WALDO: Where Unseen Model-based 6D Pose Estimation Meets Occlusion

Sajjad Pakdamansavoji Yintao Ma Amir Rasouli Tongtong Cao

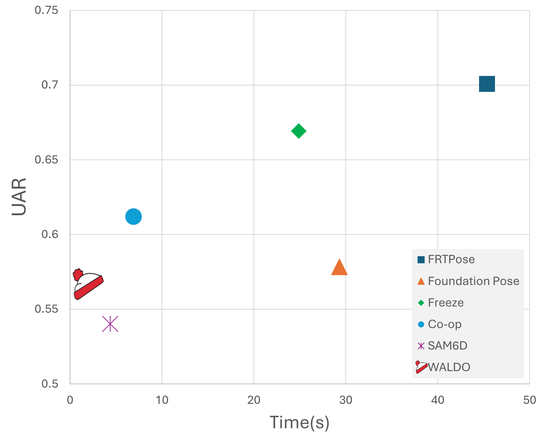
Huawei Technologies Canada

Stop occlusion errors from cascading: sample where the object is visible, keep several pose hypotheses and refine them, for +6 UAR on ICBIN and ≈3× faster inference than SAM6D.

Motivation

- **Unseen-object 6D pose** from a CAD model: detect → segment → hypothesize → refine.
- Under **occlusion**, early mistakes in masks and correspondences **propagate** through every later stage.
- Standard **average recall (AR)** is dominated by easy, highly visible instances, so it hides this failure.

Accuracy vs. speed

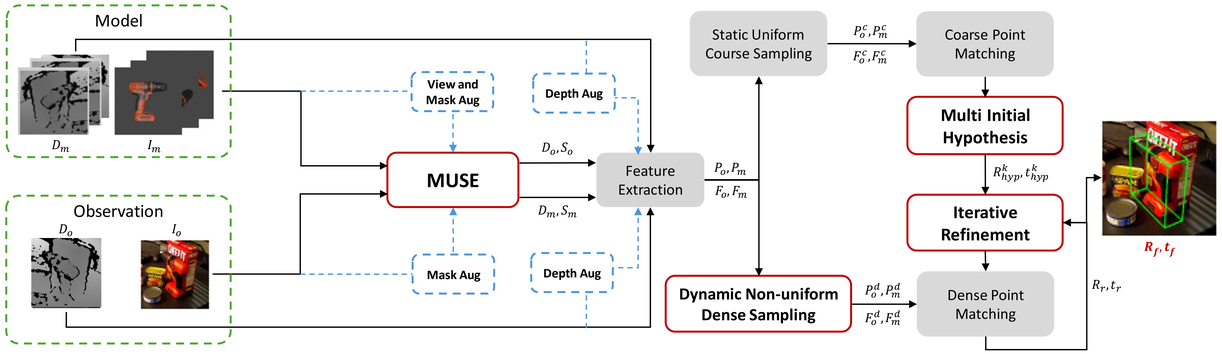


Unbiased recall (UAR) vs. runtime on BOP-Core*.

A fairer metric: visibility-weighted UAR

- Weight each instance by its **visibility decile**, so heavily occluded objects count as much as easy ones.
- On BOP-Core*, AR overstates WALDO's UAR by **0.17 (+30%)**: standard AR rewards the easy, visible cases.

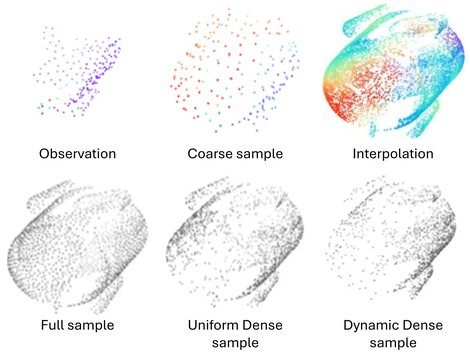
Method: four occlusion-aware extensions



1. **Dynamic non-uniform dense sampling**: estimate occlusion probabilities, then spend points on **visible regions**.
2. **Multi-hypothesis inference**: keep several confidence-ranked poses instead of one brittle path.
3. **Iterative refinement** of the surviving candidates.
4. **Occlusion-focused augmentation**: realistic depth noise and mask cut-outs during training.

Takeaways

- Make early stages **occlusion-aware** instead of trusting one mask.
- Point matching + SVD lifts UAR from 0.33 to 0.54; augmentation adds more.
- Report **UAR**, not only AR, for occluded scenes.



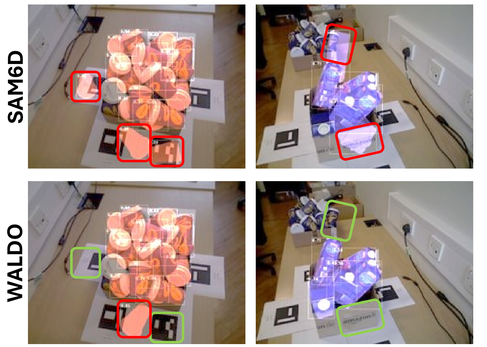
Occlusion probabilities (red = likely hidden) guide where to sample.

Results on BOP-Core* (UAR)

Model	LMO	TLESS	ICBIN	YCBV	Core*	s/img
FRTPose	0.68	0.61	0.67	0.61	0.70	45.4
FoundationPose	0.58	0.45	0.42	0.59	0.57	29.3
FreeZe	0.61	0.54	0.64	0.60	0.66	24.9
Co-op	0.55	0.45	0.52	0.60	0.61	6.9
SAM6D (base)	0.49	0.31	0.50	0.52	0.54	4.37
WALDO	**0.50**	**0.32**	**0.56**	**0.56**	**0.56**	**1.53**

- Beats its base model on **every dataset**: +6 UAR on ICBIN, +2 on BOP-Core*.
- 2.9× faster than SAM6D, 30× faster than FRTPose.
- Better detection too: AP 0.508 vs. 0.491.

Qualitative: fewer false positives



SAM6D vs. WALDO on cluttered ICBIN scenes.