Supplementary Material: Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation

Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee

National University of Singapore, Singapore
tianmeng@u.nus.edu, {mpeangh, gimhee.lee}@nus.edu.sg

1 Comparison to CASS

CASS [1] is the latest work on category-level 6D object pose and size estimation. Similar to our work, they reconstruct the complete object model in the canonical space as a by-product. However, they train a variational autoencoder to generate the point cloud, while we estimate the deformation field of the corresponding shape prior. In addition, they directly regresses the pose and size by comparing pose-independent and pose-dependent features, while we recover the pose by establishing dense correspondences. As shown in Table 1, our approach significantly outperforms CASS in pose accuracy. This demonstrates the superiority of our correspondence-based approach over direct pose regression.

Table 1. Quantitative comparison with CASS on REAL275.

Method	**mAP**						
	$3D_{25}$	$3D_{50}$	$3D_{75}$	5° 2cm	5° 5cm	10° 2cm	10° 5cm
Baseline [4]	**84.8**	**78.0**	30.1	7.2	10.0	13.8	25.2
CASS [1]	84.2	77.7	–	–	13.0	–	37.6
Ours	83.4	77.3	**53.2**	**19.3**	**21.4**	**43.2**	**54.1**

2 Comparison to 6-PACK

6-PACK [3] is the state-of-the-art category-level 6D pose tracker. Although our approach does not require pose initialization nor leverages on temporal consistency, we still achieve comparable accuracy on REAL275 at 5° 5cm (30.4% compared to 33.3%). More importantly, the accuracy of 6-PACK drops below 30% when the first 40 frames of a sequence (460 frames on average) are excluded from evaluation. This indicates that 6-PACK is highly dependent on pose initialization for higher accuracy. In contrast, the accuracy of our method remains stable since it is a pose estimation method.

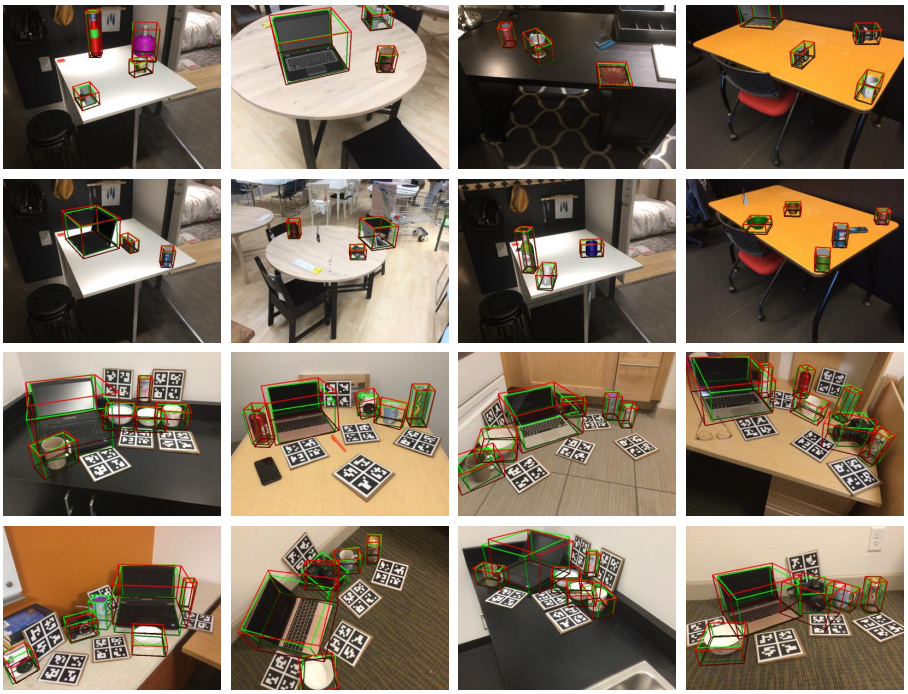

Fig. 1. We show some qualitative results of our approach (red) and their ground truths (green) on CAMERA25 (top two rows) and REAL275 (bottom two rows).

3 Qualitative Results

Fig. 1 shows the per-frame pose detection results of our approach. The results are better on synthetic data (top two rows) than on real data (bottom two rows). This performance gap is mainly induced by the observation noise, which has greater influence on objects with complicated geometric shape (e.g. camera).

4 Runtime Analysis

Given RGB-D images with resolution of 640×480 and mean object count of 4, our implementation approximately runs at 4 FPS on a desktop with an Intel Core i7-5960X CPU (3.0 GHz) and a NVIDIA GTX 1080Ti GPU. Specifically, it takes an average time of 130 ms for instance segmentation, 100 ms for network inference, and 20 ms for pose alignment.

5 Visualization of Shape Priors

In Fig. 2, we visualize the different categorical shape priors used in our ablation studies.

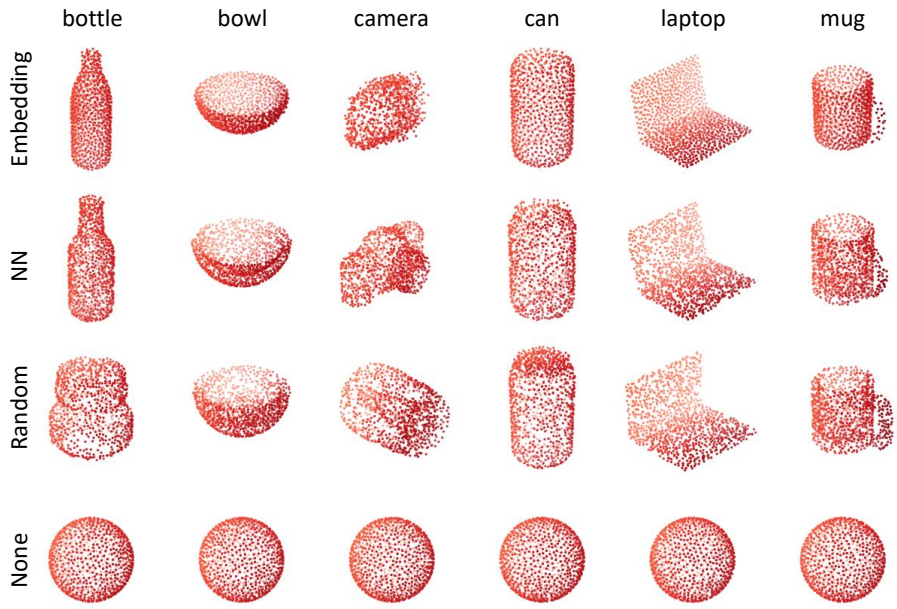

Fig. 2. Categorical shape priors used in ablation studies.

6 Derivation of the Map Operator

We first give the proposition from [2], and then derive the Map operator used in our work as a corollary. Given an object M_c^i, the proper symmetry group $\mathcal{S}(M_c^i)$ is defined as:

$$\mathcal{S}(M_c^i) = \{\mathbf{s} \in SO(3) \mid \forall\, \mathbf{p} \in SO(3),\ \mathcal{I}(M_c^i, \mathbf{p}) = \mathcal{I}(M_c^i, \mathbf{s} \cdot \mathbf{p})\}, \tag{1}$$

where $\mathcal{I}(M_c^i, \mathbf{p})$ is the image of object M_c^i under pose $\mathbf{p}$. Intuitively, $\mathcal{S}(M_c^i)$ consists of rotations which preserve the appearance of a given object.

Proposition 1. *Given a proper symmetry group* $\mathcal{S}(M_c^i)$, $\forall\, R \in SO(3)$, *the* Map *operator is defined as:*

$$\text{Map}(R) = R\hat{S},\ \textit{with}\ \hat{S} = \underset{S \in \mathcal{S}(M_c^i)}{\arg\min} \|RS - I_3\|_F\ , \tag{2}$$

where I_3 *is an* $3{\times}3$ *identity matrix. Then,* $\text{Map}(R_1) = \text{Map}(R_2) \iff \mathcal{I}(M_c^i, R_1) = \mathcal{I}(M_c^i, R_2)$.

The proof is omitted for brevity, refer to [2] for the details. The Map operator used in our work can then be derived directly from Proposition 1.

Corollary 1. *The Map operator for symmetrical objects around the y-axis is given by:*

$$\text{Map}(R) = R\hat{S}, \quad \forall\, R \in SO(3), \tag{3}$$

where

$$\hat{S} = \begin{bmatrix} \cos\hat{\theta} & 0 & -\sin\hat{\theta} \\ 0 & 1 & 0 \\ \sin\hat{\theta} & 0 & \cos\hat{\theta} \end{bmatrix}, \text{ with } \hat{\theta} = \arctan 2(R_{13} - R_{31}, R_{11} + R_{33}). \quad (4)$$

Proof. Assuming the object M_c^i is symmetrical around the y-axis of the object coordinate system, then S has the following form:

$$S = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}. \quad (5)$$

The Froebenius norm can be rewritten as:

$$\begin{aligned} \|RS - I_3\|_F^2 &= 6 - 2\text{Trace}(RS) \\ &= 6 - 2[R_{11}\cos\theta + R_{13}\sin\theta + R_{22} + R_{33}\cos\theta - R_{31}\sin\theta]. \end{aligned} \quad (6)$$

We minimize the Froebenius norm over θ to solve for the Map:

$$\begin{aligned} \hat{\theta} &= \underset{\theta\in[0,2\pi)}{\arg\min} \|RS - I_3\|_F \\ &= \underset{\theta\in[0,2\pi)}{\arg\max} (R_{11} + R_{33})\cos\theta + (R_{13} - R_{31})\sin\theta \\ &= \arctan 2(R_{13} - R_{31}, R_{11} + R_{33}) \,. \end{aligned} \quad (7)$$

Hence,

$$\hat{S} = \begin{bmatrix} \cos\hat{\theta} & 0 & -\sin\hat{\theta} \\ 0 & 1 & 0 \\ \sin\hat{\theta} & 0 & \cos\hat{\theta} \end{bmatrix}, \text{ with } \hat{\theta} = \arctan 2(R_{13} - R_{31}, R_{11} + R_{33}). \quad (8)$$

□

References

1. Chen, D., Li, J., Wang, Z., Xu, K.: Learning canonical shape space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
2. Pitteri, G., Ramamonjisoa, M., Ilic, S., Lepetit, V.: On object symmetries and 6d pose estimation from images. In: 2019 International Conference on 3D Vision (3DV). pp. 614–622. IEEE (2019)
3. Wang, C., Martín-Martín, R., Xu, D., Lv, J., Lu, C., Fei-Fei, L., Savarese, S., Zhu, Y.: 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In: International Conference on Robotics and Automation (ICRA) (2020)
4. 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 Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019)