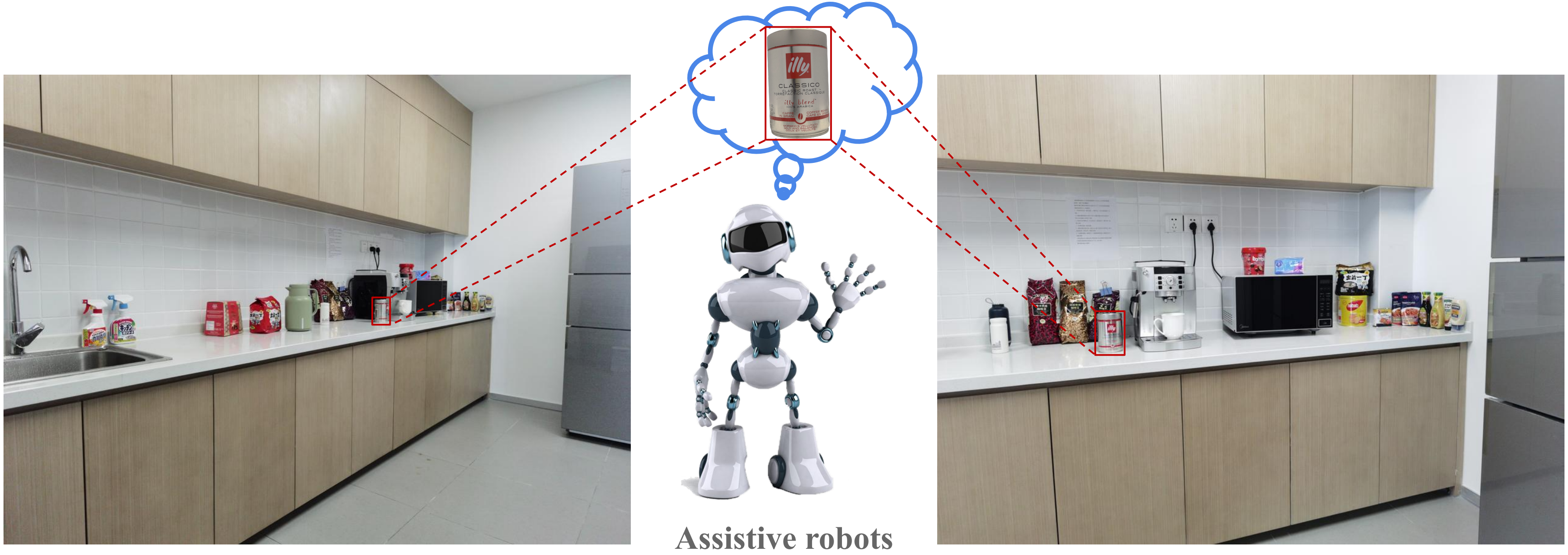
NEURAL INFORMATION
PROCESSING SYSTEMS

A High-Resolution Dataset for Instance Detection with Multi-View Object Capture

Qianqian Shen, Yunhan Zhao, Nahyun Kwon, Jeeeun Kim, Yanan Li, Shu Kong

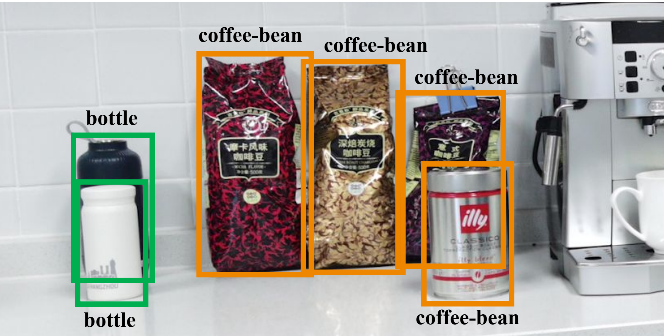
Instance Detection



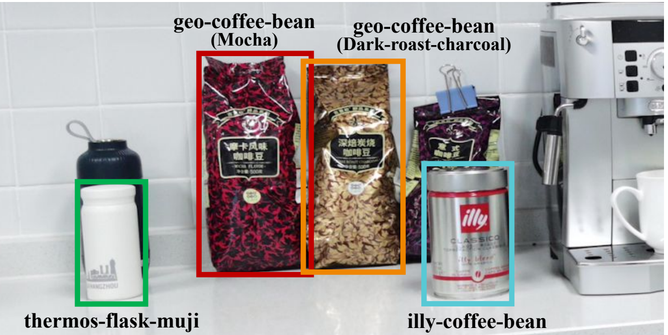
locating the **wanted** object at distance!

Object Detection (ObjDet) *vs.* Instance Detection (InsDet)

{ coffee-bean,
bottle,
cup,
… }



ObjDet aims to detect all objects belonging to some predefined classes.

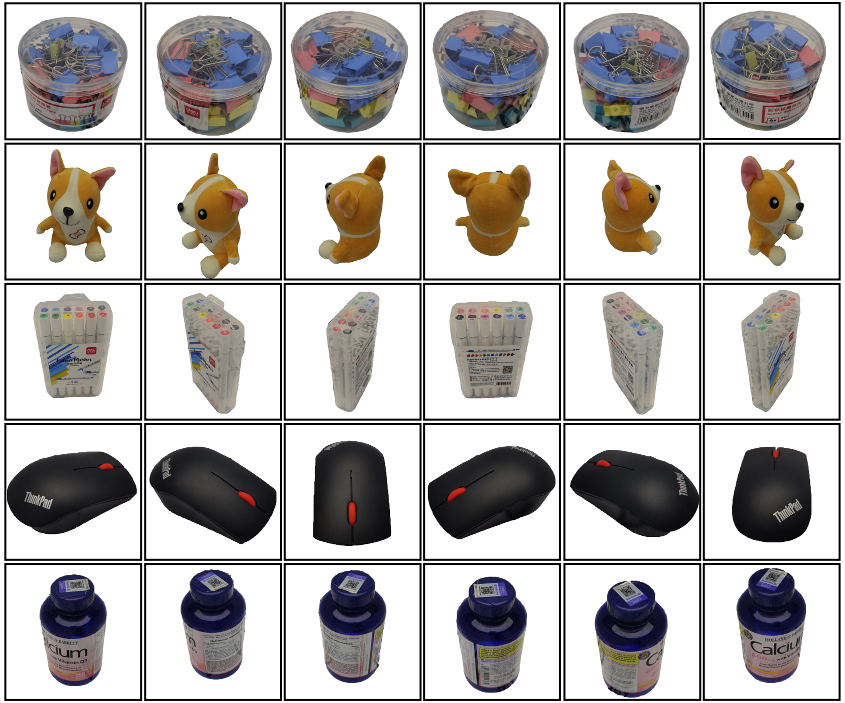


InsDet requires detecting specific object instances defined by some visual examples.

Our dataset: InsDet

Instance's profile images

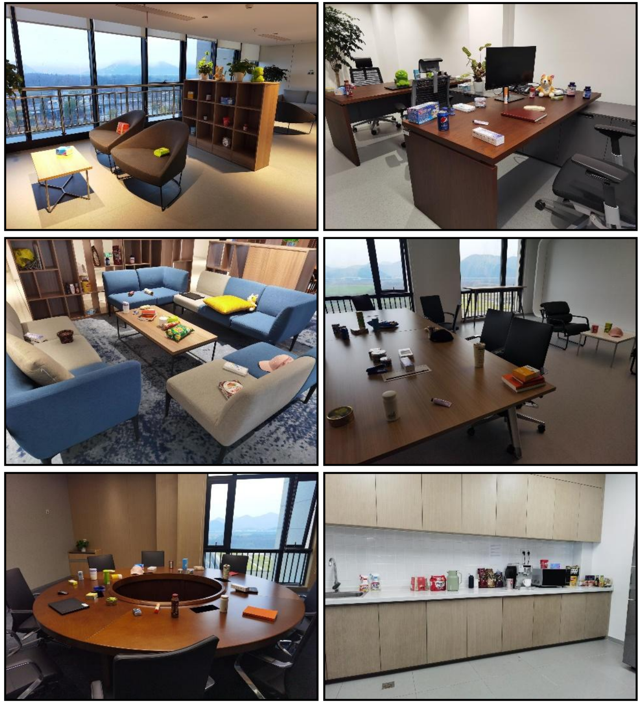
- **100** object instances
- **24** samples per instance



3072×3072

Real-World Scenes

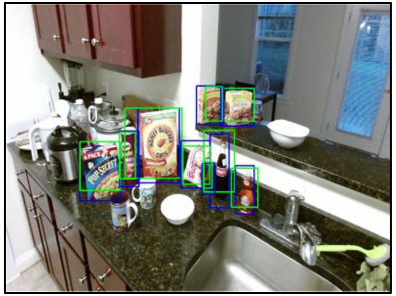
- **Diverse** scenes
- **High-resolution** images



6144×8192

Comparison against existing datasets

- **23** instances
- **9** scenes
- resolution: **1080×1920**
- publicly available



GMU dataset
[Georgakis et al. 2016]

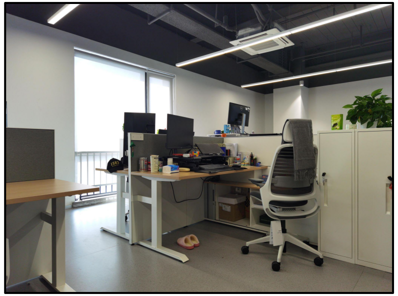
- **33** instances
- **9** scenes
- resolution: **1080×1920**
- publicly available

AVD dataset
[Ammirato et al. 2017]

- **100** instances
- **10** scenes
- resolution: **unknown**
- publicly unavailable

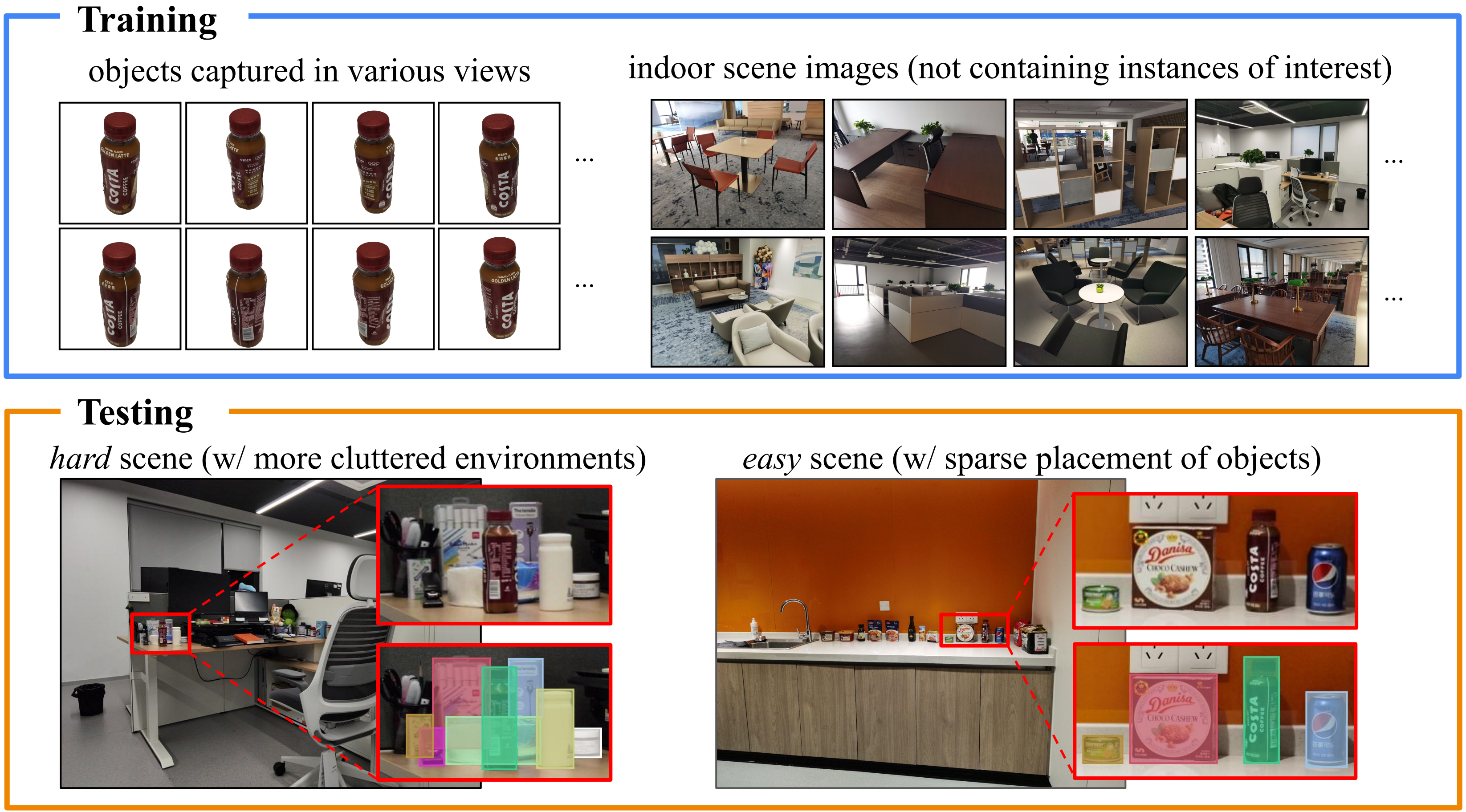
Grocery dataset
[Bormann et al. 2021]

- **100** instances
- **14** scenes
- resolution: **6144×8192**
- publicly available



InsDet dataset
[Shen et al. 2023]

A unified InsDet protocol



Training

objects captured in various views

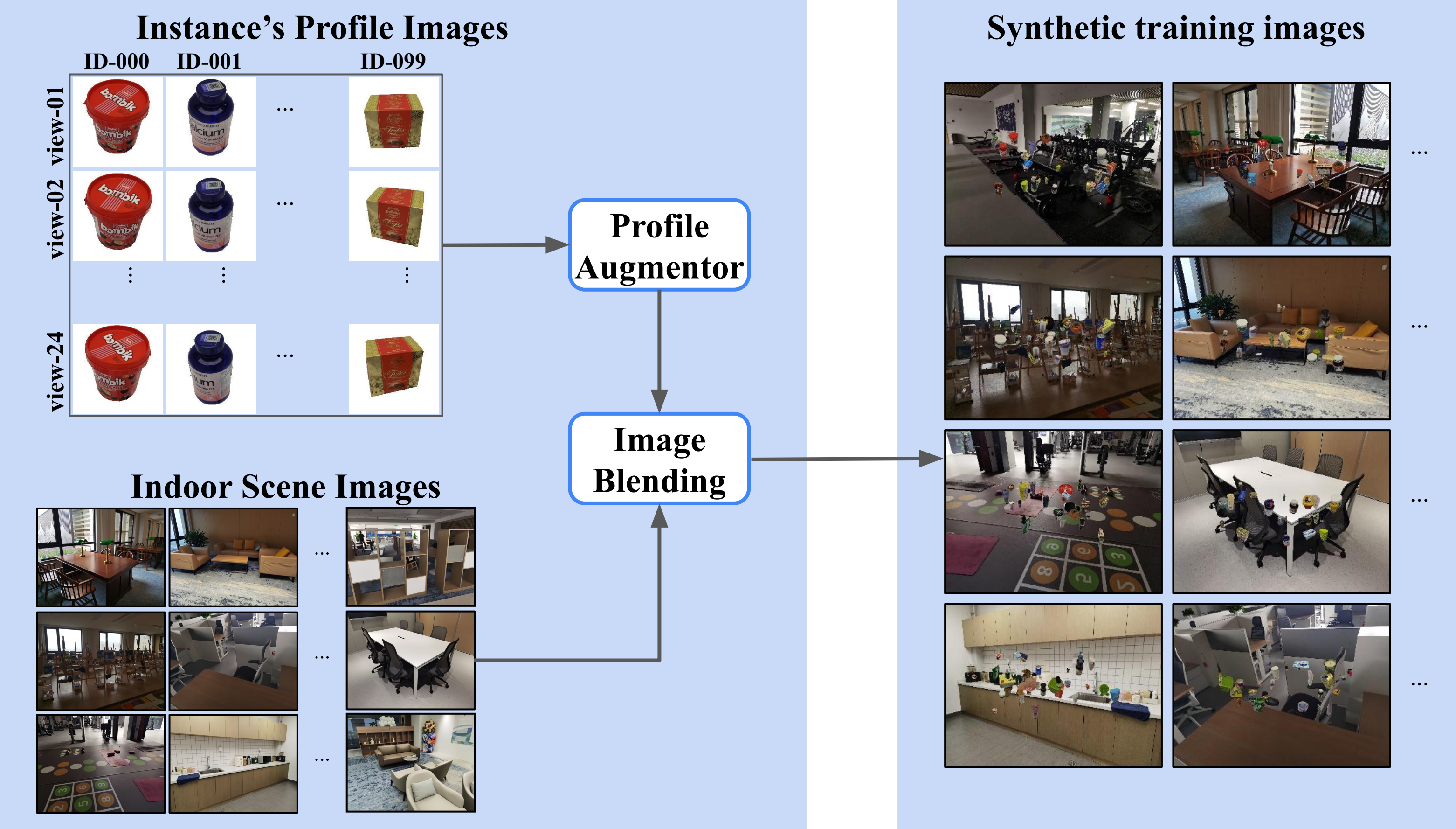
indoor scene images (not containing instances of interest)

Testing

hard scene (w/ more cluttered environments)

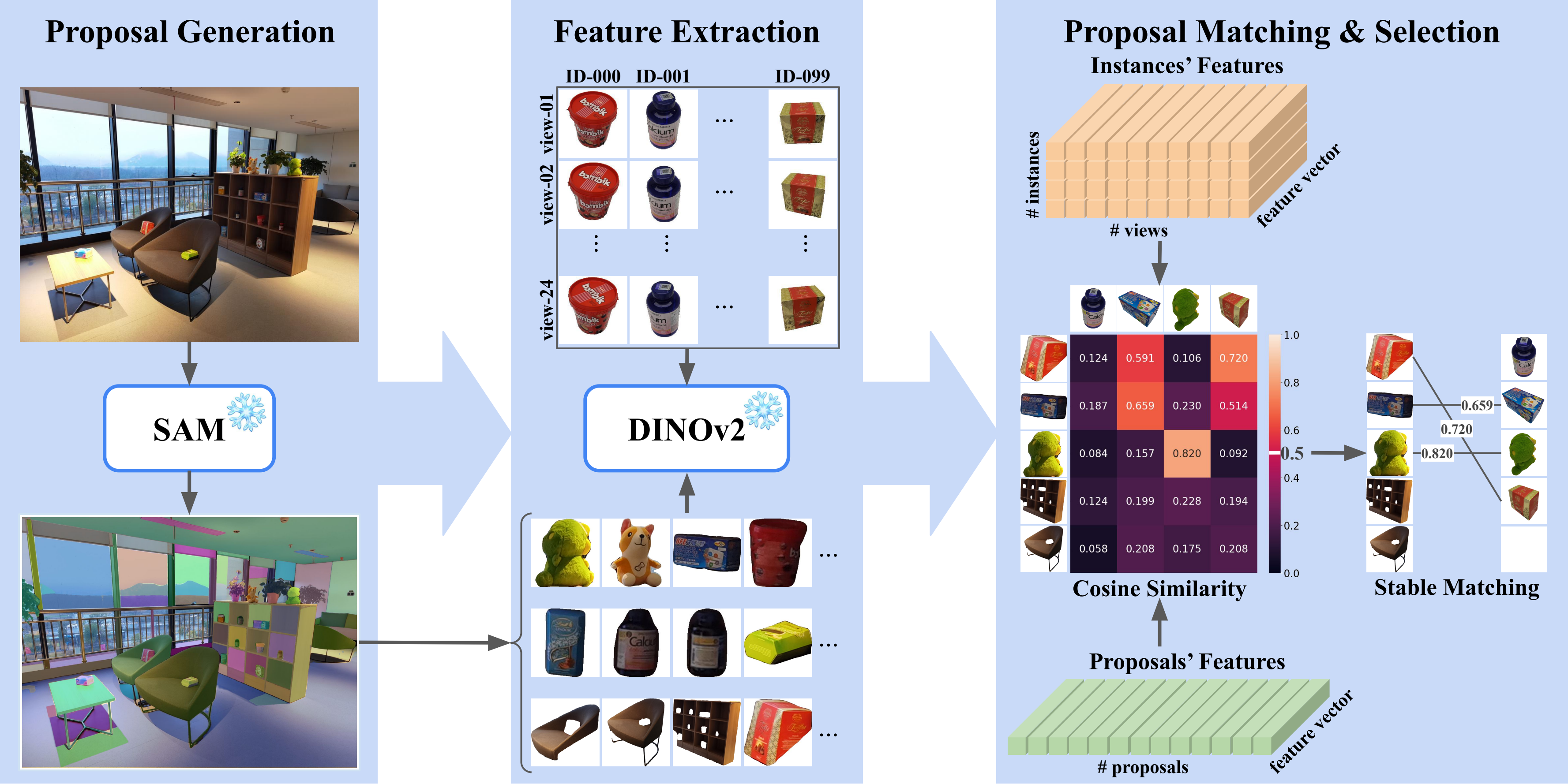
easy scene (w/ sparse placement of objects)

Baseline: Cut-Paste-Learn



[1] Cut, paste and learn: Surprisingly easy synthesis for instance detection. In ICCV, 2017.

A simple, non-learned method



Proposal Generation

SAM

Feature Extraction

DINOv2

Proposal Matching & Selection

Instances' Features

Cosine Similarity

0.124	0.591	0.106	0.720
0.187	0.659	0.230	0.514
0.084	0.157	0.820	0.092
0.124	0.199	0.228	0.194
0.058	0.208	0.175	0.208

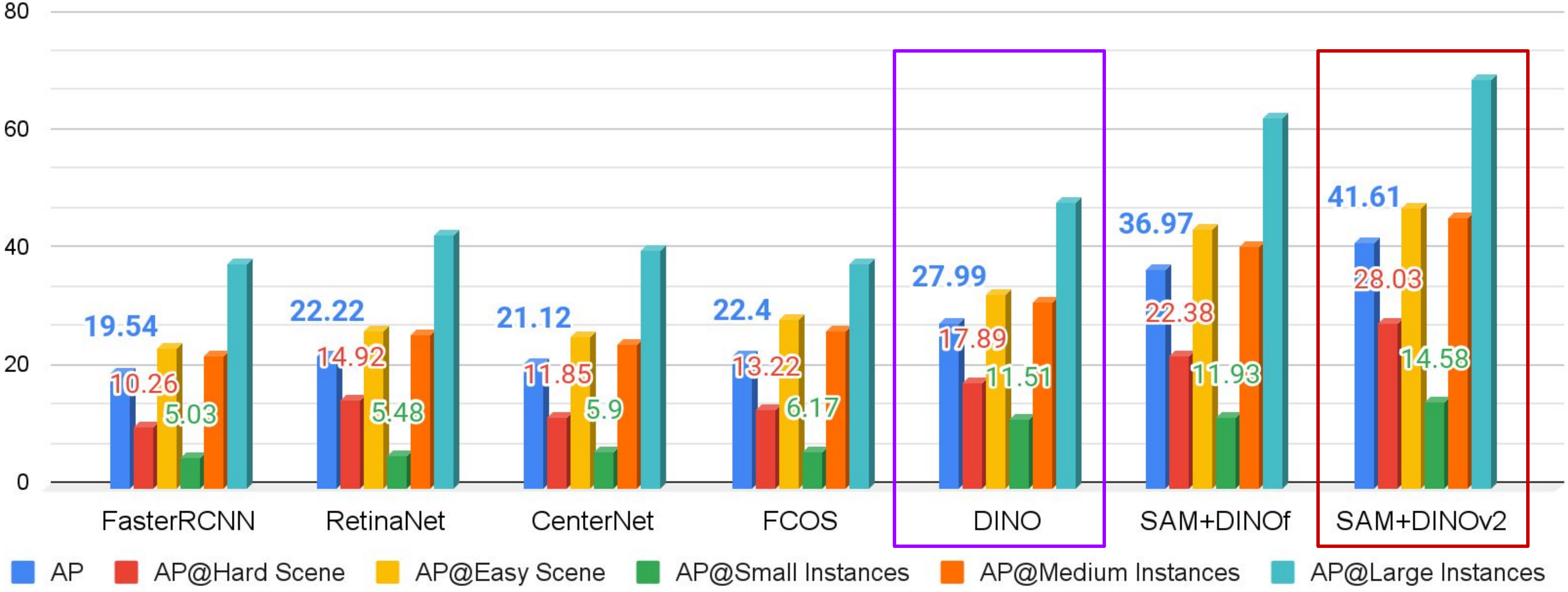
Stable Matching

Proposals' Features

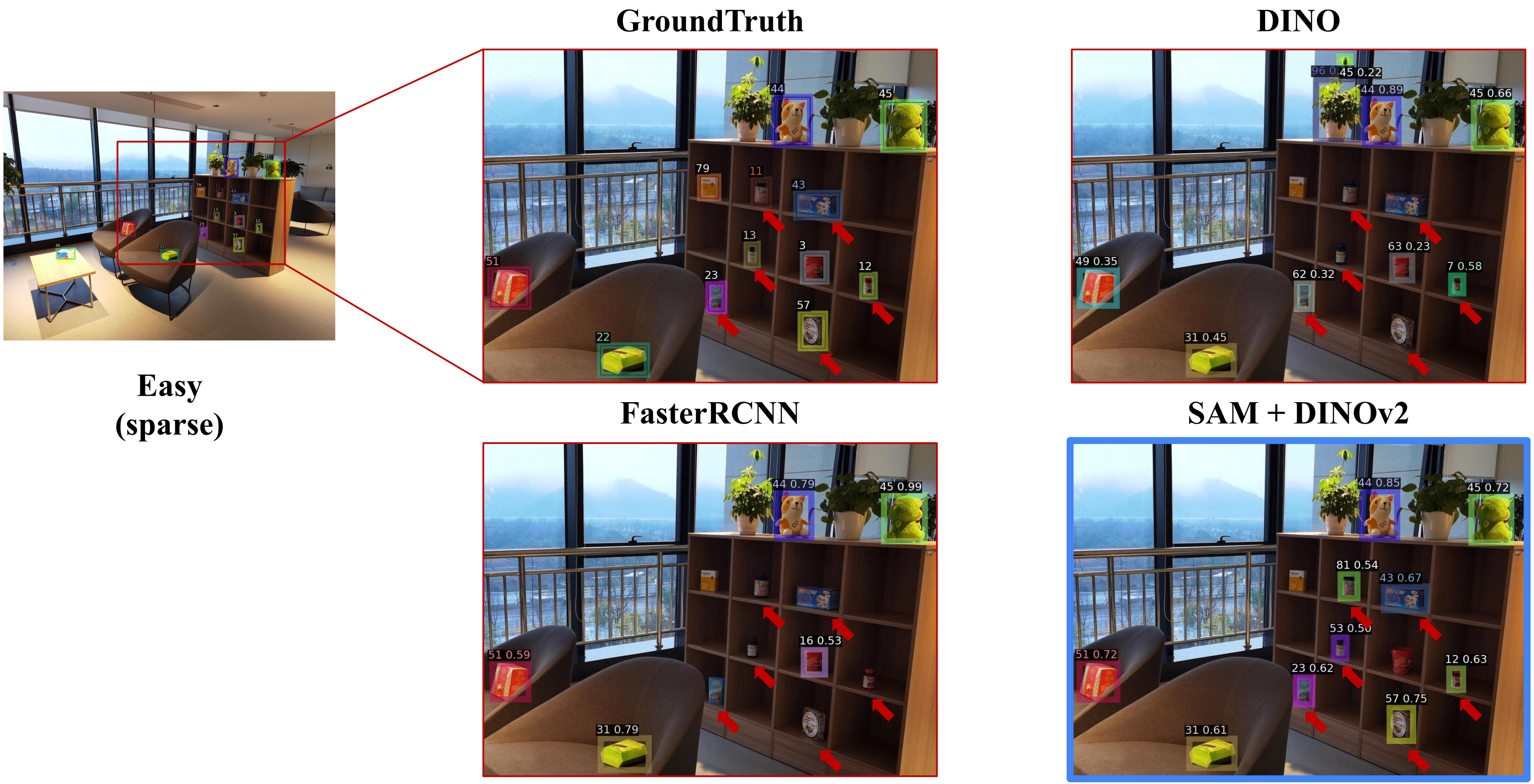
[1] Segment anything. In ICCV, 2023.

[2] DINOv2: Learning robust visual features without supervision. In arXiv, 2023

Benchmarking results



Qualitative evaluations on *easy* scenes



Easy
(sparse)

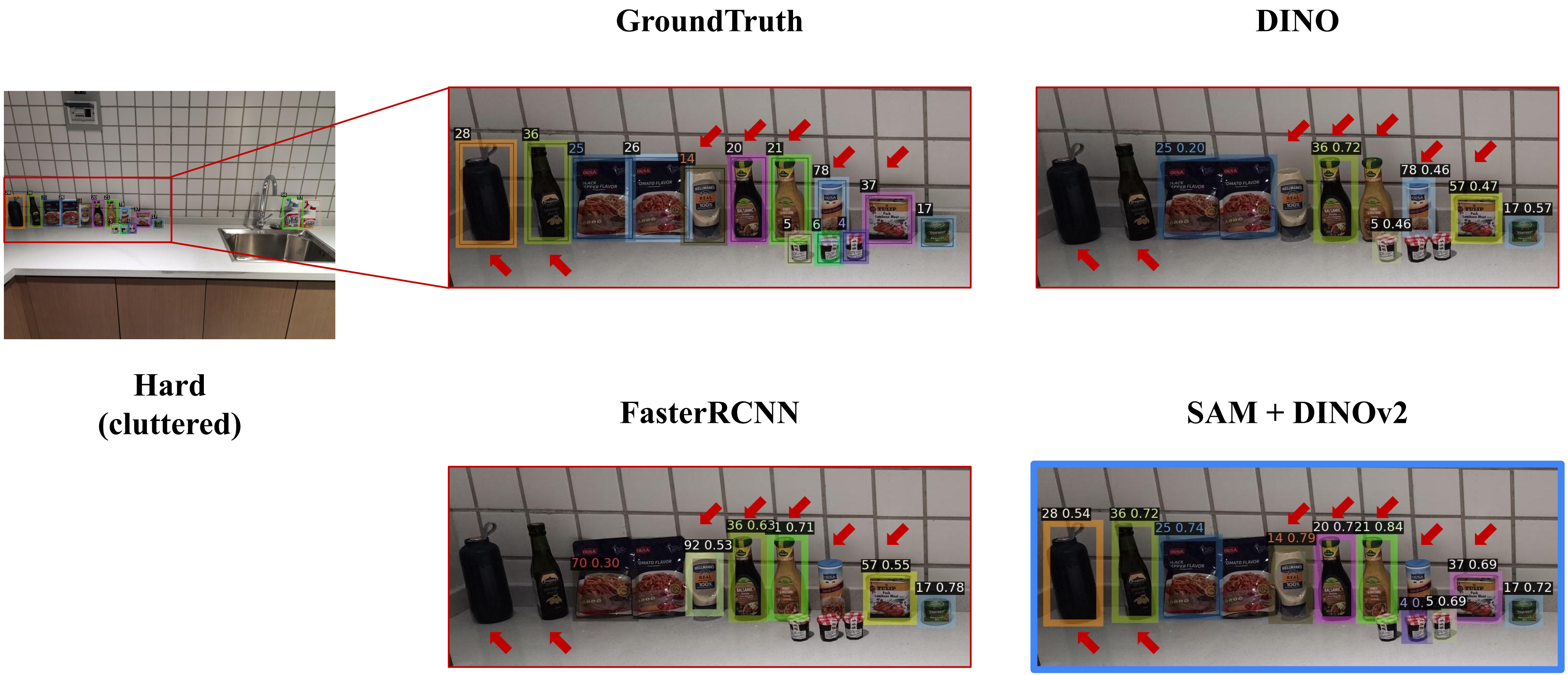
GroundTruth

DINO

FasterRCNN

SAM + DINOv2

Qualitative evaluations on *hard* scenes



GroundTruth

DINO

Hard (cluttered)

FasterRCNN

SAM + DINOv2

Suggested further directions

- **Exploring high-resolution images**

  Leverage high-resolution visual signals to help detect small objects.

- **Exploring faster algorithms**

  Building multi-scale detectors for more efficient processing.

- **Exploring more foundational models**

  Learn lightweight adaptors to bridge pretrained foundational models for better performance.

Thank You!

https://github.com/insdet