AnyBox: Efficient Zero-Shot 9DoF Pose Estimation of Boxes for Robotic Manipulation

Yintao Ma[1,*] Sajjad Pakdamansavoji[1,*] Charles Eret[1,2] Rui Heng Yang[1]
Xuan Zhao[1] Yingxue Zhang[1] Tongtong Cao[1] Amir Rasouli[1]

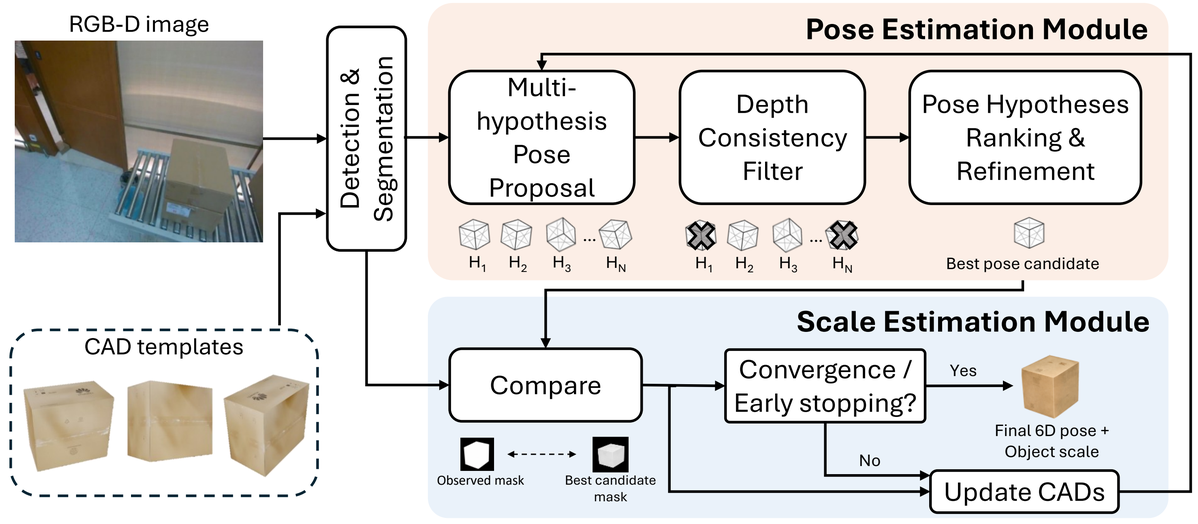
[1]Huawei Technologies Canada [2]University of Waterloo [*]Equal contribution

No per-box CAD model needed: one canonical template plus a mask-driven search recovers 6D pose *and* size, more than doubling detection AP and lifting robot box-shelving success by 28%.

Motivation

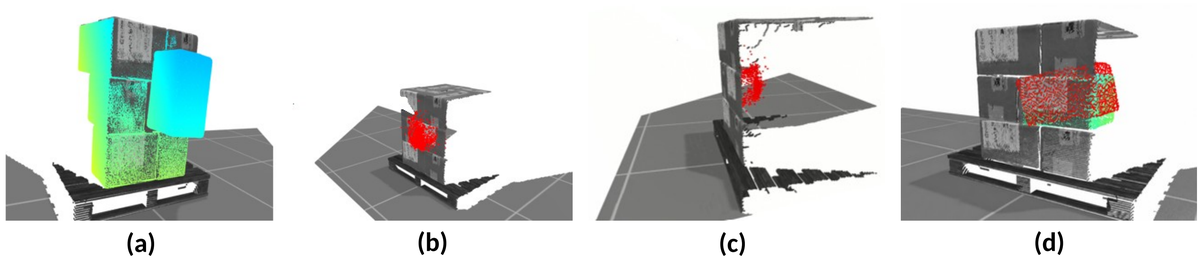
- Warehouses need the **9DoF pose** of every box: 6D pose **plus** 3D size.
- **Model-based** methods need an exact CAD model per box, which breaks as inventory changes.
- **Category-level** methods struggle with what boxes have most: **symmetry, weak texture and heavy occlusion** in stacks.
- But boxes are **geometrically simple**, a strong prior that existing methods ignore.

Method

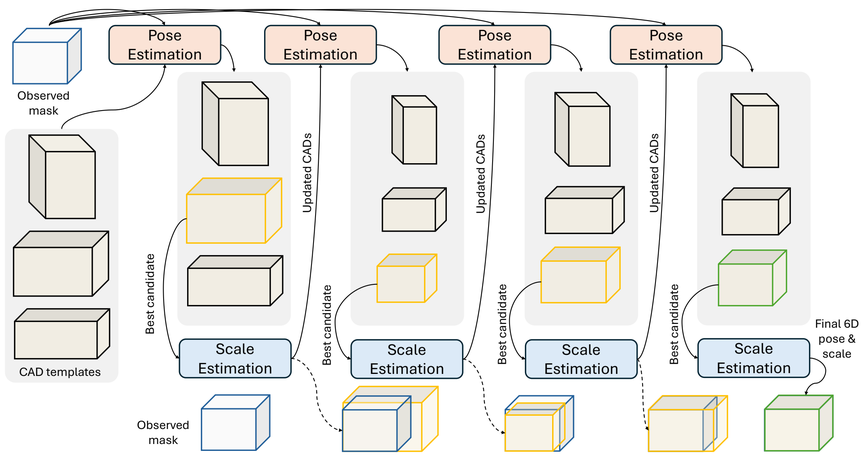


Detect with a canonical box template, hypothesize poses, filter, refine, then search over size.

1. **Pose–scale alternation**: start from one canonical template; the gap between the reprojected template and the observed mask drives a **binary search over box dimensions**.
2. **Depth-consistency filter**: reject symmetric hypotheses whose reprojected box protrudes from the observed depth.
3. **Early stopping**: once rotation converges, finish with one **closed-form scale update** instead of more search steps.



Depth-consistency filter: (a) a rejected hypothesis protruding from the stack; (b–c) hypotheses before and after filtering.



Binary search: rank templates, estimate scale, rescale until the mask converges.

Ablations

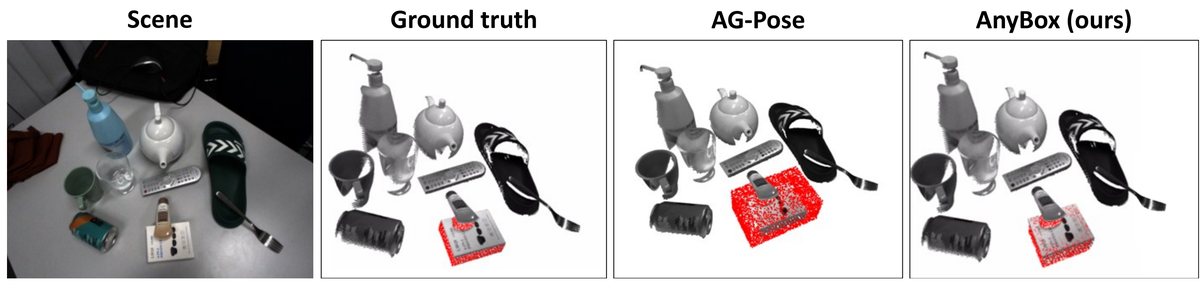
- Depth filter: precision at IoU 0.8 rises from **0.53 to 0.86**.
- Early stopping: **4.9 s** $\rightarrow$ **1.2 s** per object (4$\times$ faster) for a 0.02 drop at IoU 0.9.

Results: public benchmarks

HouseCat6D, 6D pose AP

Model	IoU.25	IoU.50
NOCS	43.3	6.5
VI-Net	44.8	12.7
SecondPose	54.5	23.7
AG-Pose	57.2	7.7
AnyBox	**88.8**	**58.9**

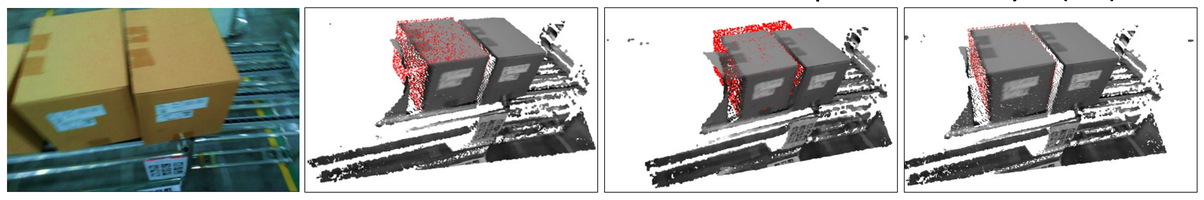
- **HouseCat6D**: **+31.6** AP at IoU 0.25, and 2.5$\times$ the best AP at IoU 0.50.
- **PACE**: detection AP 63.2 vs. 27.2 for CPPF++ (**+36 points**); ADD-S AP 49.2 vs. 5.0 at 0.1d.



Predicted boxes on HouseCat6D.

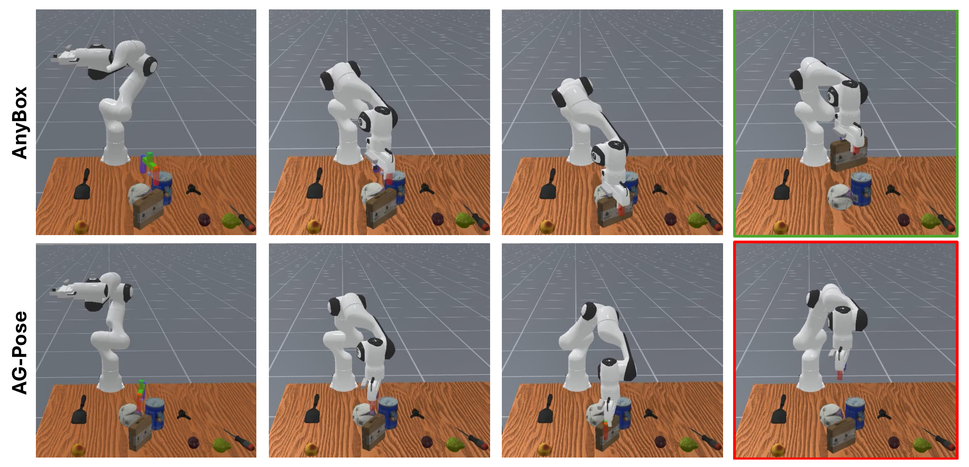
Results: Warehouse6D

- With only a canonical CAD, SAM6D collapses (det. AP@0.5 = **0.06**).
- AnyBox reaches **1.00**, matching SAM6D **with ground-truth CAD**, and ADD-S AP of **82.8%** at 0.15d.



Baseline vs. AnyBox on stacked warehouse boxes.

Downstream: robot box shelving



The robot grasps the box with AnyBox's pose but fails with AG-Pose's wrong orientation.

Planner	Pose	2 distr.	4	8	32
RRT	AG-Pose	0.45	0.40	0.42	0.03
	AnyBox	**0.59**	**0.68**	**0.56**	**0.12**
π_0 VLA	AG-Pose	0.62	0.43	0.26	0.07
	AnyBox	**0.71**	**0.49**	**0.30**	**0.08**

Conclusion

- Exploit box geometry: one template, **zero-shot** across sizes.
- Near GT-CAD accuracy, and better pose means more successful grasps.