"CATRE: Iterative Point Clouds Alignment for Category-level Object Pose Refinement" Supplementary Material

Xingyu Liu[1,*] Gu Wang[2,*] Yi Li[3] Xiangyang Ji[1]

[1]Tsinghua University, BNRist [2]JD.com [3]University of Washington
liuxy21@mails.tsinghua.edu.cn, guwang12@gmail.com,
yili.matrix@gmail.com, xyji@tsinghua.edu.cn

Abstract. In this supplementary material, we provide more implementation details and additional results.

A More Implementation Details

A.1 Architecture of Encoder

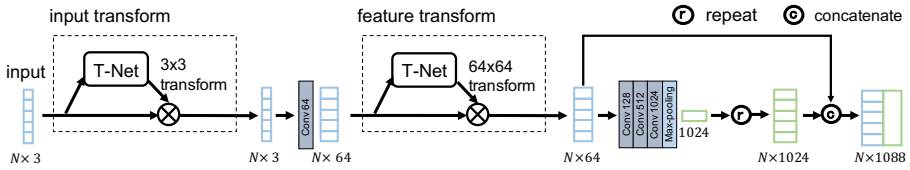


Fig. A.1: Architecture of the PointNet-based [6] geometry encoder, where N denotes the number of input points.

Fig. A.1 depicts the detailed architecture of our employed PointNet-based [6] geometry encoder as mentioned in Sec. 3.3 of the main text. In specific, the encoder first maps each point from $\mathbb{R}^3$ into a high-dimensional space, and generates an additional global feature by aggregating features from all points. The global feature is then repeated N times and concatenated to each point-wise feature, where N denotes the number of input points.

A.2 Instance-level Pose Refinement

As shown in Fig. A.2, since the instance-level refiner (see Sec. 4.4 of the main text) only estimates 6DoF pose transformation, we make some modification

* Equal contribution.

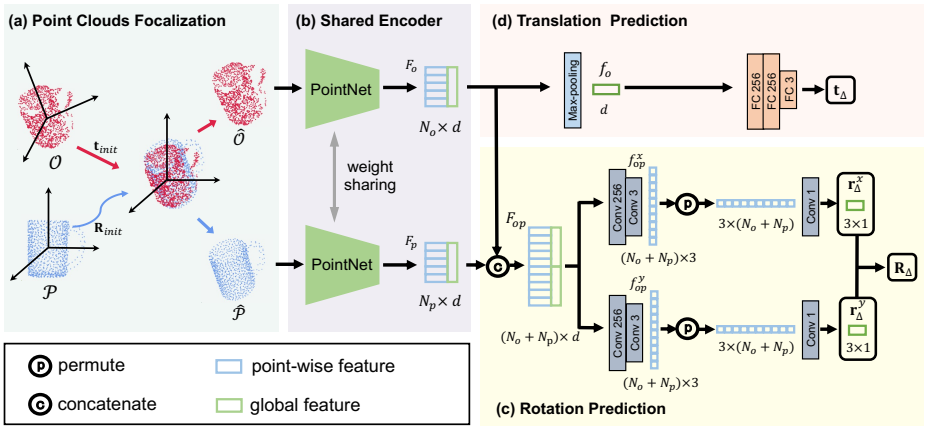


Fig. A.2: Network architecture of the instance-level refiner, notice that $\mathbf{s}_{init}$ and $\mathbf{s}_{\Delta}$ related components are removed compared to the category-level refiner CATRE.

based on our CATRE network. Specifically, the initial size information $\mathbf{s}_{init}$ is removed in both the focalization operation and the original TS-Head, and the output FC layer for relative size prediction is also discarded, making TS-Head into T-Head. Accordingly, the loss function is changed to

$$\mathcal{L} = \mathcal{L}_{pm} + \mathcal{L}_{\mathbf{R}} + \mathcal{L}_{\mathbf{t}}, \tag{1}$$

thereby

$$\begin{cases} \mathcal{L}_{pm} &= \underset{\mathbf{x}\in\mathcal{P}}{\text{avg}} \|(\mathbf{R}_{gt}\mathbf{x} + \mathbf{t}_{gt}) - (\mathbf{R}_{est}\mathbf{x} + \mathbf{t}_{est})\|_1, \\ \mathcal{L}_{\mathbf{R}} &= \frac{3 - \text{Tr}(\mathbf{R}_{gt}\mathbf{R}_{est}^T)}{4}, \\ \mathcal{L}_{\mathbf{t}} &= \|\mathbf{t}_{gt} - \mathbf{t}_{est}\|_1. \end{cases} \tag{2}$$

B Additional Results

B.1 Accuracy on NOCS *vs*. Iterations

Apart from the results in Sec. 4.2 of our paper, in Table B.1, we show the additional results for accuracy w.r.t. iterations, where the initial poses are provided by our reproduced results from NOCS [9]. Since the initial predictions from NOCS are poor, more iterations of refinement improve the results substantially. Considering the trade-off between performance and speed, we empirically set 4 as the maximum number of iteration in all the refine experiments.

B.2 Category-level Refinements on Each Category

Table B.2 evinces category-level pose refinement results for each category. Notably, our method achieves consistent enhancement on 6 categories w.r.t. each

Table B.1: Accuracy w.r.t. iterations. We initialize the predictions with the reproduced results from NOCS [9].

Iteration	0	1	2	3	4
IoU_{75}	9.0	21.8	30.2	39.9	**42.6**
5° 2 cm	7.3	29.0	34.2	39.1	**40.7**
5° 5 cm	9.9	30.7	36.9	44.2	**47.8**
5°	10.8	31.9	38.6	46.1	**49.7**
2 cm	37.8	71.6	74.5	74.2	**74.8**

Table B.2: Detailed results of category-level pose refinement for each category on REAL275. * denote our re-implementation of the method.

Method		NOCS*	w/ Ours	SPD*	w/ Ours	DualPoseNet*	w/ Ours
bottle	IoU_{75}	8.8	25.9	12.2	24.5	12.7	16.4
	5° 2 cm	2.6	42.9	21.6	57.0	27.8	43.1
	5° 5 cm	2.6	49.2	23.2	61.6	33.4	56.7
bowl	IoU_{75}	42.8	88.0	76.1	88.2	80.8	92.6
	5° 2 cm	40.2	84.5	50.5	87.1	72.0	91.4
	5° 5 cm	49.8	85.8	54.0	91.1	78.8	97.4
camera	IoU_{75}	0.8	4.1	3.4	5.9	1.4	4.1
	5° 2 cm	0.0	0.5	0.0	0.8	0.0	0.1
	5° 5 cm	0.0	0.6	0.0	0.9	0.0	0.1
can	IoU_{75}	1.5	14.4	29.6	33.5	5.3	16.2
	5° 2 cm	0.3	62.4	37.9	66.0	32.3	49.1
	5° 5 cm	0.3	70.7	42.7	79.3	47.2	67.4
laptop	IoU_{75}	0.4	78.7	32.7	71.6	66.3	71.3
	5° 2 cm	0.6	39.5	4.6	51.8	36.7	56.8
	5° 5 cm	6.5	65.8	7.0	81.1	48.8	83.8
mug	IoU_{75}	0.0	44.3	8.0	37.6	21.6	65.9
	5° 2 cm	0.0	14.4	0.3	11.8	7.0	23.0
	5° 5 cm	0.0	14.7	0.3	12.6	7.1	23.5
overall	IoU_{75}	9.0	42.6	27.0	43.6	31.4	44.4
	5° 2 cm	7.3	40.7	19.1	45.8	29.3	43.9
	5° 5 cm	9.9	47.8	21.2	54.4	35.9	54.8

metric. However, the refined pose for camera is still poor and the reasons are twofold. First, the initial pose prediction is far from ground truth, making the refinement procedure extremely hard. Moreover, the shape variance among the camera category is huge, so that the categorical mean shape is not certainly reliable.

Since CATRE is meant for real-time applications, it is vital whether the model is able to scale with different sizes and resolutions. The resolution and size have minimal impact on the inference speed because we uniformly sample a fixed number of points (*i.e.*, 1024) from the input point cloud. This strategy is applied vastly [1,7,10,2] and scales well with CATRE. Concretely, there is about

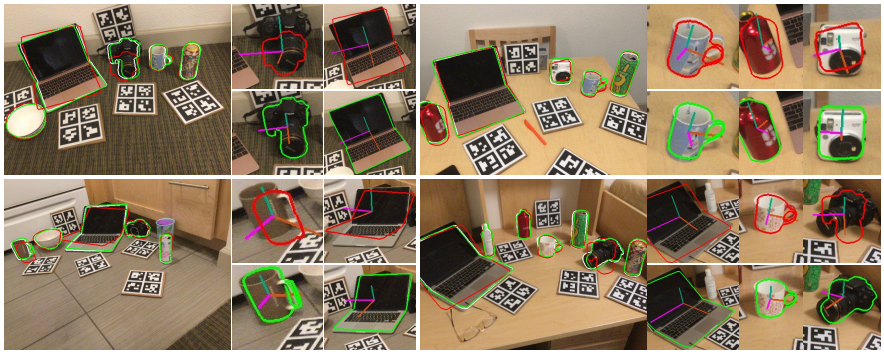


Fig. B.3: Qualitative comparison, where white, red and green contours demonstrate ground-truth, initial (SPD [7]) and refined (Ours) poses, respectively. The axis length reflects predicted scale meanwhile.

ten times the difference between the resolution of "laptop" and "bottle", but CATRE can refine their poses with one model (Table B.2).

B.3 Qualitative Results of Category-level Pose Refinement on REAL275

Fig. B.3 presents the qualitative examples of category-level pose refinement on REAL275.

B.4 More Ablation Studies on REAL275

More ablation studies on category-level pose refinement are presented in Table B.3.

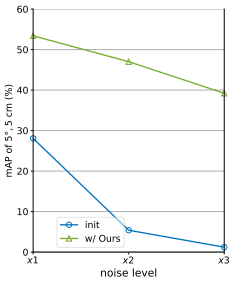
We conduct refinement on the initial poses obtained by perturbing the ground truth with Gaussian noise to explore the sensitivity of CATRE w.r.t. large noise. As shown in Table B.3a, thanks to the iterative testing strategy, even with harsh noise existing in initial poses, CATRE can still achieve satisfying results.

Table B.3b shows some additional results. By removing T-Net in PointNet-based [6] geometry encoder, the accuracy w.r.t. rotation and translation decreases. Since rare occlusion exists in REAL275 dataset, we cropped a 25% block around a random corner of predicted bounding boxes to imitate occlusion. We find that CATRE can exceed the baseline (Table B.3b C1 *vs*. A0), but occlusion still has a side effect on performance (Table B.3b C1 *vs*. B0).

B.5 BOP Results on LM

In the main text, we evaluate our method on LM by traditional protocol following the compared methods [8,3,5]. Recently, BOP metric [4] proves to be more suited

Table B.3: Additional ablation studies on REAL275. *(a)* Accuracy w.r.t. additional noises, where $\times m$ means adding m times of training noise to grund-truth poses during inference. *(b)* Quantitative results, where SPD* denotes our re-implementation of SPD [7].



(a)

Row	Method	IoU_{75}	5°2 cm	5°5 cm	5°	2 cm
A0	SPD*	27.0	19.1	21.2	23.8	68.6
B0	SPD*+Ours	43.6	**45.8**	**54.4**	**58.0**	**75.1**
C0	B0: w/o T-Net in PointNet	**43.7**	41.8	50.0	55.3	74.1
C1	B0: w/ occlusion	36.5	33.4	38.6	43.8	67.3

(b)

Table B.4: Results on LM referring to the Average Recall (%) of BOP metric. * denotes symmetric objects. Ours(B) and Ours(F) use 8 bounding box corners and 128 FPS model points as shape priors respectively.

Method	AR_{MSPD}	AR_{MSSD}	AR_{VSD}	AR
PoseCNN [11]	86.1	73.5	59.3	73.0
w/ Ours(B)	74.7	88.4	91.4	84.8
w/ Ours(F)	84.2	95.1	96.4	91.9

for the evaluation of pose estimation, specifically for symmetric objects. It report an average recall (AR) by calculating the mean score of three metrics: $AR = (AR_{MSPD} + AR_{MSSD} + AR_{VSD})/3$. Details of these metrics can be referred to [4].

As shown in Table B.4, our method still achieve distinct enhancement w.r.t. AR against the initial prediction provided by PoseCNN [11].

B.6 Failure Cases Analyses

As shown in Fig. B.4, failure conditions can be generally grouped into 4 categories: (a) severe occlusion or truncation, (b) poor initial prediction, (c) inaccurate ground-truth label, and (d) ambiguity of symmetry (especially for mug and camera).

References

1. Chen, W., Jia, X., Chang, H.J., Duan, J., Leonardis, A.: G2L-Net: Global to local network for real-time 6D pose estimation with embedding vector features. In:

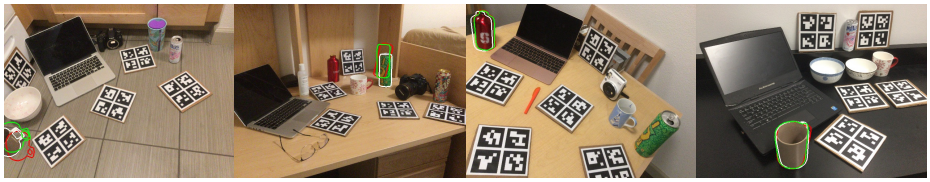


Fig. B.4: Several failure cases, where white, red and green contours demonstrate ground-truth, initial (SPD [7]) and refined (Ours) poses, respectively.

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4233–4242 (2020)
2. Chen, W., Jia, X., Chang, H.J., Duan, J., Linlin, S., Leonardis, A.: FS-Net: Fast shape-based network for category-level 6D object pose estimation with decoupled rotation mechanism. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1581–1590 (June 2021)
3. Gao, G., Lauri, M., Hu, X., Zhang, J., Frintrop, S.: CloudAAE: Learning 6D object pose regression with on-line data synthesis on point clouds. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 11081–11087 (2021)
4. Hodan, T., Sundermeyer, M., Drost, B., Labbe, Y., Brachmann, E., Michel, F., Rother, C., Matas, J.: BOP Challenge 2020 on 6D object localization. In: Bartoli, A., Fusiello, A. (eds.) European Conference on Computer Vision Workshops (ECCVW). pp. 577–594 (2020)
5. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1521–1529 (2017)
6. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) **1**(2), 4 (2017)
7. Tian, M., Ang Jr, M.H., Lee, G.H.: Shape prior deformation for categorical 6D object pose and size estimation. In: European Conference on Computer Vision (ECCV) (August 2020)
8. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: DenseFusion: 6D object pose estimation by iterative dense fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3343–3352 (2019)
9. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2642–2651 (2019)
10. Weng, Y., Wang, H., Zhou, Q., Qin, Y., Duan, Y., Fan, Q., Chen, B., Su, H., Guibas, L.J.: CAPTRA: Category-level pose tracking for rigid and articulated objects from point clouds. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13209–13218 (2021)
11. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. Robotics: Science and Systems Conference (RSS) (2018)