y] \in \mathbb{R}^6$ and $T \in \mathbb{R}^3$ denote rotation and translation vectors, respectively. And $[R_x|R_y]$ is an continuous rotation representation proposed by [2] to address the discontinuity of quaternions and Euler angles in Euclidean space. As introduced in [2], the mapping from SO(3) to the 6D representation of rotation is:

$$g_{GS}\left(\begin{bmatrix}\mathbf{a_1} & \mathbf{a_2} & \mathbf{a_3}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{a_1} & \mathbf{a_2}\end{bmatrix} \tag{1}$$

The mapping form the 6D representation to SO(3) is:

$$f_{GS}\left(\begin{bmatrix}\mathbf{a_1} & \mathbf{a_2}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{b_1} & \mathbf{b_2} & \mathbf{b_3}\end{bmatrix} \tag{2}$$

$$b_i = \left[\begin{cases} N(\mathbf{a_1}) & \text{if } i = 1 \\ N(\mathbf{a_2} - (\mathbf{b_1} \cdot \mathbf{a_2})\mathbf{b_1}) & \text{if } i = 2 \\ \mathbf{b_1} \times \mathbf{b_2} & \text{if } i = 3 \end{cases}\right] \tag{3}$$

Here $N(\cdot)$ denotes a normalization function.

Architecture of the Energy Network The energy network $\mathbf{\Psi}_\phi$ shares exactly the same architecture with the score network $\mathbf{\Phi}_\theta$. The inputs are first fed into $\mathbf{\Phi}_\phi$ to obtain a score-shaped vector $\mathbf{\Phi}_\phi(\mathbf{p}, t|O) \in \mathbb{R}^{|\mathcal{P}|}$. Then, the output energy is calculated by the dot product between the input pose and the score-shaped vector $\mathbf{\Psi}_\phi(\mathbf{p}, t|O) = \langle \mathbf{p}, \mathbf{\Phi}_\phi(\mathbf{p}, t|O)\rangle \in \mathbb{R}^1$.

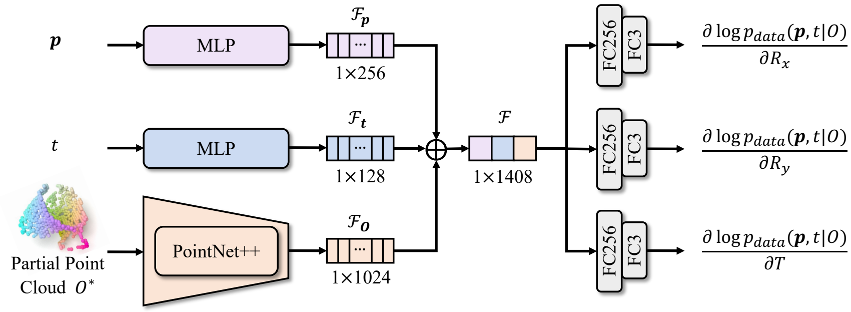

$\boldsymbol{p}$
MLP
$\mathcal{F}_{\boldsymbol{p}}$
1×256
t
MLP
$\mathcal{F}_t$
1×128
Partial Point Cloud O^*
PointNet++
$\mathcal{F}_O$
1×1024
$\oplus$
$\mathcal{F}$
1×1408
FC256
FC3
$\frac{\partial \log p_{data}(\boldsymbol{p}, t|O)}{\partial R_x}$
$\frac{\partial \log p_{data}(\boldsymbol{p}, t|O)}{\partial R_y}$
$\frac{\partial \log p_{data}(\boldsymbol{p}, t|O)}{\partial T}$

Figure 1: **Architecture of the score network $\mathbf{\Phi}_\theta$.** $\mathbf{p}$ denotes sampled 6D object poses. O^* denotes partially observed 3D point cloud condition. t denotes timestep. $\oplus$ denotes the concatenation operator.

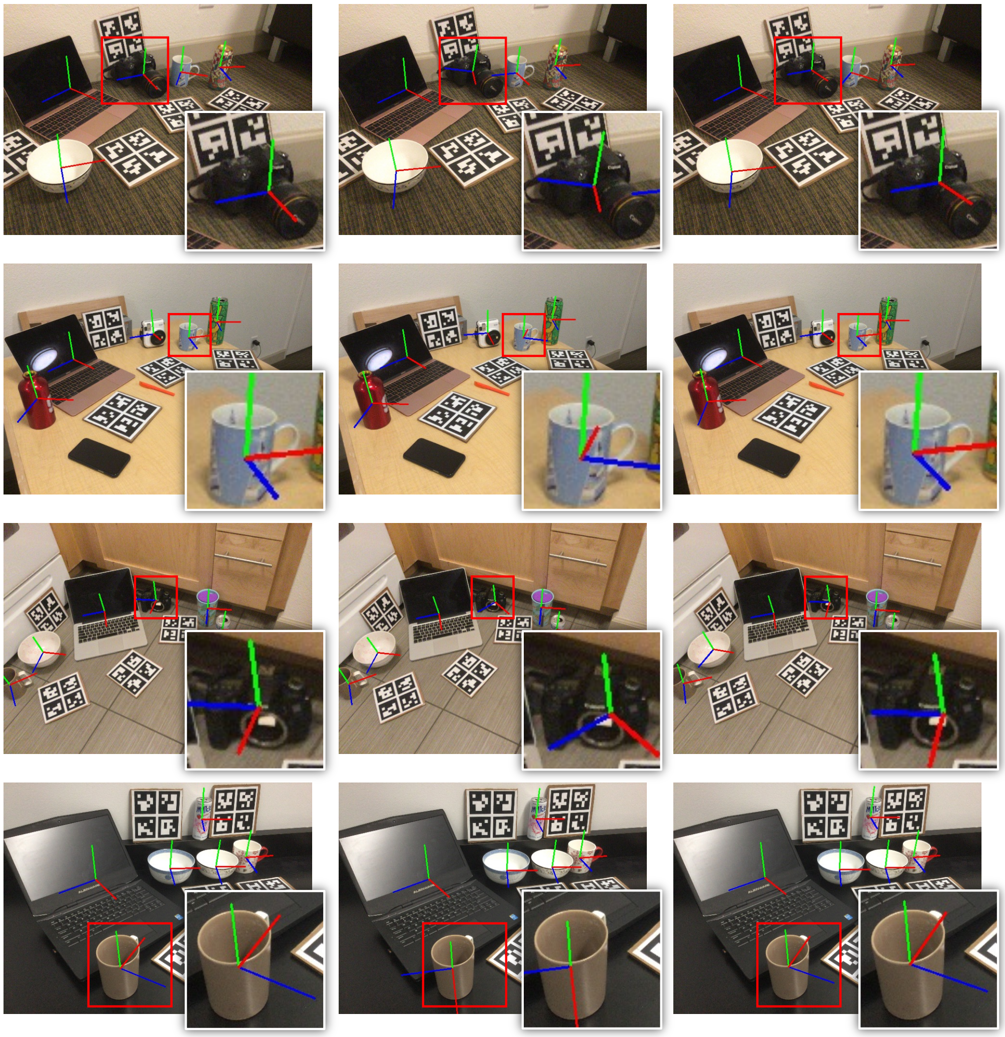
2 Qualitative Comparison on REAL275

Figure 2 illustrates the qualitative comparison results between our method and RBP-Pose [3] on the REAL275 dataset. The images are accompanied by red boxes highlighting objects that exhibit noticeable differences in the predicted results. Additionally, the bottom-right corner of each image provides an enlarged view of the highlighted object, showing the ground truth pose as well as the poses estimated by RBP-Pose and our approach. Our method demonstrates a significant performance improvement compared to RBP-Pose, particularly in the case of objects such as mugs. Notably, in the fourth row of the figure, it can be observed that our method achieves highly accurate poses even when only a small portion of the mug handle is visible. This success can be attributed to the fact that, during the training process, a unique pose exists when the mug handle is visible. However, when the mug handle becomes occluded, a multi-hypothesis problem arises, which our generative formulation effectively handles.

3 More Results and Analysis

3.1 Per-category Results

Figure 3 demonstrates a quantitative comparison between our method and the state-of-the-art depth-based approach, RBP-Pose [3], for various object categories at different thresholds. The results



GT RBP-Pose Ours

Figure 2: **Qualitative comparison with RBP-Pose [3] on REAL275.** The left column represents the ground truth pose, the middle column represents the results of RBP-Pose, the right column represents the results of our approach.

clearly indicate that our method outperforms RBP-Pose in all metrics, despite the fact that we do not incorporate augmentation specifically designed for symmetric objects during the training phase, unlike RBP-Pose. Our approach exhibits significant improvements, particularly in regions with stringent threshold requirements. This emphasizes the superior performance of our generative category-level object 6D pose estimation approach in effectively addressing the multi-hypothesis challenges posed by symmetric objects and partial observations, thereby enabling its successful application in robot manipulation tasks demanding precise object pose prediction.(*e.g.*, pouring liquids.)

3.2 Results on CAMERA

Table 1 illustrates a quantitative comparison between our method and the baselines on the CAMERA [4] dataset. The results clearly demonstrate the remarkable performance enhancement achieved by our method. When compared to approaches that rely solely on depth data as network input, as well as those that utilize RGB-D and shape priors as network input, our method consistently outperforms them, surpassing the current state-of-the-art performance. Notably, our method exhibits a particularly pronounced advantage when stricter accuracy requirements are imposed, such as the $5°2cm$ metric. In this case, our method outperforms the current SOTA method, RBP-Pose, by an impressive margin of 6.4%. This significant improvement highlights the efficacy of our approach.

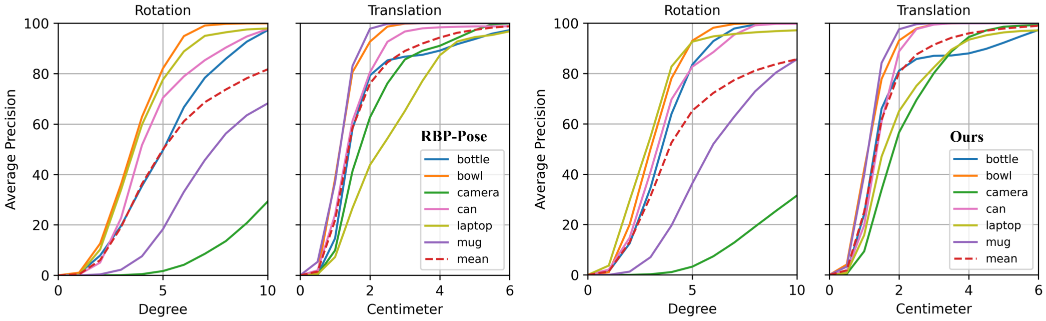


Figure 3: **Per-category quantitative comparison with RBP-Pose [3] on REAL275.** The left represents the results of RBP-Pose, while the right represents the results of our approach.

Table 1: **Quantitative comparison of category-level object pose estimation on CAMERA [4] dataset.** We summarize the results reported in the original paper for the baseline method. ↑ represents a higher value indicating better performance, while ↓ represents a lower value indicating better performance. **Data** refers to the format of the input data used by the method, and **Prior** indicates whether the method requires category prior information. '-' indicates that the metrics are not reported in the original paper. **K** represents the number of hypotheses."

Method		Data	Prior	5°2cm↑	5°5cm↑	10°2cm↑	10°5cm↑	Parameters(M)↓
Deterministic	NOCS [4]	RGB-D	×	32.3	40.9	48.2	64.6	-
	DualPoseNet [5]	RGB-D	×	64.7	70.7	77.2	84.7	67.9
	SPD [6]	RGB-D	✓	54.3	59.0	73.3	81.5	18.3
	CR-Net [7]	RGB-D	✓	72.0	76.4	81.0	87.7	-
	SGPA [8]	RGB-D	✓	70.7	74.5	82.7	88.4	-
	GPV-Pose [9]	D	×	72.1	79.1	-	89.0	-
	SAR-Net [10]	D	✓	66.7	70.9	75.3	80.3	6.3
	SSP-Pose [11]	D	✓	64.7	75.5	-	87.4	-
	RBP-Pose [3]	D	✓	73.5	79.6	82.1	89.5	-
Probabilistic	Ours	D	×	**79.9**	**84.4**	**84.6**	**89.6**	**4.4**
	Ours(K=10)	D	×	90.8	93.0	93.4	95.7	2.2
	Ours(K=50)	D	×	95.5	96.4	97.2	98.2	2.2

3.3 Real World Experiments

We have also successfully integrated our approach with robot manipulation capabilities, as demonstrated through various experiments conducted with the UFACTORY xArm6 equipped with RealSense D435. **The demonstrations can be found in the supplementary video or on the project website.** As shown in Figure 4, we illustrate the following three tasks:

Pouring Task. This task involves transferring the contents (*e.g.*, water) from one container to another. The demonstration highlights the potential of combining our approach with heuristic strategies, enabling functional robot operations.

Stacking Task. In this task, we focused on piling up objects of the same category, like organizing scattered bowls on a tabletop. This demonstrates the precision of the estimated object pose, as accurate knowledge of object poses is crucial for completing this task.

Handover Task. This task involved either receiving objects from human hands to perform tasks or passing objects to person. The demonstration exemplified one form of human-robot interaction empowered by our method.

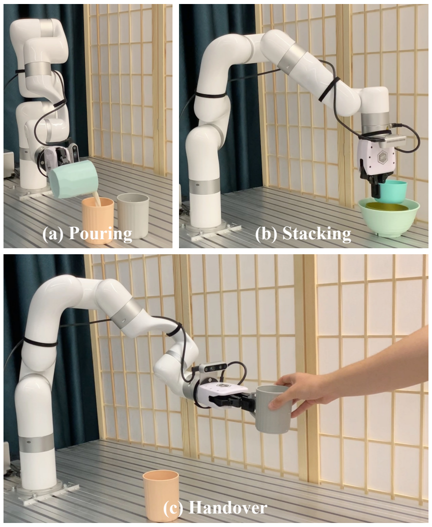


Figure 4: **Pose estimation for robot manipulation tasks.** We demonstrate three types of tasks.

4 Ethics Statement and Boarder Impact

Our method has the potential to develop home-assisting robot, thus contributing to social welfare. We evaluate our method in synthesized or human-collected datasets, which may introduce data bias. However, similar studies also have such general concerns. We do not see any possible major harm in our study.

References

[1] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[2] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.

[3] 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] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.

[5] 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/CVF International Conference on Computer Vision*, pages 3560–3569, 2021.

[6] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020.

[7] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021.

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

[9] 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.

[10] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: shape alignment and recovery network for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2022.

[11] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7452–7459. IEEE, 2022.