COG: Confidence-aware Optimal Geometric Correspondence for Unsupervised Single-reference Novel Object Pose Estimation

Supplementary Material

Contents

A. Model Details

A.1. Sinkhorn and Affinity Kernel

To compute differentiable cross-view correspondences under non-uniform confidence marginals, we adopt the entropy-regularized optimal transport (OT) framework with a log-domain Sinkhorn [7] solver. Let $\mathbf{K} \in \mathbb{R}^{n\times n}$ denote the affinity matrix, and let $\mathbf{w}_{\text{row}} = \mathbf{w}_p, \mathbf{w}_{\text{col}} = \mathbf{w}_q$ be the target marginals obtained by normalizing point-wise confidences from the query and reference point clouds. The Sinkhorn algorithm iteratively updates the dual variables $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ and recovers a transport plan satisfying the prescribed marginals (Alg. 1). The log-sum-exp implementation ensures numerical stability and prevents underflow in high-dimensional spaces.

For explanatory decomposition, we first define the OT cost as the negative similarity, and decompose the OT cost into a geometric similarity term and a nonlinear scaled semantic consistency term:

$$\mathbf{C}_{[i,j]} = -\langle \mathbf{G}_{p[i]}, \mathbf{G}_{q[j]} \rangle_{\cos} - \lambda \log\big(1 + \langle \mathbf{S}_{p[i]}, \mathbf{S}_{q[j]} \rangle_{\cos}\big), \tag{1}$$

where the semantic term is expressed in a logarithmic form. This design yields a semantic modulation that encourages semantically coherent matches without overwhelming the geometric similarity when semantic features are relatively similar. On the other hand, when the semantic similarity becomes negative, it yields a strong penalty for semantically inconsistent matches. This nonlinear shape prevents excessive amplification for already similar semantic pairs, while still sharply suppressing mismatched or unrelated regions. In practice, since cosine similarity may take negative values close to -1 and $\log(1+x)$ close to $-\infty$, a small stability constant $\epsilon = 10^{-6}$ is added inside the logarithmic semantic term for stable training. And the regularized OT problem is written as:

$$\min_{\mathbf{\Pi}\geq 0} \langle \mathbf{C}, \mathbf{\Pi} \rangle \; + \; \tau \sum_{i,j} \mathbf{\Pi}_{[i,j]} \log \mathbf{\Pi}_{[i,j]}, \tag{2}$$

subject to the constraints $\mathbf{\Pi}\,\mathbf{1} = \mathbf{w}_{\text{row}}$ and $\mathbf{\Pi}^\top \mathbf{1} = \mathbf{w}_{\text{col}}$. Under entropy regularization, the minimizer admits the Gibbs form $\mathbf{\Pi}_{[i,j]} \propto \exp\big(-\frac{1}{\tau}\mathbf{C}_{[i,j]}\big)$. Substituting Eq. (1) yields the following factorized affinity kernel:

$$\begin{aligned}\mathbf{K}_{[i,j]} &= \exp\big(-\tfrac{1}{\tau}\mathbf{C}_{[i,j]}\big) \\ &= \exp\big(\tfrac{1}{\tau}\langle \mathbf{G}_{p[i]}, \mathbf{G}_{q[j]} \rangle_{\cos}\big)\,\big(1 + \langle \mathbf{S}_{p[i]}, \mathbf{S}_{q[j]} \rangle_{\cos}\big)^{\lambda/\tau}.\end{aligned} \tag{3}$$

which injects geometric affinity with semantic coherence, providing a soft semantic prior.

Algorithm 1 Log-Domain Sinkhorn with Target Marginals

Input: affinity $\mathbf{K}$, marginals $\mathbf{w}_{\text{row}}, \mathbf{w}_{\text{col}}$, iterations T
Output: transport matrix $\mathbf{\Pi}$
Initialize dual variables: $\mathbf{u} \leftarrow \mathbf{0}^n, \quad \mathbf{v} \leftarrow \mathbf{0}^n$
for $t = 1, \ldots, T$ **do**
  $\mathbf{u}_{[i]} \leftarrow \log \mathbf{w}_{\text{row}[i]} - \log \sum_j \exp\big(\log \mathbf{K}_{[i,j]} + \mathbf{v}_{[j]}\big)$
  $\mathbf{v}_{[j]} \leftarrow \log \mathbf{w}_{\text{col}[j]} - \log \sum_i \exp\big(\log \mathbf{K}_{[i,j]} + \mathbf{u}_{[i]}\big)$
end for
Recover transport plan:

$$\mathbf{\Pi}_{[i,j]} = \exp\big(\log \mathbf{K}_{[i,j]} + \mathbf{u}_{[i]} + \mathbf{v}_{[j]}\big)\,.$$

A.2. Semantic Denoising

As we indicated in main script, we follow STEGO [2] for semantic denoising. Fig. 1 shows the t-SNE embeddings of the raw DINO [6] features F_p, F_q and the denoised semantic features S_p, S_q from the query and reference views. After semantic denoising, features belonging to the same semantic parts from different viewpoints become significantly closer in the embedding space, indicating improved cross-view semantic consistency. This demonstrates that the lightweight semantic head effectively filters view-dependent noise in DINO features and provides more stable guidance for the subsequent OT correspondence.

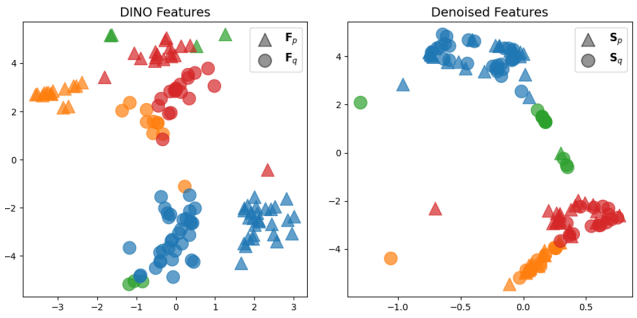

Figure 1. t-SNE results of DINO features on query and reference $\mathbf{F}_p, \mathbf{F}_q$, and denoised semantic features $\mathbf{S}_p, \mathbf{S}_q$, colored by k-means part segmentation. With semantic denoising, the distance between points from different views (triangle and circle) with the same semantic part is reducing dramatically.

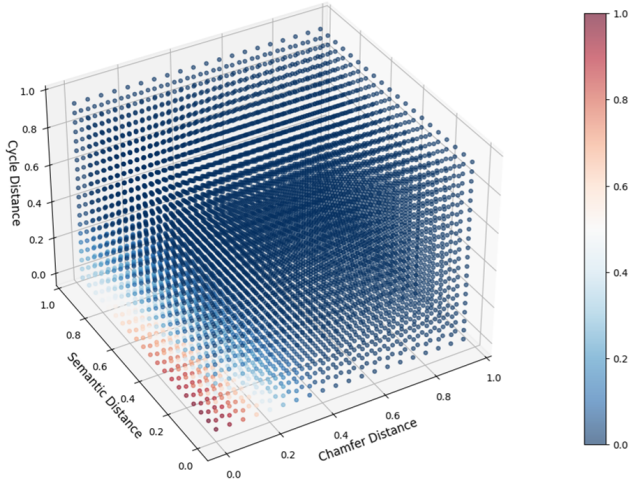

Figure 2. Visualization of pseudo confidence label generation. Points with smaller Chamfer, cycle, and semantic distances receive higher confidence labels.

A.3. Pseudo Confidence Labels

As an essential supervision signal of point validity, pseudo confidence labels are generated from Gaussian kernels defined over geometric ($\phi_{\text{cycl}}, \phi_{\text{pose}}$) and semantic ($\phi_{\text{sem}}$) distances. As illustrated in Fig. 2, points exhibiting small Chamfer, semantic, and cycle distances are assigned high confidence labels, while inconsistent or noisy points which contains larger distance receive low confidence labels.

With such design, the training process can be interpreted as an EM-like process that unfolds across epochs rather than within a single optimization step. Specifically, the previous training iterations act as the *E-step*, where the network estimates latent variables—soft correspondences and poses—and derives pseudo confidence labels based on the current model state. The current iteration then plays the role of the *M-step*, in which these pseudo confidence labels supervise the confidence branch, guiding the model toward more accurate correspondence and pose estimation within high confidence points. Through this implicit alternating process across iterations, COG progressively reinforces the consistency between confidence, correspondence, and pose, leading to stable performance without external supervision.

Method	LM-O [1]	TUD-L [3]	YCB-V [8]	Mean ↑
DINO+PHS [6]	24.7	16.2	45.2	28.7
COG (1% data)	52.9	66.9	73.6	64.5
COG (5% data)	55.7	70.5	75.7	67.3
COG (25% data)	55.1	70.7	75.4	67.2
COG (50% data)	55.4	71.8	74.7	67.3
COG (100% data)	56.7	73.8	75.9	68.8

Table 1. Quantitative comparison of DINO with pose-hypothesis-scoring (PHS) and COG trained with different fractions of the training data.

Stability. **Locality consistency**: Intuitively, once a point receives a high pseudo confidence label, its neighboring points, being geometrically close and semantically consistent, naturally inherit similar high confidence labels. This locality-aware consistency propagates stable supervision across nearby regions, making subsequent confidence predictions more coherent over iterations. **Normalized confidence**: To further ensure robustness in early training (when confidence labels are close to zero), all losses except $\mathcal{L}_{\text{conf}}$ are weighted by the normalized confidence $\mathbf{w} = \mathbf{c}/\overline{\mathbf{c}}$ rather than the raw confidence $\mathbf{c}$. Consequently, even if all predicted confidences start near zero, the optimization reduces to a uniform weighting scheme, preventing degenerate gradients and maintaining stable learning dynamics. **Detached labels**: In addition, the pseudo confidence labels used in $\mathcal{L}_{\text{conf}}$ are detached from backpropagation, avoiding harmful feedback loops that would push low confidence points farther.

In practice, with such design, the training process consistently behavior across all datasets, and we observe no signs of collapse or instability.

B. Experiment Details

B.1. Data Efficiency Analysis

As a task requires dataset-agnostic generalization, the data efficiency is important. To evaluate the data efficiency of our model, we train COG with progressively smaller subsets of the training data. Specifically, we randomly sample $(1\%, 5\%, 25\%, 50\%)$ of the full training set (2M images) and train our unsupervised model on each subset. For comparison with training-free pure semantic features, we replace the OT module with one-to-one correspondence based on DINO [6] feature similarity and apply a pose-hypothesis-scoring (PHS) scheme, which generates $6,000$ pose candi-

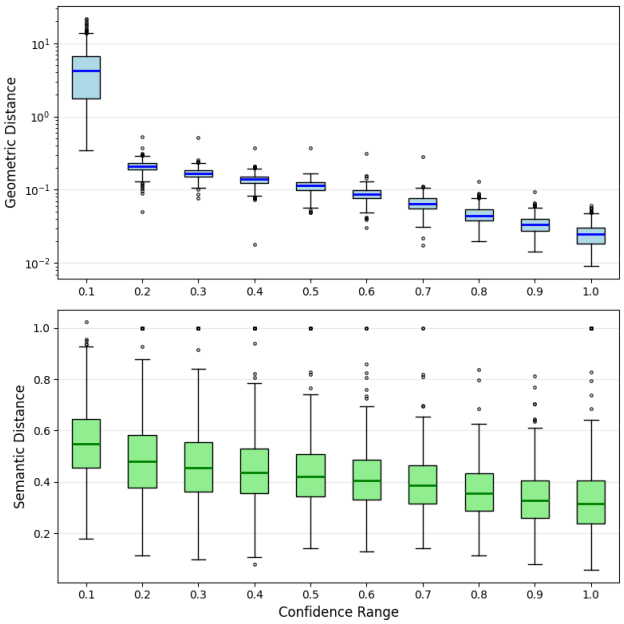

Figure 3. Relationship between predicted confidence and point-level distances for all test query-reference pairs in TUD-L [3] benchmark. The geometric distance (log scale) and semantic cosine distance are measured between each point and its nearest neighbor.

dates and selects the one with the highest matching score, following the coarse pose selection strategy used in [4, 5].

The results in Tab. 1 show that even with only 1% of the data (about 20k images), our method achieves a substantial performance gain over the DINO baseline, highlighting the importance of geometric reasoning and confidence learning for accurate pose estimation. Interestingly, increasing the data volume beyond 5% leads to only marginal improvements on LM-O [1] and YCB-V [8] benchmarks, while performance on TUD-L [3] benchmark exhibits a more noticeable gain. This difference reflects the characteristics of each benchmark: TUD-L contains objects with complex shapes that benefit from high-precision geometric alignment, whereas LM-O and YCB-V include more texture-rich objects, where semantic priors already provide strong guidance.

Overall, these findings demonstrate that unlike prior approaches relying on large-scale data, CAD models, or pose supervision, COG generalizes effectively even under extremely limited data and unsupervised setting, highlighting the data efficiency of our unsupervised framework.

B.2. Confidence Analysis

To further examine what the predicted confidence captures, we use all query-reference pairs in TUD-L [3], and measure for each point its nearest corresponding point in the other

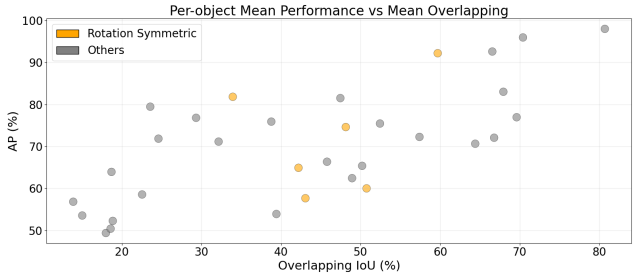

Figure 4. Per-object mean performance and overlapping.

view and compute two distances: geometric Euclidean distance in 3D space and semantic cosine distance $(1-\langle\cdot,\cdot\rangle_{\cos})$ in feature space.

As shown in Fig. 3, both geometric and semantic distances show a consistent downward trend as confidence increases, confirming that the confidence branch effectively learns to identify reliable correspondences, by jointly informed from geometric and semantic consistency.

B.3. Symmetry Analysis

For rotationally symmetric objects, such as bottles and bowls, the geometric and semantic features of different points often become indistinguishable. This ambiguity can lead to multiple valid transport solutions and, consequently, inaccurate pose estimation.

To mitigate this risk, we incorporate positional embeddings into the fine-phase geometry encoder to sharpen diffused features. These coordinate-based embeddings provide a soft constraint, biasing the model toward selecting correspondences within a local neighborhood rather than distant, yet geometrically similar, points. Furthermore, points with diffused correspondences typically fail the cycle consistency constraint, resulting in lower confidence scores.

As shown in Fig. 4, our evaluation of objects with identical overlapping ratios demonstrates no significant performance degradation attributable to symmetry.

B.4. Extended Results and Visualizations

In this section, we provide additional quantitative and qualitative results that complement the main experiments presented in the paper. Specifically, we report complete per-object results for all test objects on the LM-O, TUD-L, and YCB-V [1, 3, 8] benchmarks, as shown in Tab. 2. We also include extended visualizations of estimated poses in Fig. 5, illustrating the robustness of COG under varying occlusion levels, viewpoint changes, and object shapes. These extended results further validate the conclusions drawn in the main paper, demonstrating that COG achieves stable and accurate performance across diverse scenarios.

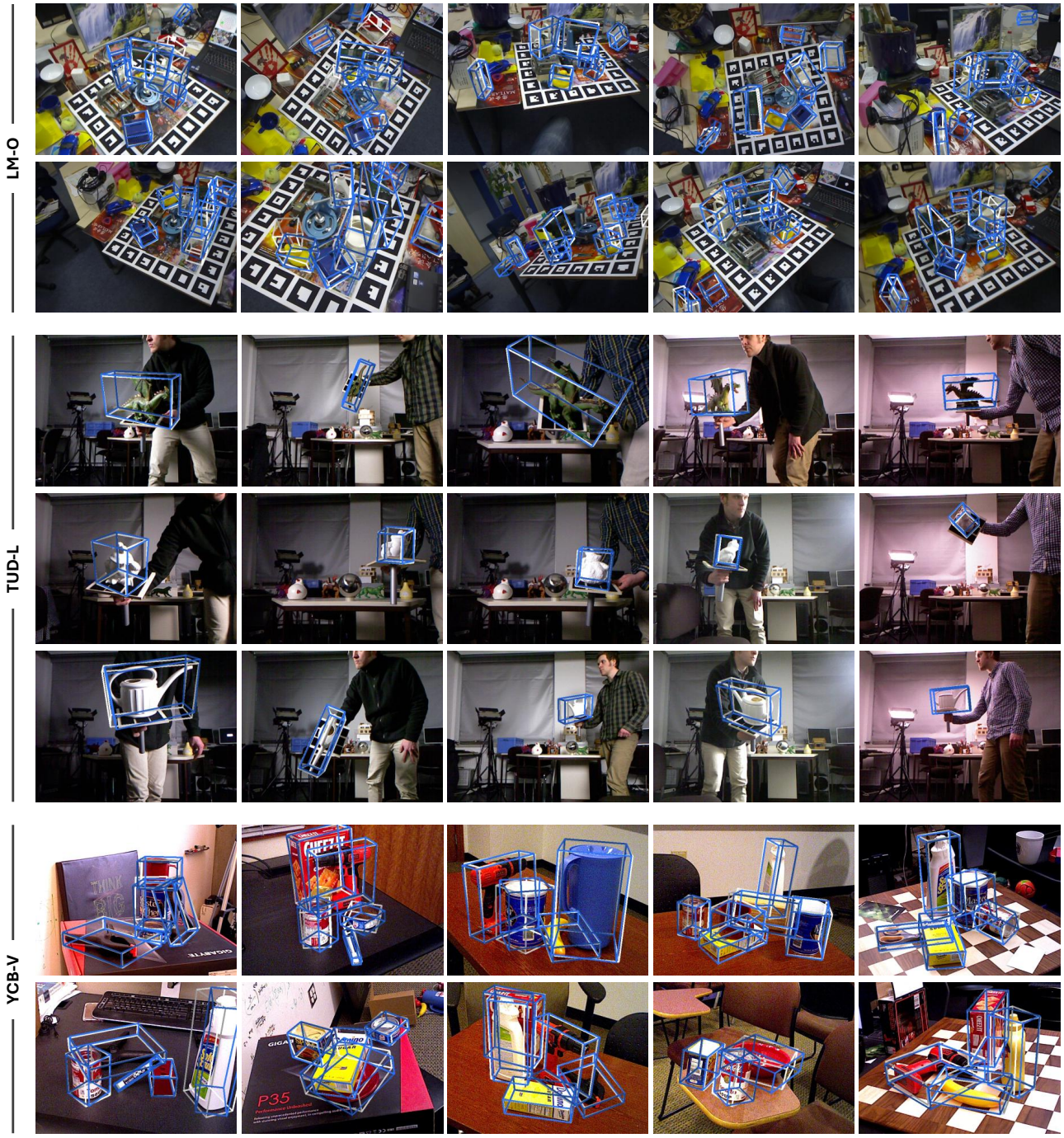

Figure 5. Results on 3 BOP benchmarks. Blue bounding boxes represent the estimated poses, while white boxes denote ground-truth poses.

C. Limitations

C.1. Segmentation Failure

Our framework relies on segmentation model to extract the masked depth for the point cloud inputs in the preprocessing. Therefore, segmentation errors can directly propagate to downstream pose estimation. As shown in Fig. 6 (a), if the segmentation model mistakenly includes different objects rather than the query object, the resulting point cloud may contain irrelevant geometry that cannot be recovered during pose estimation phase, leading to false correspondences and inaccurate. In future work, integrating joint segmentation-pose optimization may mitigate such cascading errors.

Benchmark	Object	VSD ↑	MSSD ↑	MSPD ↑	Mean ↑
LM-O [1]	ape	40.0	54.7	65.5	53.4
	can	51.1	70.2	68.0	63.1
	cat	37.6	53.1	58.1	49.6
	driller	47.7	56.1	51.6	51.8
	duck	53.1	59.2	68.6	60.3
	eggbox	42.9	59.8	63.6	55.4
	glue	38.7	55.2	54.8	49.6
	holepuncher	59.5	69.5	71.7	66.9
	Average	46.9	60.1	63.0	56.7
TUD-L [3]	dragon	64.2	83.3	81.0	76.2
	frog	72.5	74.8	79.4	75.6
	can	58.9	78.3	72.3	69.8
	Average	65.2	78.8	77.6	73.8
YCB-V [8]	master_chef_can	83.0	70.2	42.9	65.4
	cracker_box	71.0	79.6	62.5	71.0
	sugar_box	92.6	97.6	90.5	93.6
	tomato_soup_can	84.9	75.5	64.8	75.0
	mustard_bottle	81.7	89.5	74.6	81.9
	tuna_fish_can	72.8	56.1	53.8	60.9
	pudding_box	92.3	97.6	94.7	94.9
	gelatin_box	97.3	100.0	98.7	98.7
	potted_meat_can	57.4	62.7	55.9	58.7
	banana	75.7	87.7	64.7	76.0
	pitcher_base	87.4	86.5	59.9	77.9
	bleach_cleanser	80.1	88.2	68.0	78.8
	bowl	87.1	97.3	91.1	91.8
	mug	65.5	71.6	67.3	68.1
	power_drill	69.3	87.0	71.3	75.9
	wood_block	86.5	93.7	81.2	87.1
	scissors	51.0	86.4	71.2	69.5
	large_marker	62.6	93.3	89.8	81.9
	large_clamp	62.0	87.5	79.7	76.4
	extra_large_clamp	46.2	78.9	57.9	61.0
	foam_brick	80.8	86.4	83.3	83.5
	Average	76.5	82.0	69.1	75.9

Table 2. Per-object results on the LM-O [1], TUD-L [3], and YCB-V [8] benchmarks.

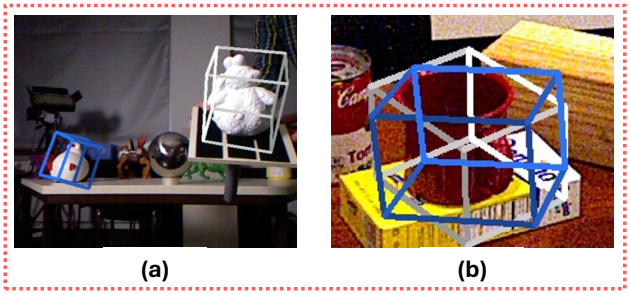

Figure 6. Failure cases of our method. (a) indicates the segmentation failure where segmentation mask is incorrect, and (b) indicates the unsupervised limitation discussed in Sec. C.2.

C.2. Unsupervised Limitation

As an unsupervised framework, COG optimizes pose implicitly, mainly through confidence-weighted Chamfer distance. This objective encourages minimizing high confidence point distances rather than explicitly enforcing a globally correct pose transformation. Consequently, when parts of the object contain sparse or noisy points, the network may prioritize aligning dense regions at the expense of small but semantically important parts. For example, as shown in Fig. 6 (b), when estimating the pose of a mug, the handle often contributes only few points; the network may achieve lower overall loss by ignoring this region and aligning only the main body of the mug. Although the introduction of semantic priors alleviates this issue to some extent, such priors remain soft constraints and cannot fully prevent these trade-offs. Future work could explore hybrid supervision or structural regularization (*e.g.* semantic part hard constraints) to better preserve critical fine-grained geometry during unsupervised optimization.

References

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 536–551. Springer, 2014. 2, 3, 5

[2] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. 1

[3] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 19–34, 2018. 2, 3, 5

[4] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27906–27916, 2024. 3

[5] Xingyu Liu, Gu Wang, Ruida Zhang, Chenyangguang Zhang, Federico Tombari, and Xiangyang Ji. Unopose: Unseen object pose estimation with an unposed rgb-d reference image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22023–22034, 2025. 3

[6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2

[7] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964. 1

[8] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2, 3, 5