# RACE-6D: Real-time Accurate Coarse-to-finE object 6D Pose Transformer

Supplementary Material

## 1. Detection Performance

Since our model performs object detection and 6D pose estimation in an end-to-end manner, evaluating its detection performance is equally crucial. Fundamentally, the True Positive (TP) counts and Recall values used in BOP metric calculations heavily rely on this underlying detection capability. Therefore, we evaluated the detection metrics using the same data employed for the BOP metric computation, with the results summarized in Table 1.

While the overall estimation performance is robust, the detection accuracy on certain datasets falls short when compared to state-of-the-art 2D detection methods from the BOP challenge. This limitation is a well-known inherent drawback of whole-image-based Vision Transformer architectures, which typically struggle to accurately localize relatively small objects, particularly in densely cluttered environments. Furthermore, we postulate that this issue is exacerbated by task interference inherent in our multi-task learning framework (e.g., balancing 2D localization with complex 6D pose regression).

## 2. Additional analysis of performance

Our proposed **bbox-norm** approach significantly enhances the stability of the model against bounding box inaccuracies. Specifically, a 0.1 decrease in Intersection over Union (IoU)—provided the initial IoU remains $> 0.7$—results in a marginal **1.9% average drop in the ADD-S metric**. Furthermore, in contrast to standard image-level normalization (img-norm), our scale-invariant **anchor+residual** representation simplifies the overall optimization process. This design choice leads to faster convergence on the LMO dataset (**37 epochs compared to 44 epochs**), alongside improvements in both Average Recall (AR) (**0.668 vs. 0.655**) and Keypoint (KPT) accuracy (increasing from 0.467 to **0.492**).

Table 1. Detection performance across BOP datasets.

| Dataset | Detection Performance | | |
|---|---|---|---|
| | **mAP 50:95** | **mAP 50** | **mAP 75** |
| LMO | 69.99 | 93.09 | 81.00 |
| TLESS | 81.19 | 90.82 | 89.49 |
| TUDL | 91.05 | 99.20 | 98.83 |
| IC-BIN | 66.81 | 88.37 | 79.33 |
| HB | 77.64 | 92.14 | 85.89 |
| YCBV | 87.43 | 99.21 | 97.96 |
| TLESS(D) | 81.87 | 90.33 | 88.03 |
| YCBV(D) | 87.13 | 97.96 | 96.52 |

Table 2. Impact of refinement iterations across BOP datasets.

| Dataset | Refinement iteration (AR / FPS) | | |
|---|---|---|---|
| | **Iter 1** | **Iter 2** | **Iter 3** |
| LMO | 0.651 / 107 | 0.667 / 95 | **0.668** / 87 |
| TLESS | 0.667 / 83 | 0.685 / 77 | **0.686** / 71 |
| TUDL | 0.789 / 108 | 0.793 / 96 | **0.795** / 88 |
| IC-BIN | 0.564 / 105 | 0.583 / 94 | **0.587** / 85 |
| HB | 0.590 / 102 | 0.598 / 92 | **0.603** / 83 |
| YCBV | 0.776 / 105 | **0.782** / 94 | 0.769 / 86 |
| **Avg.**$_{\text{RGB}}$ | 0.673 / 102 | 0.685 / 91 | **0.685** / 83 |
| TLESS(D) | 0.738 / 81 | 0.749 / 75 | **0.752** / 69 |
| YCBV(D) | 0.802 / 101 | 0.803 / 91 | **0.804** / 83 |

Regarding the hyperparameter $N$, we observed that increasing it beyond a sufficient threshold ($\approx 2\times$ the object count) provides negligible gains in AR, while strictly degrading the Frames Per Second (FPS). Consequently, we chose to keep $N$ fixed to maintain computational efficiency. Table 2 provides a detailed analysis of the trade-off between AR and FPS across different refinement iterations (up to a maximum of 3).

As shown in the table, the average RGB AR effectively saturates at refinement iteration 2, achieving an optimal balance between accuracy and speed. We even observe a marginal performance drop on the YCB-V dataset at iteration 3, likely due to error accumulation in highly occluded scenes. Conversely, depth-assisted evaluations (T-LESS(D) and YCB-V(D)) demonstrate continued, albeit slight, improvements up to iteration 3. Based on these observations, we suggest that the number of refinement iterations can be selectively configured between 2 and 3, depending on the specific application's latency constraints and available sensor modalities (RGB vs. RGB-D). Nevertheless, our runtime breakdown intentionally utilizes the 3-iteration setting to establish a rigorous upper bound for computational latency.

**Runtime Breakdown.** We conducted a detailed latency profiling on the YCB-V dataset using 3 refinement iterations to evaluate the real-time capability of our approach. The total inference latency strictly included in the FPS calculation is **11.7 ms** per frame, which translates to approximately **85 FPS**. This latency is composed of two main stages: **11.2 ms** for the network forward pass (which encompasses the backbone, decoder, and top-k bounding box post-processing), and **0.5 ms** for subsequent filtering and 3D coordinate transformation. It should be explicitly noted that the image pre-processing time (**5.3 ms**, primarily spent on image resizing and tensor conversion) and file I/O op-


Figure 5. **Qualitative results on BOP datasets (LM-O, IC-BIN, YCB-V).** The results demonstrate that our proposed projection consistency is strictly maintained even under heavy occlusion and complex backgrounds.

erations are excluded from this calculation. This exclusion aligns with standard real-time evaluation protocols in the field, isolating the core algorithmic efficiency.

## 3. Qualitative Results

To better understand the practical impact of our approach, we conducted a qualitative analysis of challenging cases within the LMO dataset, specifically focusing on instances with a confidence score $> 0.7$ but a translation error $> 10$ mm. Our findings indicate that models incorporating Keypoint (KPT) injection exhibited lower errors in **132 cases**, compared to only **90 cases** for models without it. Visually, the injection of KPT helps maintain projection consistency, ensuring that the 2D alignment remains robust even in the presence of severe 6D pose errors, as demonstrated in Fig. 6. However, while this robustness significantly aids in accurate 2D projection alignment, resolving the inherent depth scale ambiguity remains a challenge for future work. Qualitative results on other datasets can be seen in Fig. 5.

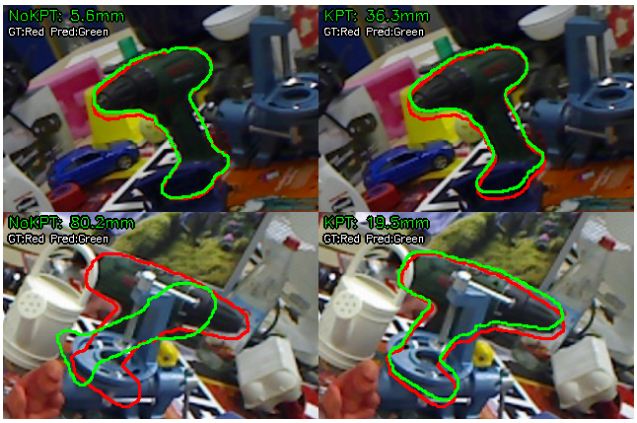


Figure 6. **Qualitative results on the Driller object.** The visualizations demonstrate how keypoint inference actively guides and maintains projection consistency despite severe 6D pose ambiguities.