Supplementary Material of LaPose: Laplacian Mixture Shape Modeling for RGB-Based Category-Level Object Pose Estimation

Ruida Zhang[1], Ziqin Huang[1], Gu Wang[1], Chenyangguang Zhang[1], Yan Di[2], Xingxing Zuo[3], Jiwen Tang[1], and Xiangyang Ji[1]

[1] Tsinghua University
[2] Technical University of Munich
[3] California Institute of Technology
{zhangrd23@mails. xyji@}tsinghua.edu.cn

A Details of Loss Terms

We adopt L1 distance to supervise the learning of the scale-agnostic pose parameters,

$$\begin{aligned} \mathcal{L}_R &= ||\mathbf{R}_{out} - \mathbf{R}_{out}^{(gt)}||_1, \\ \mathcal{L}_t &= ||\mathbf{t}_{out} - \mathbf{t}_{out}^{(gt)}||_1, \\ \mathcal{L}_s &= ||\mathbf{s}_{out} - \mathbf{s}_{out}^{(gt)}||_1, \end{aligned} \tag{1}$$

where $\mathbf{R}_{out}^{(gt)}, \mathbf{t}_{out}^{(gt)}, \mathbf{s}_{out}^{(gt)}$ is the ground truth value of $\mathbf{R}_{out}, \mathbf{t}_{out}, \mathbf{s}_{out}$ respectively.

Additionally, we adopt point matching loss [7] to supervise rotation.

$$\mathcal{L}_{pm} = \text{avg}_{\mathbf{x}\in\mathbf{M}}||\mathbf{R}_{pred}\mathbf{x} - \mathbf{R}_{gt}\mathbf{x}||_1, \tag{2}$$

where $\mathbf{x}$ is a point on the object model $\mathbf{M}$, $\mathbf{R}_{pred}$ is the predicted rotation matrix and $\mathbf{R}_{gt}$ is the ground truth one.

We define $\mathcal{L}_{pose}$ in Eq. 9 in the main text as follows,

$$\mathcal{L}_{pose} = \lambda_R\mathcal{L}_R + \lambda_t\mathcal{L}_t + \lambda_s\mathcal{L}_s + \lambda_{pm}\mathcal{L}_{pm}, \tag{3}$$

where $\{\lambda_R, \lambda_t, \lambda_s, \lambda_{pm}\} = \{1, 1, 1, 1\}$ are the balancing hyper-parameters.

B Errors in Evaluation Script

We have identified two main errors in the evaluation script provided by [8].

Firstly, the computation of 3D IoU is erroneous, which is also identified in [5]. Further details on this issue can be found in this link.

Secondly, when calculating the mean Average Precision under n° m *cm* metric, the script employs a 3D IoU threshold of 10% to filter out negative predictions (see this link). However, it mistakenly filters the ground truth as well, as indicated in Lines 1837 - 1840 (here). The filtered ground truth is used to compute

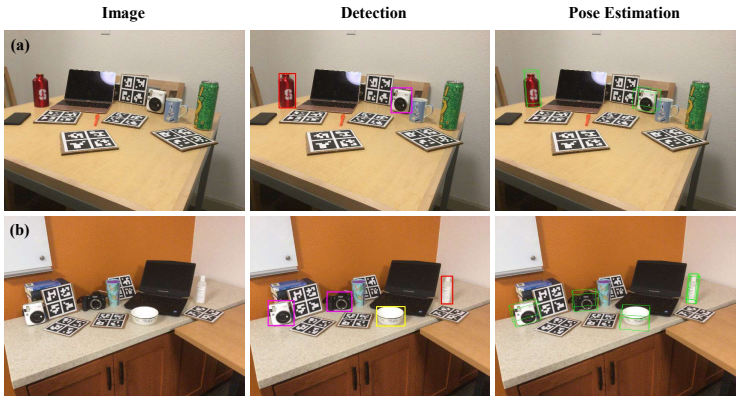

Fig. D.1: Failure cases caused by missing detections.

recall for mAP (see this link), leading to falsely high recall values. This error has been rectified in our code.

After correction, the IoU mAP decreases for all methods, while $n°m\,cm$ of DMSR and MSOS increases and $n°m\,cm$ of OLD-Net decreases. The method ranking remains roughly the same.

C Predicting Metric Scale

In order to recover the object metric scale, we employ a MobileNet [3] as the scale net to predict the diagonal length d of the object tight bounding box. We compute the average metric scale d_{avg} of the category beforehand and predict the delta value $d_{out} = d - d_{avg}$. We use L1 loss to supervise the learning of the scale net.

D Limitations

- Our pose estimation method relies on the accuracy of object detection results. As illustrated in Fig. D.1, our method encounters failures when the detector fails to accurately detect the target object. Employing more advanced foundation models for object detection could potentially improve the precision of pose estimation.
- While LaPose exhibits superior performance against all competitors under absolute-scale metrics, it does not completely resolve the issue of scale ambiguity. A notable example of this challenge can be observed by comparing results in Tab. E.2 and Tab. E.1, where the absolute-scale metric of *bottle* experiences a significant drop due to the inherent variability in the scale of *bottle* instances, making the scale prediction particularly challenging. In practical applications such as robotic manipulation, absolute-scale poses are often

Table E.1: Per-category results of LaPose (Ours) on NOCS-REAL275 using scale-agnostic evaluation metrics.

Category	$NIoU_{25}$	$NIoU_{50}$	$NIoU_{75}$	$10°0.2d$	$10°0.5d$	$0.2d$	$0.5d$	$10°$
bottle	41.1	16.3	1.2	16.7	45.5	18.6	50.4	51.1
bowl	100.0	96.9	46.9	89.6	97.4	92.3	100.0	97.4
camera	43.3	6.5	0.0	0.7	14.2	5.8	54.4	21.3
can	60.3	23.0	2.4	27.6	84.0	28.0	85.6	90.5
laptop	80.3	71.6	27.0	52.4	60.0	63.1	83.6	60.6
mug	96.4	73.1	17.0	37.6	43.1	73.4	98.9	43.2

Table E.2: Per-category results of LaPose (Ours) on NOCS-REAL275 using absolute evaluation metrics.

Category	IoU_{25}	IoU_{50}	IoU_{75}	$5°5cm$	$10°5cm$	$10°10cm$
bottle	0.0	0.0	0.0	0.0	0.0	0.3
bowl	47.8	24.8	8.9	24.8	27.8	52.6
camera	44.6	12.0	0.1	0.3	5.4	15.3
can	33.2	11.7	1.6	7.7	26.9	63.9
laptop	84.1	41.2	4.2	2.3	6.1	30.8
mug	37.7	15.4	0.9	2.8	9.1	20.3

required for precise actions. To address this challenge, additional sources of information such as extra viewpoints [1] and active perception methodologies [6] could be incorporated into the framework.

E Per-Category Results on NOCS Datasets

We present per-category results of LaPose (Ours) in Tab. E.1, Tab. E.2, Tab. E.3 and Tab. E.4.

F Additional Ablation Studies

In Tab. F.5, we respectively remove the conv-branch, DINO-branch and SAP from the full version (Tab. 4 (G) in the main text). Removing any component results in decreased performance. This demonstrates the effectiveness of our design.

G Additional Qualitative Results

In Fig. G.2 and Fig. G.3, we provide additional qualitative results on NOCS-Real275 and CAMERA25.

Table E.3: Per-category results of LaPose (Ours) on NOCS-CAMERA25 using scale-agnostic evaluation metrics.

Category	$NIoU_{25}$	$NIoU_{50}$	$NIoU_{75}$	$10°0.2d$	$10°0.5d$	$0.2d$	$0.5d$	$10°$
bottle	75.3	49.4	14.1	52.5	81.5	52.6	82.2	86.3
bowl	94.0	77.4	22.9	72.8	94.7	73.2	95.2	95.9
camera	65.1	29.3	2.9	18.8	56.0	25.7	77.1	66.0
can	74.3	42.6	8.5	31.9	76.9	31.9	77.0	85.6
laptop	90.3	71.3	31.5	63.0	85.2	67.7	93.7	88.5
mug	52.3	23.5	4.6	15.2	44.5	21.1	62.2	57.6

Table E.4: Per-category results of LaPose (Ours) on NOCS-CAMERA25 using absolute evaluation metrics.

Category	IoU_{25}	IoU_{50}	IoU_{75}	$5°5cm$	$10°5cm$	$10°10cm$
bottle	37.3	11.8	1.7	6.2	7.5	27.6
bowl	30.7	8.6	1.1	10.7	11.9	43.4
camera	16.9	4.3	0.3	1.9	6.9	25.1
can	16.1	4.7	0.7	9.3	13.0	44.6
laptop	70.7	31.3	5.2	4.5	9.7	31.2
mug	27.7	7.5	0.6	4.7	9.3	32.5

Table F.5: Additional ablation studies on NOCS-REAL275.

Method	$NIoU_{25}$	$NIoU_{50}$	$10°0.2d$	$10°0.5d$
w/o conv	65.4	37.4	21.1	43.3
w/o DINO	65.5	43.4	33.6	51.2
w/o SAP	52.4	25.8	15.4	41.6
Ours	**70.7**	**47.9**	**37.4**	**57.4**

References

1. Chen, K., James, S., Sui, C., Liu, Y.H., Abbeel, P., Dou, Q.: Stereopose: Category-level 6d transparent object pose estimation from stereo images via back-view nocs. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2855–2861. IEEE (2023) 3
2. Fan, Z., Song, Z., Xu, J., Wang, Z., Wu, K., Liu, H., He, J.: Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In: ECCV. pp. 220–236. Springer (2022) 6, 7
3. Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q.: Searching for mobilenetv3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1314–1324 (2019) 2
4. Lee, T., Lee, B.U., Kim, M., Kweon, I.S.: Category-level metric scale object shape and pose estimation. IEEE RA-L **6**(4), 8575–8582 (2021) 6, 7

5. Liu, X., Wang, G., Li, Y., Ji, X.: Catre: Iterative point clouds alignment for category-level object pose refinement. In: European Conference on Computer Vision. pp. 499–516. Springer (2022) 1
6. Ren, X., Luo, J., Solowjow, E., Ojea, J.A., Gupta, A., Tamar, A., Abbeel, P.: Domain randomization for active pose estimation. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7228–7234. IEEE (2019) 3
7. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: CVPR (June 2021) 1
8. 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: CVPR. pp. 2642–2651 (2019) 1
9. Wei, J., Song, X., Liu, W., Kneip, L., Li, H., Ji, P.: Rgb-based category-level object pose estimation via decoupled metric scale recovery. arXiv preprint arXiv:2309.10255 (2023) 6, 7

Fig. G.2: Qualitative results of Ours (green line), DMSR [9] (blue), OLD-Net [2] (red) and MSOS [4] (pink) on **NOCS-REAL275**.

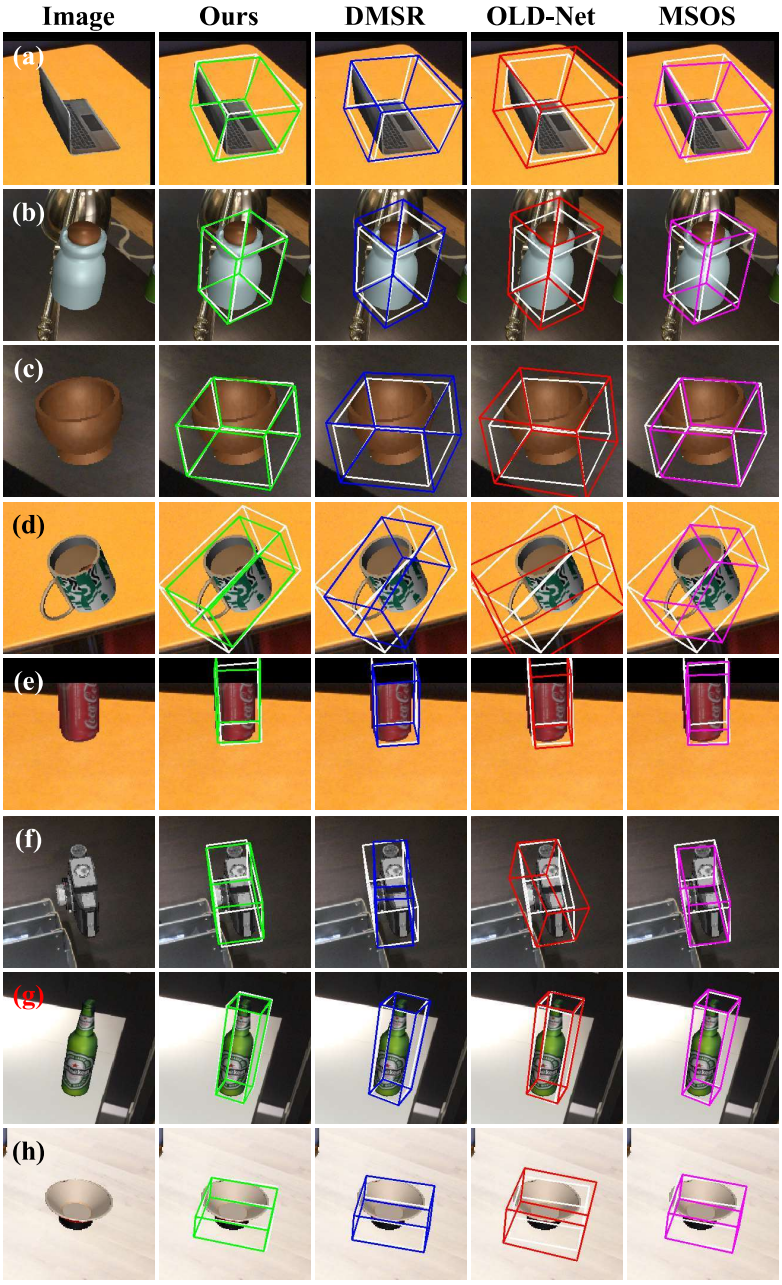

Fig. G.3: Qualitative results of Ours (green line), DMSR [9] (blue), OLD-Net [2] (red) and MSOS [4] (pink) on **NOCS-CAMERA25**.