CleanPose: Category-Level Object Pose Estimation via Causal Learning and Knowledge Distillation

Supplementary Material

A. More Loss Function Details

The backbone of our method is based on AG-Pose [6]. In addition to $\mathcal{L}_{pose}$ (Eq. (21)), there are some additional loss functions to balance keypoints selection and pose prediction. First, to encourage the keypoints to focus on different parts, the diversity loss $\mathcal{L}_{div}$ is used to force the detected keypoints to be away from each other, in detail:

$$\mathcal{L}_{div} = \sum_{i=1}^{N_{kpt}} \sum_{j=1, j\neq i}^{N_{kpt}} \mathbf{d}\left(\mathcal{P}_{kpt}^{(i)}, \mathcal{P}_{kpt}^{(j)}\right) \tag{1}$$

$$\mathbf{d}\left(\mathcal{P}_{kpt}^{(i)}, \mathcal{P}_{kpt}^{(j)}\right) = \max\left\{th_1 - \left\|\mathcal{P}_{kpt}^{(i)} - \mathcal{P}_{kpt}^{(j)}\right\|_2, 0\right\}, \tag{2}$$

where th_1 is a hyper-parameter and is set as 0.01, $\mathcal{P}_{kpt}^{(i)}$ means the i-th keypoint. To encourage the keypoints to locate on the surface of the object and exclude outliers simultaneously, an object-aware chamfer distance loss $\mathcal{L}_{ocd}$ is employed to constrain the distribution of $\mathcal{P}_{kpt}$. In formula:

$$\mathcal{L}_{ocd} = \frac{1}{|\mathcal{P}_{kpt}|} \sum_{x_i \in \mathcal{P}_{kpt}} \min_{y_j \in \mathcal{P}_{obj}^{'}} \|x_i - y_j\|_2, \tag{3}$$

where $\mathcal{P}_{obj}^{'}$ denotes the point cloud of objects without outlier points. Moreover, we also use MLP to predict the NOCS coordinates of keypoints $\mathcal{P}_{kpt}^{nocs} \in \mathbb{R}^{N_{kpt}\times 3}$. Then, we generate ground truth NOCS of keypoints $\mathcal{P}_{kpt}^{gt}$ by projecting their coordinates under camera space $\mathcal{P}_{kpt}$ into NOCS using the ground truth $\mathcal{R}_{gt}, t_{gt}, s_{gt}$. And we use the $SmoothL_1$ loss to supervise the NOCS projection:

$$\mathcal{P}_{kpt}^{gt} = \frac{1}{\|s_{gt}\|_2} \mathcal{R}_{gt}\left(\mathcal{P}_{kpt} - t_{gt}\right) \tag{4}$$

$$\mathcal{L}_{nocs} = SmoothL_1(\mathcal{P}_{kpt}^{gt}, \mathcal{P}_{kpt}^{nocs}). \tag{5}$$

Hence, the complete form of overall loss (Eq. (22)) is as follows:

$$\begin{aligned}\mathcal{L}_{all} = \lambda_1\mathcal{L}_{ocd} + \lambda_2\mathcal{L}_{div} + \lambda_3\mathcal{L}_{nocs} \\ +\lambda_4\mathcal{L}_{pose} + \alpha_2\mathcal{L}_{KD},\end{aligned} \tag{6}$$

where the parameters are set as $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \alpha_2) = (1.0, 5.0, 1.0, 0.3, 0.01)$ according to AG-Pose [6] and following ablations.

B. More Details of Using ULIP-2

ULIP-2 [12] is a large-scale 3D foundation model with strong perception capabilities for the point cloud modality. It offers multiple pre-trained versions of point cloud encoders. In our model, there are two key steps that involve the use of different pre-trained encoders of ULIP-2. **(i)** In the knowledge distillation process, we leverage the pre-trained PointBert [13], which achieves the best zero-shot classification performance across all versions. Therefore, it can provide comprehensive category knowledge guidance for our model. In the ablation study Tab. 5d, we also compared it with PointNet++ [10], which is more similar in architecture to our model. The experimental results demonstrate that our distillation method focuses more on category knowledge rather than feature similarity. **(ii)** However, during the initial construction of the dynamic queue, we use the pre-trained PointNet++ [10], as the front-door adjustment primarily focuses on the differences between samples. We aim to avoid introducing confounders due to feature discrepancies from different encoders. The additional ablation study results in Tab. S2b also support our analysis.

C. Additional Ablations

Effect of varying queue lengths N_q. Tab. S1a ablates the different lengths of dynamic queue N_q. The queue that is too short results in insufficient sample diversity, while too long affect memory efficiency and feature consistency. We observe that the estimation performance achieves the peak at the length of around 80, with slight declines upon further increases. We speculate that the queue length is closely related to task characteristics and data scale of COPE. We select $N_q = 80$ in our model to balance between efficiency and accuracy.

Effect of different sampling quantities N_s. In Sec. 4.2 of main manuscript, we sample N_s features for the specific network design to perform *front-door adjustment*. We study the influence with response to different sampling quantities N_s in Tab. S1b. It can be found that a large N_s leads to a slight performance degradation. We speculate that a larger sample size may introduce noise and redundant information that affects key features in causal inference. An appropriate sample size can balance valid and redundant information, prompting the model to focus on learning more representative causal correlations. The results demonstrate that $N_s = 12$ yields the most significant performance gains.

Varying balanced coefficient α_2 for loss $\mathcal{L}_{KD}$. In Sec. 4.3 of main manuscript, we introduce L2 loss to supervise the feature-based distillation and use α_2 to balanced its contribution in overall loss function. We investigate the impact of different α_2 in Tab. S1c. We observe that the better performance is achieved when α_2 is small, possibly because

N_q	5°2cm	5°5cm	10°2cm	10°5cm
20	57.0	64.6	75.1	84.7
50	60.8	67.3	77.9	**86.4**
80	**61.7**	**67.6**	**78.3**	86.3
200	59.4	66.1	78.0	85.9
500	58.8	65.3	76.8	85.8
1000	58.3	66.8	76.3	86.2
3000	57.7	65.5	75.6	85.0
10000	57.0	65.0	75.7	85.4

(a) Effect of varying queue lengths N_q

N_s	5°2cm	5°5cm	10°2cm	10°5cm
6	60.7	66.3	77.8	85.8
12	**61.7**	**67.6**	78.3	86.3
18	59.4	66.5	**78.8**	**86.8**
24	58.5	65.8	77.8	85.9
48	56.8	64.9	76.3	85.6
80	56.9	64.4	76.2	85.8

(b) Effect of varying queue lengths N_s

α_2	5°2cm	5°5cm	10°2cm	10°5cm
0.005	59.3	66.8	78.0	**86.4**
0.01	**61.7**	**67.6**	**78.3**	86.3
0.1	58.4	65.1	77.6	85.9
0.5	57.3	63.9	76.2	86.0
1	56.9	63.4	76.4	85.4

(c) Effect of varying balanced coefficient α_2

Table S1. Additional ablation studies on some hyper-parameters. Default settings are colored in gray.

$\mathcal{L}_{KD}$ is comparable in magnitude to the pose loss function, which is favorable for regression. The results show that our method performs well under $\alpha_2 = 0.01$.

Different queue initialization approaches. By default, we construct confounders queue with features extracted by 3D encoders of ULIP-2 [12]. Alternatively, we can randomly initialize the queue, which should achieve the same effect ideally. Therefore, we evaluate the performance between two initialization approaches in Tab. S2a. The results indicate the degraded performance with "Random" initialization strategy. We speculate that the randomly initialized queue may introduce additional and uncontrollable confounders, limiting the model's optimization potential.

Effect of different 3D encoder for initial construction of the queue. Tab. S2b ablates the different 3D encoders of ULIP-2 [12] for initial construction of the dynamic queue. The results exhibit that using PointNet++ [10] yields the most performance gains. As mentioned in Sec. B, the dynamic queue is utilized in the cross-attention phase of front-door adjustment, thus primarily focusing on the differences between samples. Using encoders with similar architectures helps avoid introducing extra confounders.

Init.	5°2cm	5°5cm	10°2cm	10°5cm
Random	58.0	65.7	75.8	85.1
Extract	**61.7**	**67.6**	**78.3**	**86.3**

(a) Effect of different queue initialization approaches

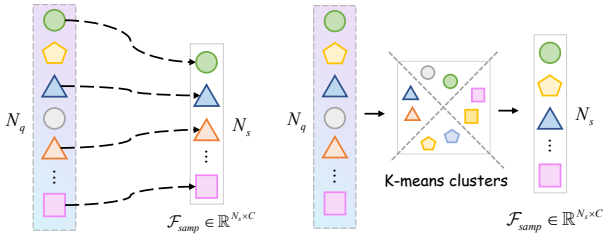
3D Encoder	5°2cm	5°5cm	10°2cm	10°5cm
PointNet++[10]	**61.7**	**67.6**	**78.3**	86.3
PointMLP[9]	58.1	65.8	76.3	85.2
PointBert[13]	58.8	65.6	77.9	**86.4**

(b) Effect of different 3D encoders of ULIP-2 [12] for initial construction of the queue.

Selector	5°2cm	5°5cm	10°2cm	10°5cm
Random	**61.7**	**67.6**	**78.3**	86.3
K-means	58.5	65.5	78.0	**86.5**
K-means (simi)	58.7	66.2	77.7	86.0

(c) Effect of distinct feature selectors

Table S2. Additional ablation studies on confounders queue. Default settings are colored in gray.



(a) Random feature selector (b) K-means feature selector

Figure S1. Illustration of different feature selectors in ablations Tab. S2c.

Various feature selection strategies. In Sec. 4.2 of main manuscript, we randomly sample N_s features from queue by default, as shown in Fig. S1(a). Optionally, we can first use K-means to cluster the features of the queue, and then select features from each cluster to form $\mathcal{F}_{samp}$, as illustrated in Fig. S1(b). For fair comparison, the number of clusters is set equal to N_s. We investigate the impact of these two feature selector in Tab. S2c. As shown in the table, we observe that K-means-based feature sampling strategy shows a decline in performance on strict metrics (5°2cm and 5°5cm). We argue that k-means, which clusters by Euclidean distance, may lose important boundary information, thus affecting the model performance. Moreover, k-means clustering needs to be performed after each queue update, which increases computational load and training costs. Therefore, we added a comparative experiment using similarity-based updates, denoted as 'K-means (simi)', where (simi) refers to the 'Similarity' defined in Tab. 5a of main text. In this case, clustering is only performed once

Category	IoU^*_{25}↑	IoU^*_{50}↑	IoU^*_{75}↑	5°2*cm*↑	5°5*cm*↑	10°2*cm*↑	10°5*cm*↑
bottle	51.3	49.4	37.1	75.7	81.7	79.9	87.8
bowl	100.0	100.0	93.8	93.3	98.2	95.0	99.9
camera	90.9	83.5	39.9	2.8	3.2	33.9	40.5
can	71.3	71.1	43.2	84.2	85.9	96.9	98.6
laptop	86.3	84.0	76.1	69.2	90.9	71.9	98.5
mug	99.6	99.4	86.1	45.2	45.6	91.9	91.9
Average	83.3	81.2	62.7	61.7	67.6	78.3	86.3

Table S3. **Category-wise evaluation of CleanPose on REAL275 dataset.** '*' denotes CATRE [8] IoU metrics.

Methods	IoU_{75}↑	IoU_{25}/IoU_{50}↑										
		Average	Bottle	Box	Can	Cup	Remote	Teapot	Cutlery	Glass	Tube	Shoe
NOCS[11]	-	50.0/21.2	41.9/5.0	43.3/6.5	81.9/62.4	68.8/2.0	**81.8/59.8**	24.3/0.1	14.7/6.0	95.4/49.6	21.0/4.6	26.4/16.5
FS-Net[1]	14.8	74.9/48.0	65.3/45.0	31.7/1.2	98.3/73.8	96.4/68.1	65.6/46.8	69.9/59.8	71.0/51.6	99.4/32.4	79.7/46.0	71.4/55.4
GPV-Pose[3]	15.2	74.9/50.7	66.8/45.6	31.4/1.1	98.6/75.2	96.7/69.0	65.7/46.9	75.4/61.6	70.9/52.0	99.6/62.7	76.9/42.4	67.4/50.2
VI-Net[5]	20.4	80.7/56.4	90.6/79.6	44.8/12.7	99.0/67.0	96.7/72.1	54.9/17.1	52.6/47.3	89.2/76.4	99.1/93.7	**94.9**/36.0	85.2/62.4
SecondPose[2]	24.9	83.7/66.1	94.5/79.8	**54.5**/23.7	98.5/93.2	99.8/82.9	53.6/35.4	81.0/71.0	93.5/74.4	99.3/92.5	75.6/35.6	86.9/73.0
AG-Pose[6]	53.0	88.1/76.9	97.6/**86.0**	54.0/13.9	98.3/96.7	**100/99.9**	53.9/37.2	**99.9/98.5**	96.0/**93.3**	**100**/99.3	81.4/45.0	99.7/99.5
CleanPose	**53.9**	**89.2/79.8**	**99.9**/79.1	51.4/**28.7**	**99.9/99.7**	**100/99.9**	71.2/57.8	99.0/94.0	**97.8**/91.0	**100/99.6**	72.7/**48.4**	**99.8/99.8**

Table S4. **Overall and category-wise evaluation of 3D IoU on the HouseCat6D.** ↑: a higher value indicating better performance, '-' means unavailable statistics. Overall best results are in **bold** and the second best results are underlined.

Seed	1	42	500	1k	1w	10w	σ^2 ↓
5°2*cm*	61.4	61.5	61.7	61.4	61.3	61.7	0.03
5°5*cm*	67.2	67.3	67.5	67.2	67.1	67.6	0.04

Table S5. Effect of different sampling seed during inference.

during the initial training. However, experimental results also show that such strategy leads to further performance degradation as one clustering loses the diversity of features.

Effect of different sampling seed during inference. We have conducted additional experiments with 6 different random seeds, as shown in Tab. S5. The computed variances σ^2 for metrics demonstrate stable performance across different random seeds, indicating the robustness and reliability of our method.

D. More Experimental Results

We report category-wise results of REAL275 [11] in Tab. S3. Since there is a small mistake in the original evaluation code of NOCS [11] for the 3D IoU metrics, we present more reasonable CATRE [8] metrics following [2, 7, 14]. Further, more detailed results of HouseCat6D [4] are shown in Tab. S4. As for more restricted metric IoU_{75}, our method also demonstrates state-of-the-art performance (**53.9%**), further validating the effectiveness of CleanPose in 3D IoU evaluation. Moreover, in category-wise validation on IoU_{25} and IoU_{50}, our approach obtains state-of-the-art (*e.g.*, *Can*, *Cup*, *Glass* and *Shoe*) or competitive results across all categories. It is worth mentioning that our method exhibits more stable performance on these two metrics. For instance, compared to the current sota method AG-Pose [6] in the *Box* category, our method achieves the best performance (**28.7%**) on IoU_{50} metric when both obtain competitive results on IoU_{25} metric, with a significant reduction of the AG-Pose (13.9%).

ID	Method	Visual Enc.	Param.↓	IoU^*_{75} ↑	5°2*cm*↑	TT↓	IT↑
1	AG-Pose	ViT-S/14	**223M**	61.3	57.0	**51.3**	**35**
2	ours	ViT-S/14	246M	**62.7**	**61.7**	51.8	33
3	ours†	ViT-S/14	246M	61.8	58.1	51.6	33
4	AG-Pose	ResNet18	**220M**	60.9	56.2	**51.2**	**35**
5	ours	ResNet18	243M	62.3	60.3	51.6	33

Table S6. Detailed comparison results. '*' denotes CATRE [8] IoU metrics. '†' represents replacement of causal module with MLPs of the same number of parameters. TT: Traning Time (min/epoch), IT: Inference Speed (Frame/sec). Overall best results are in **bold** and the second best results are underlined. Default settings are colored in gray.

E. Detailed Comparison with SOTA method

We follow the domain consensus to report the metric accuracy in main manuscript. Moreover, we add more terms, *e.g.* visual encoder type, inference latency (FPS), in Tab. S6 for comprehensive comparison. As confirmed in (#1) and (#2) of Tab. S6, the front-door adjustment only increases the

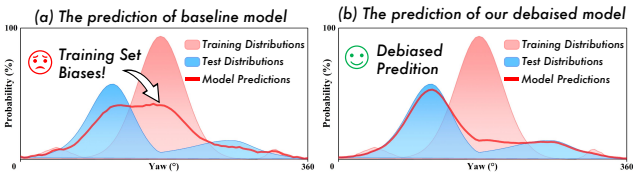


Figure S2. Effect of debiasing. Illustrated by Yaw angle distributions of the *mug* category.

number of parameters by **10%** (246M *vs*. 223M), while the running time remains nearly unchanged (33 *vs*. 35 in FPS). To ensure a fair comparison, we also replace the front-door module with MLPs that have the same number of parameters (#3). The results further demonstrate the superior effectiveness of causal learning. What's more, we include additional results with ResNet18. Specifically, our method still outperforms AG-Pose with ResNet18 setting (#4 *vs*. #5), further supporting the efficacy of our approach.

F. Effect of Debiasing.

We evaluate the debiasing effect via rotation distributions. As shown in Fig. S2, the predictions of baseline model are clearly biased toward the training set distributions, while the debiased model primarily unaffected.

G. Limitation and Broader Impact

Limitation and future work. While our method achieves superior results in various challenging benchmarks of category-level pose estimation, there are still several aspects for improvement. First, although the front-door adjustment is effective, the investigate on the application of causal learning methods remains incomplete. Therefore, exploring further use of different causal learning methods such as back-door adjustment and counterfactual reasoning may enhance the performance of CleanPose. Second, despite the guidance of the causal analysis, the network modules in actual implementation may induce inaccuracy inevitably. Such a flaw introduces a gap between the causal framework and the network design. In future work, we will further study advanced algorithm design strategies.

Broader Impact. For tasks with parameter regression properties, *e.g*., category-level pose estimation, the current mainstream approaches focus on exploring advanced network designs to perform data fitting. We believe that relying solely on learning statistical similarity can also introduce spurious correlations into parameter regression models, thereby damaging the model's generalization ability. We hope this work brings new insights for the broader and long-term research on parameter regression tasks. Besides, adapting foundation models to downstream tasks has become a dominant paradigm in machine learning. Our method also provide novel views for offering knowledge guidance in similar tasks across diverse categories.

References

[1] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1581–1590, 2021. 3

[2] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9959–9969, 2024. 3

[3] 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 (CVPR)*, pages 6781–6791, 2022. 3

[4] HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22498–22508, 2024. 3

[5] Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 14001–14011, 2023. 3

[6] Xiao Lin, Wenfei Yang, Yuan Gao, and Tianzhu Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21040–21049, 2024. 1, 3

[7] Jian Liu, Wei Sun, Chongpei Liu, Hui Yang, Xing Zhang, and Ajmal Mian. Mh6d: Multi-hypothesis consistency learning for category-level 6-d object pose estimation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2024. 3

[8] Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. Catre: Iterative point clouds alignment for category-level object pose refinement. In *European Conference on Computer Vision (ECCV)*, pages 499–516. Springer, 2022. 3

[9] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[10] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *In Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 1, 2

[11] 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 (CVPR)*, pages 2642–2651, 2019. 3

[12] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27091–27101, 2024. 1, 2

[13] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19313–19322, 2022. 1, 2

[14] Linfang Zheng, Tze Ho Elden Tse, Chen Wang, Yinghan Sun, Hua Chen, Ales Leonardis, Wei Zhang, and Hyung Jin Chang. Georef: Geometric alignment across shape variation for category-level object pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10693–10703, 2024. 3