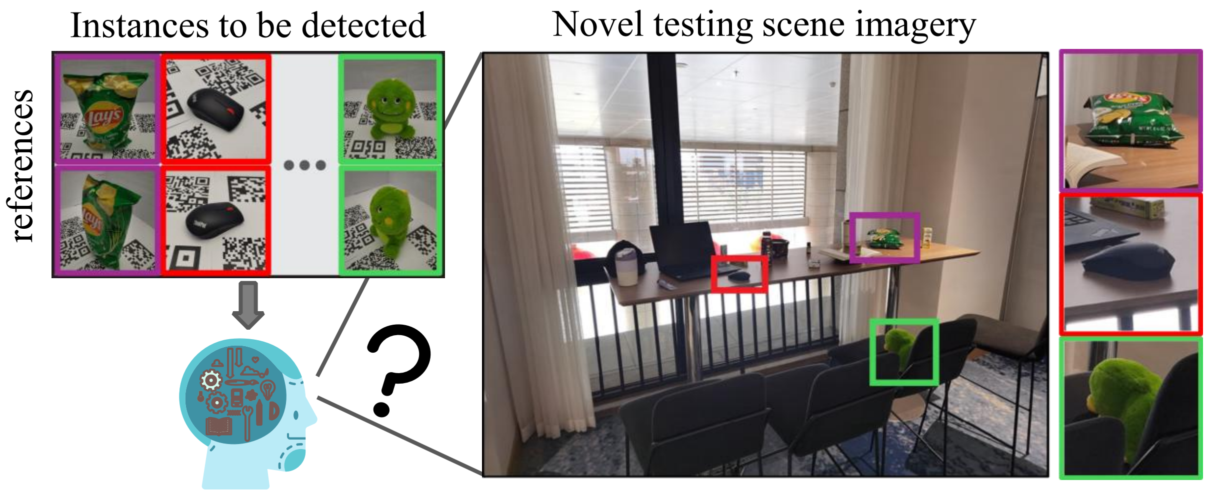
CVPR Nashville JUNE 11-15, 2025

Solving Instance Detection from an Open-World Perspective

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

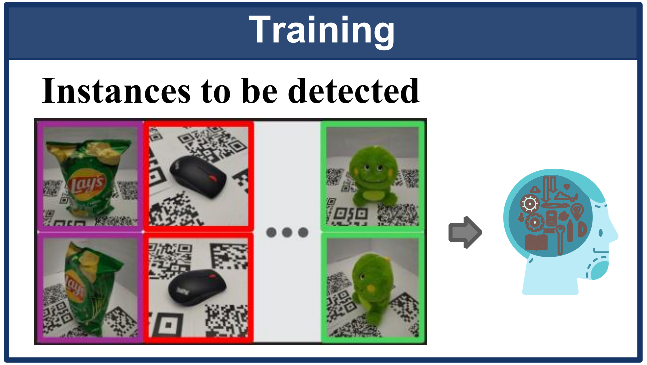
Project Page

Instance Detection (InsDet)



locating specific objects based on given visual references

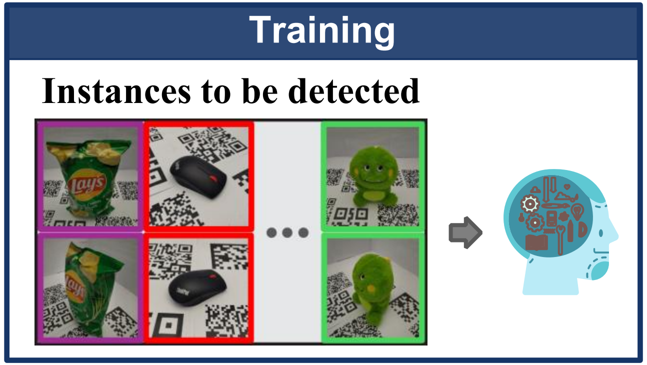
Two Settings: CID & NID



Conventional Instance Detection (CID)

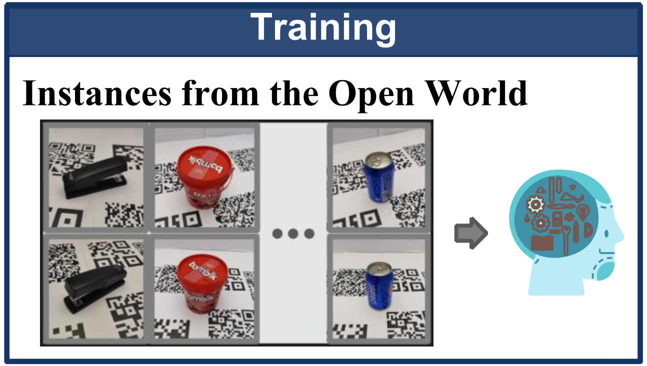
- Instances to be detected are pre-defined during training.
- The scene images are unknown in testing.

Two Settings: CID & NID



Conventional Instance Detection (CID)

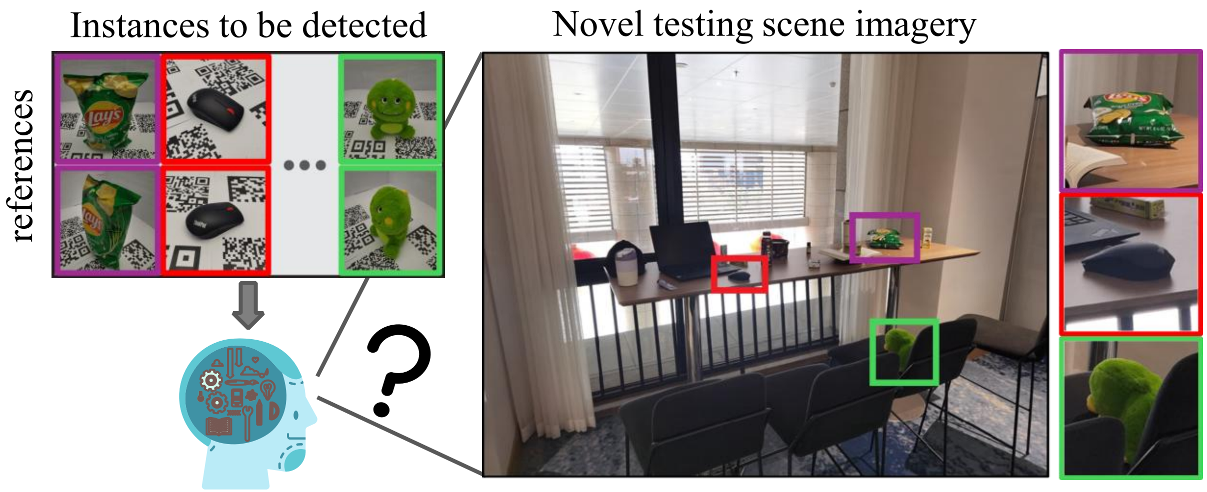
- Instances to be detected are pre-defined during training.
- The scene images are unknown in testing.



Novel Instance Detection (NID)

- Instances to be detected are defined only in testing.
- The trained models are not allowed to be finetuned further during testing.

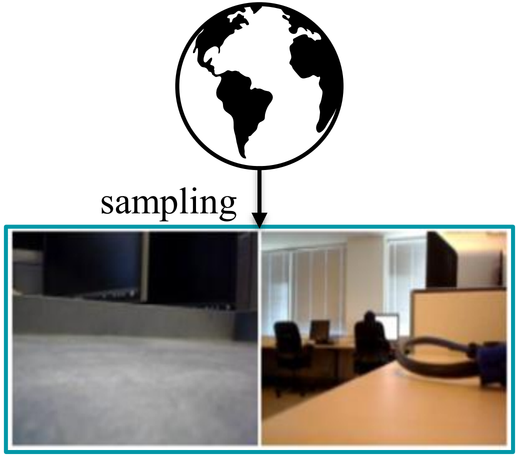
Open-World Challenges of InsDet



- The testing imagery is never-before-seen and unknown to an instance detector.
- Domain gaps between visual references and detected proposals.
- Robustness and generalization are desperately needed to detect diverse instances.

Existing methods partially exploit the open-world information.

(a) Background Imagery



synthesize scene images for training,
e.g., Cut-Paste-Learn [ICCV2017]

partially addressing **unknown** testing scene distribution

Dwibed & Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection", ICCV, 2017.

Existing methods partially exploit the open-world information.

(a) Background Imagery

sampling

synthesize scene images for training,
e.g., Cut-Paste-Learn [ICCV2017]

partially addressing **unknown** testing scene distribution

(b) Object Images

sampling

learn personalized representation,
e.g., VoxDet [NeurIPS2023]

partially addressing **domain gaps** between proposals and references

Dwibed & Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection", ICCV, 2017.
Li et al., "VoxDet: voxel learning for novel instance detection", NeurIPS, 2023.

Existing methods partially exploit the open-world information.

(a) Background Imagery

sampling

synthesize scene images for training,
e.g., Cut-Paste-Learn [ICCV2017]

partially addressing **unknown** testing scene distribution

(b) Object Images

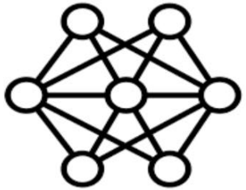
sampling

learn personalized representation,
e.g., VoxDet [NeurIPS2023]

partially addressing **domain gaps** between proposals and references

(c) Foundation Models

sampling



leverage foundation models,
e.g., OTS-FM [NeurIPS2023]

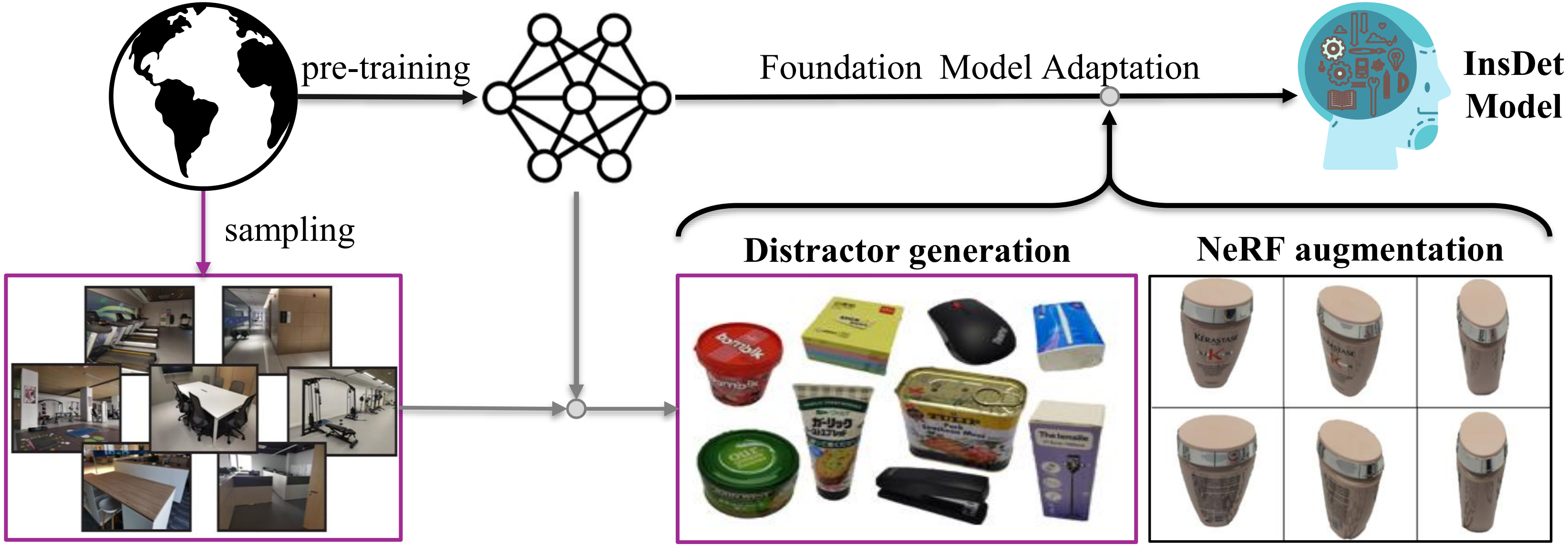
improving proposal detectors and feature representations

Dwibed & Hebert, “Cut, paste and learn: Surprisingly easy synthesis for instance detection”, ICCV, 2017.
Li et al., “VoxDet: voxel learning for novel instance detection”, NeurIPS, 2023.
Shen et al., “A high-resolution dataset for instance detection with multi-view object capture”, NeurIPS, 2023.

Our Philosophy: Addressing InsDet in the Open World (IDOW)

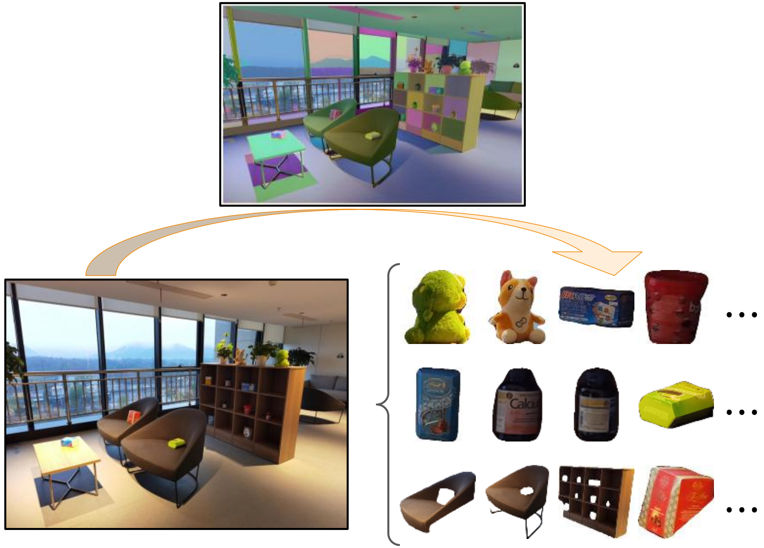
Thoughts:

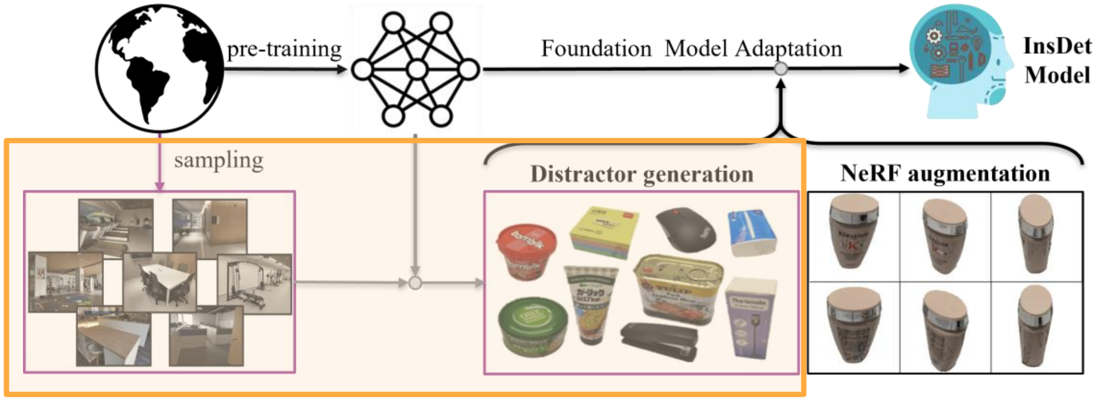
- A foundational detector yields high recall, i.e., SAM detecting all instances of interest. Let's focus on instance matching.
- Using features of DINOv2 for matching is promising but far from perfect. Let's finetune it.
- Data examples in the open world are diverse. Let's sample both synthetic and real data.



Kirillow et al, "Segment Anything", ICCV, 2023.
Liu et al, "GroundingDINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection", ECCV, 2024.
Ben et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". ECCV, 2020.
Oquab et al, "DINOv2: Learning Robust Visual Features without Supervision", Arxiv, 2023.

Sampling Distractor Instance from Real Imagery

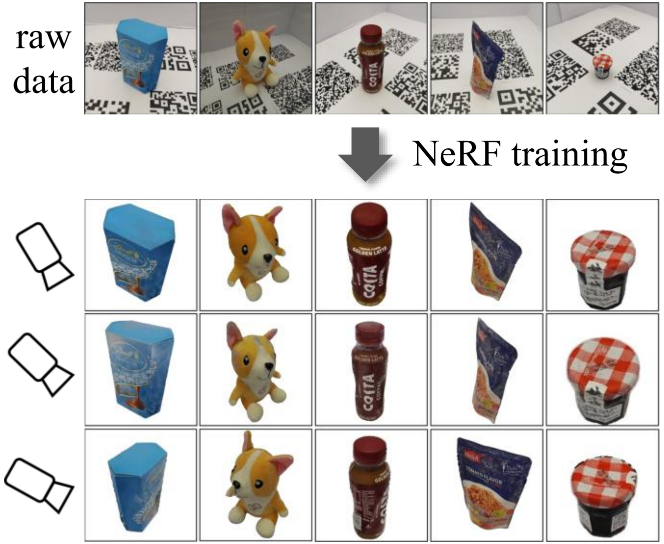


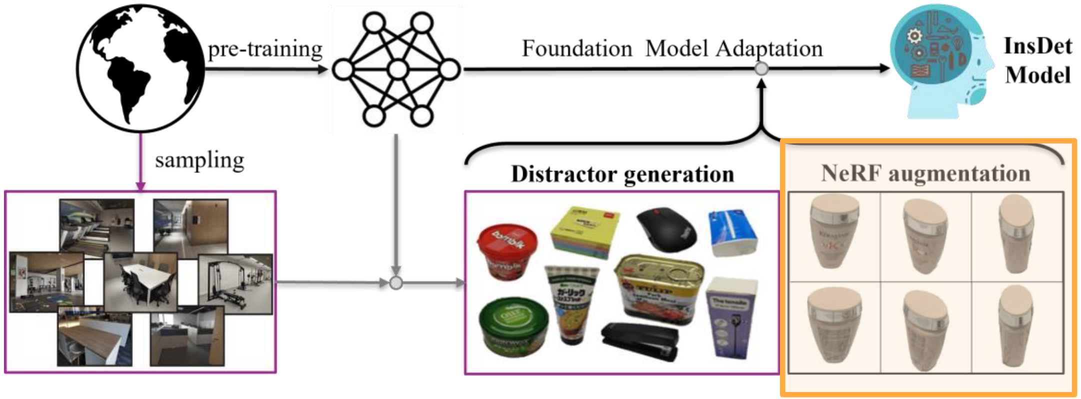


sample distractors from diverse images in the open world by SAM

Kirillow et al, "Segment Anything", ICCV, 2023.
Oquab et al, "DINOv2: Learning Robust Visual Features without Supervision", Arxiv, 2023.

Sampling More Positive Instances



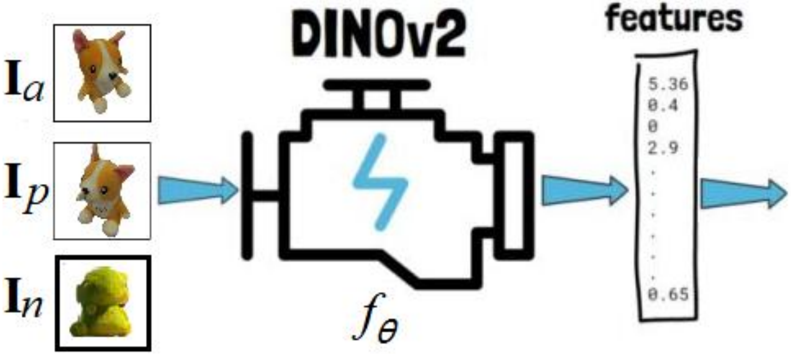


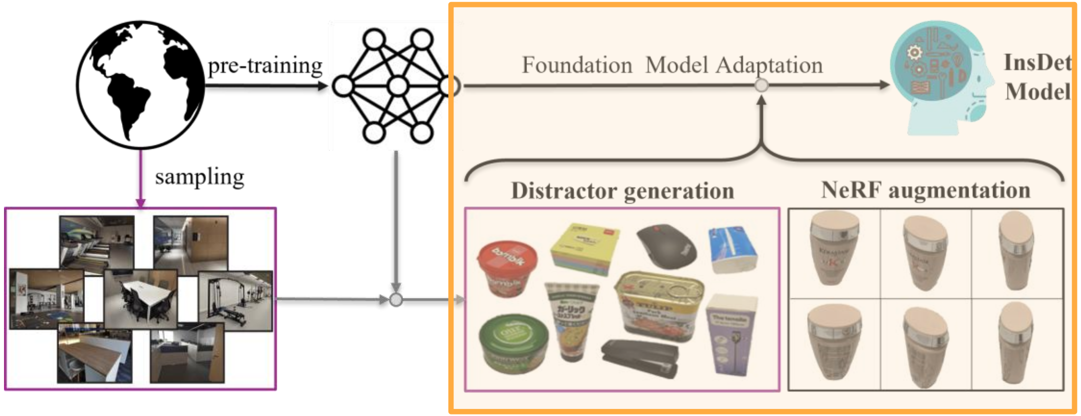
synthesize novel-view images by using NeRF on the given visual references

Ben et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". ECCV 2020.
Barron et al, "Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields", ICCV, 2023.

Adapting DINOv2 using Metric Learning

$$\ell = \left[d(f_\theta(\mathbf{I}_a), f_\theta(\mathbf{I}_p)) - d(f_\theta(\mathbf{I}_a), f_\theta(\mathbf{I}_n)) + \alpha\right]_+$$

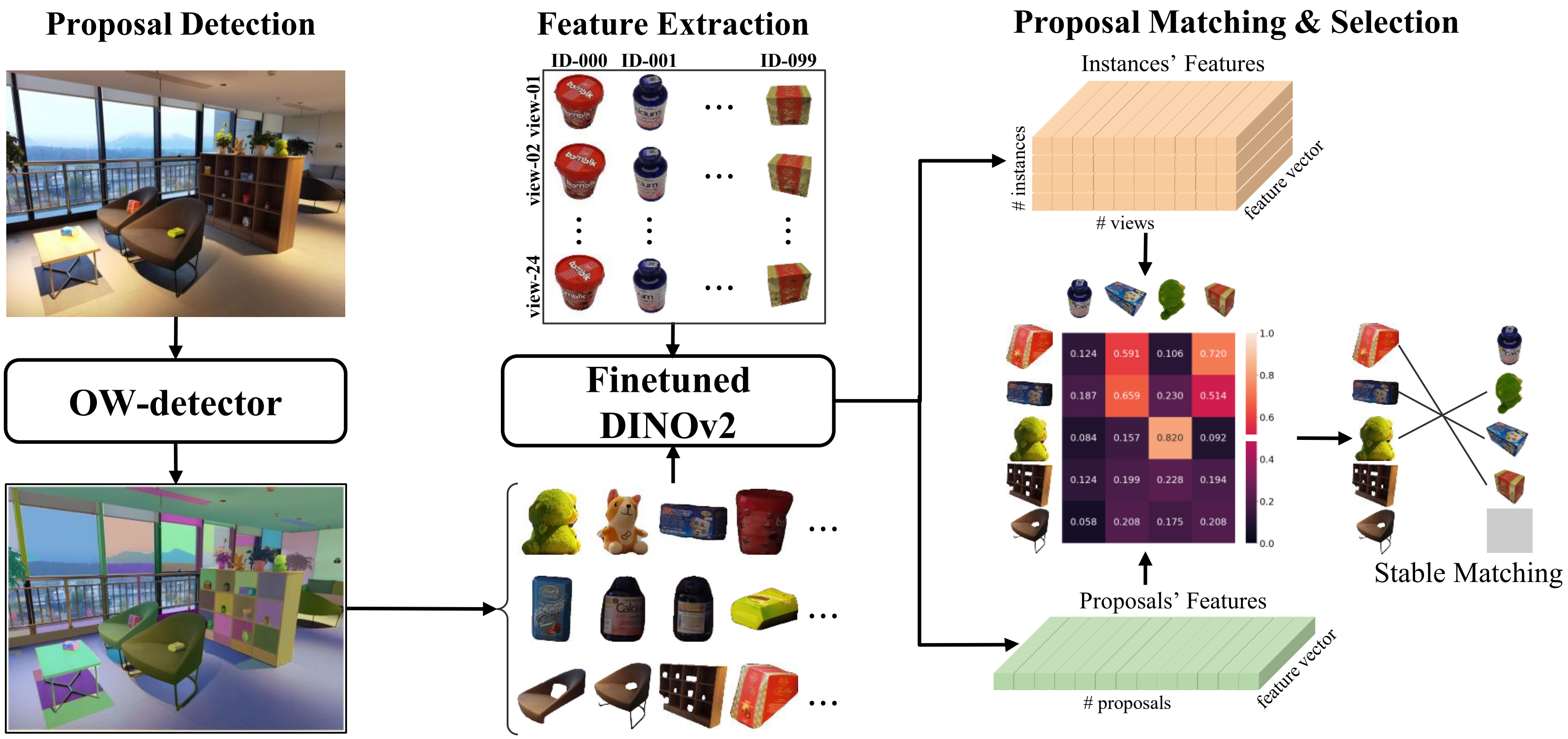




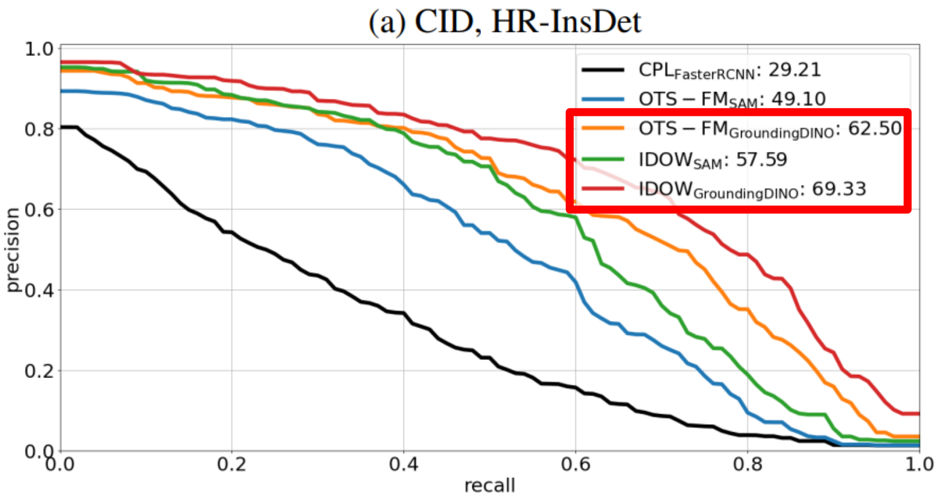
finetune foundation models (e.g., DINOv2) with metric learning

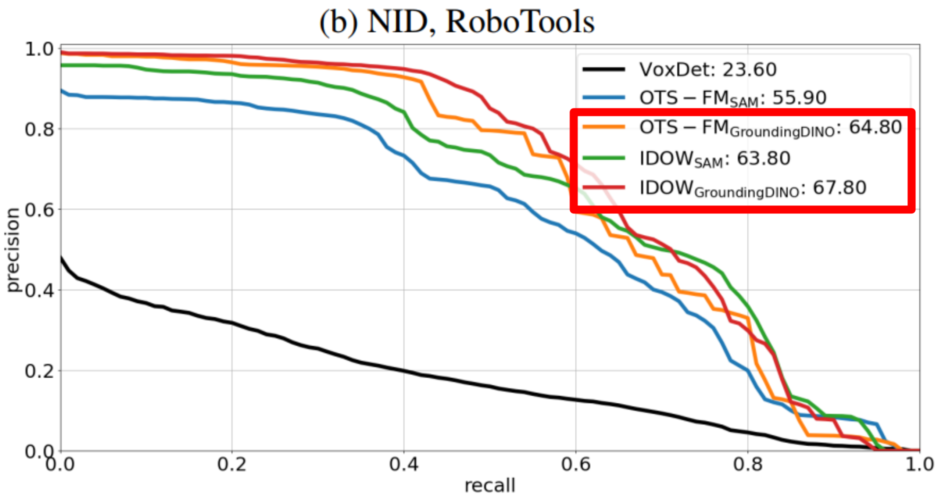
Oquab et al, "DINOv2: Learning Robust Visual Features without Supervision", Arxiv, 2023.

IDOW: Solving InsDet from an Open-World Perspective



Benchmarking results



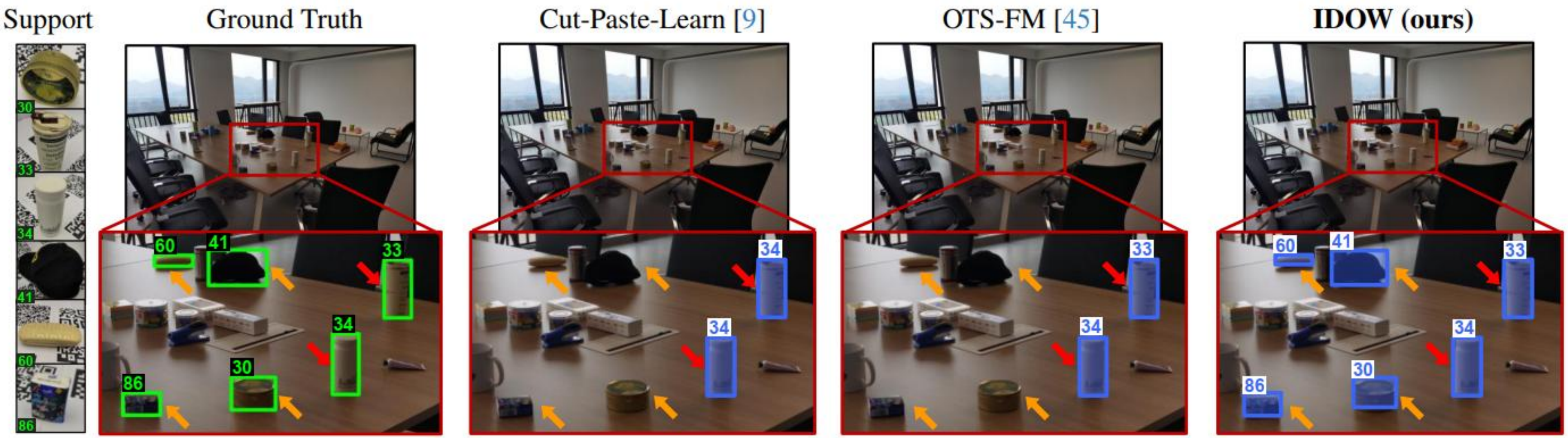


- Our IDOW significantly outperforms the compared methods in both CID and NID settings.
- Using stronger open-world detector improves InsDet performance, cf. GroundingDINO vs. SAM.
- Using stronger features improves InsDet performance, cf. finetuned DINOv2 vs. OTS-FM.

Dwibed & Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection", ICCV, 2017.
Li et al. "VoxDet: Voxel Learning for Novel Instance Detection", NeurIPS, 2023.
Shen et al. "A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture", NeurIPS, 2023.

Qualitative evaluations

HR-InsDet in the CID setting

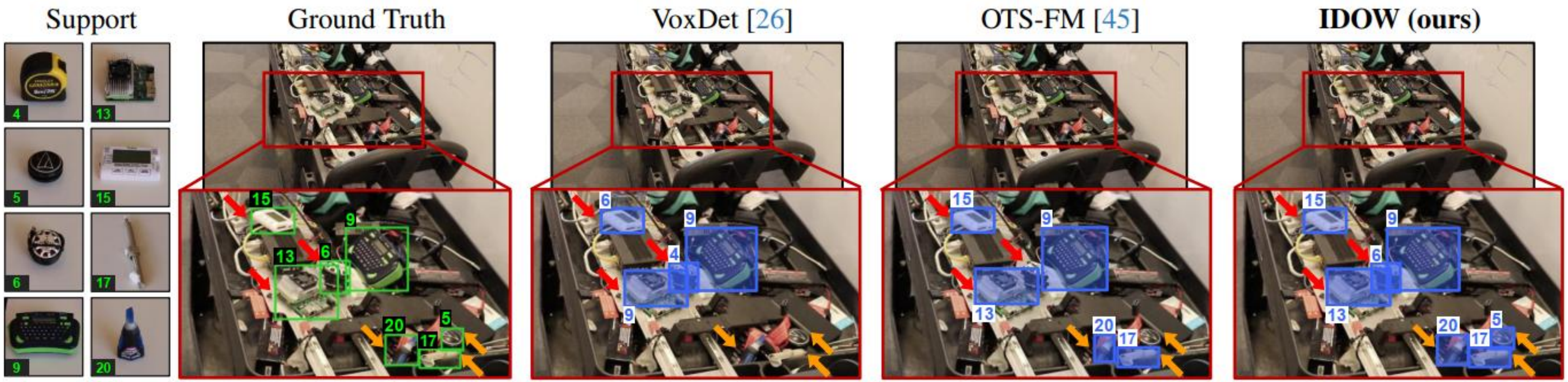


Our IDOW significantly outperforms the compared methods in both CID and NID settings.

Dwibed & Hebert, “Cut, paste and learn: Surprisingly easy synthesis for instance detection”, ICCV, 2017.
Shen et al. “A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture”, NeurIPS, 2023.

Qualitative evaluations

RoboTools in the NID setting



Our IDOW significantly outperforms the compared methods in both CID and NID settings.

Li et al. “VoxDet: Voxel Learning for Novel Instance Detection”, NeurIPS, 2023.
Shen et al. “A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture”, NeurIPS, 2023.

Thank You!

ExHall D Poster #431

https://shenqq377.github.io/IDOW