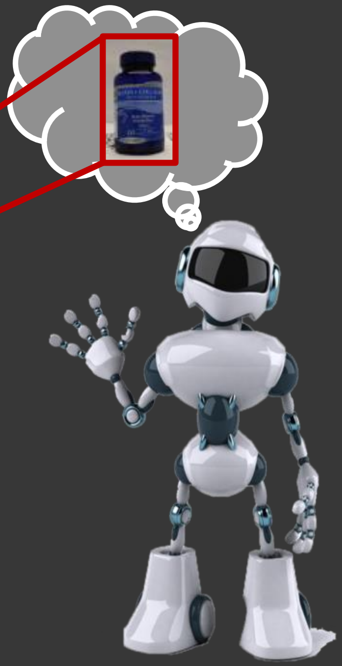
Instance Detection and Tracking in the Open World

The 1st workshop on instance detection at ACCV 2024



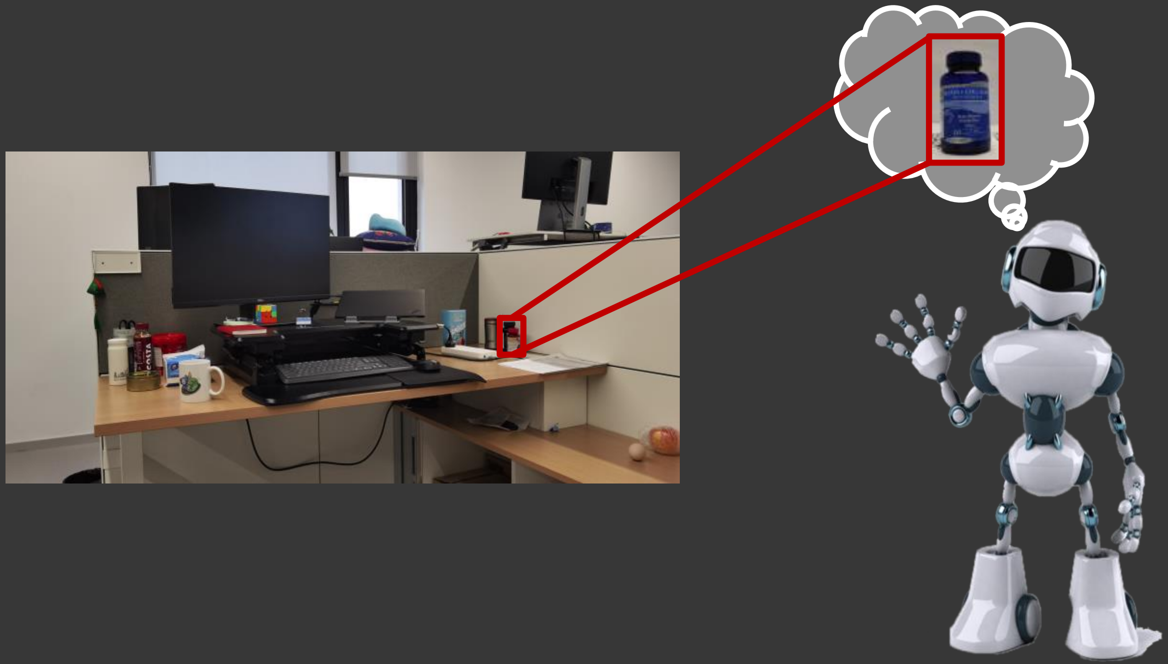
Shu Kong

University of Macau

December 9, 2024

Instance Detection

- It aims to localize the “wanted” object in distance.
- It is usually a prerequisite step in vision systems
- It is useful in robotics, AR/VR, etc.



Instance Detection

- It aims to localize the “wanted” object in distance.
- It is usually a prerequisite step in vision systems
- It is useful in robotics, AR/VR, etc.

Where is my “*screw driver*”?

Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024.

Instance Detection

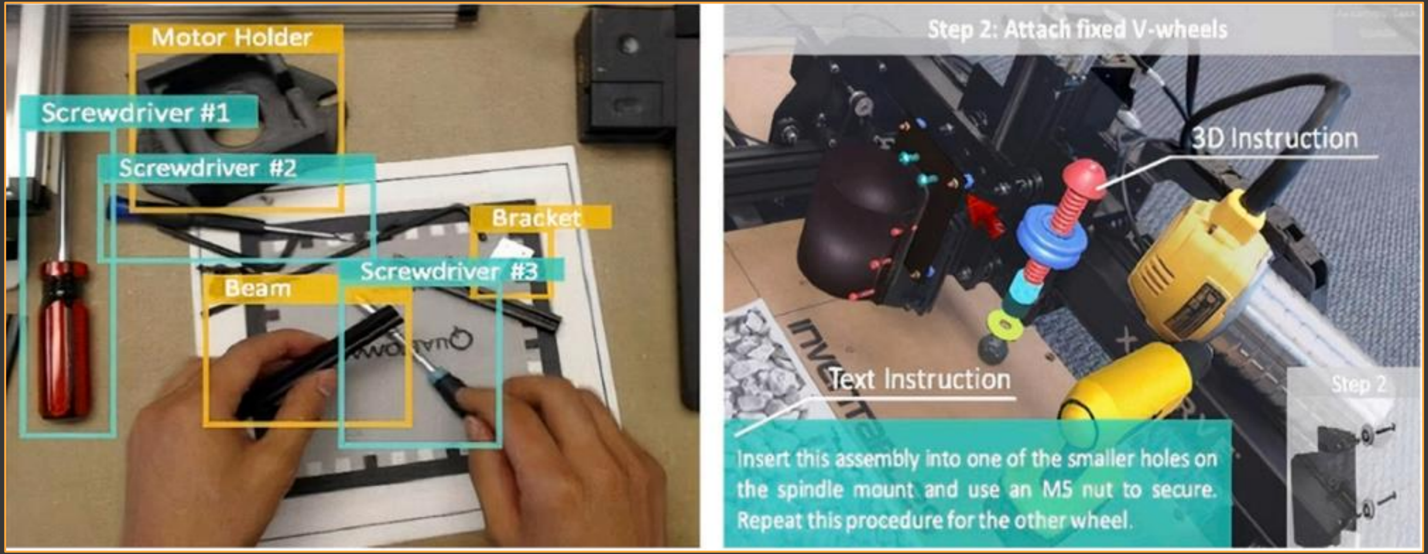
- It aims to localize the “wanted” object in distance.
- It is usually a prerequisite step in vision systems
- It is useful in robotics, AR/VR, etc.

Hi robot, get “my coffee mug” to me!

Loghmani et al., “Recognizing Objects In-the-wild: Where Do We Stand?”, ICRA 2018

Instance Detection

- It aims to localize the “wanted” object in distance.
- It is usually a prerequisite step in vision systems
- It is useful in robotics, AR/VR, etc.



Well, what to do next?

Tao et al., “A self-aware and active-guiding training & assistant system for worker-centered intelligent manufacturing”, Manufacturing Letters, 2019

Outline

Outline

1. InsDet: problem definition and settings
2. InsDet: the state of the art
3. InsDet in the open world
4. InsTrack in 3D scenes from egocentric videos
5. Remarks

Instance Detection *vs.* Related Problems

- proposal detection
  detect all possible objects agnostic to classes

- object detection
  detecting objects of pre-defined classes

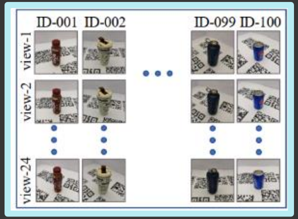
- instance detection
  detecting object instances specified by some visual references

coffee-bean
bottle
cup
…

Two Settings of Instance Detection

- Conventional Instance Detection (CID) / pre-enrollment
  instances are pre-defined that support training;
  applications: AR/VR device helps customers answer "*where is my key?*".

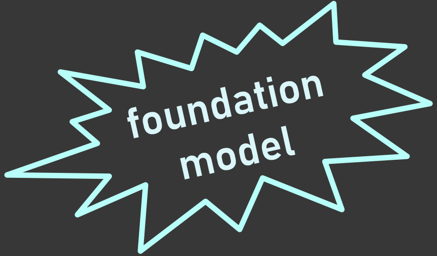
training



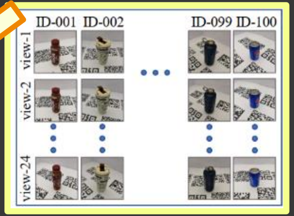
testing

- Novel Instance Detection (NID) / online enrollment
  instances are defined online during testing and the trained detector cannot be finetuned;
  applications: robots search for a novel luggage of a customer at an airport.

training

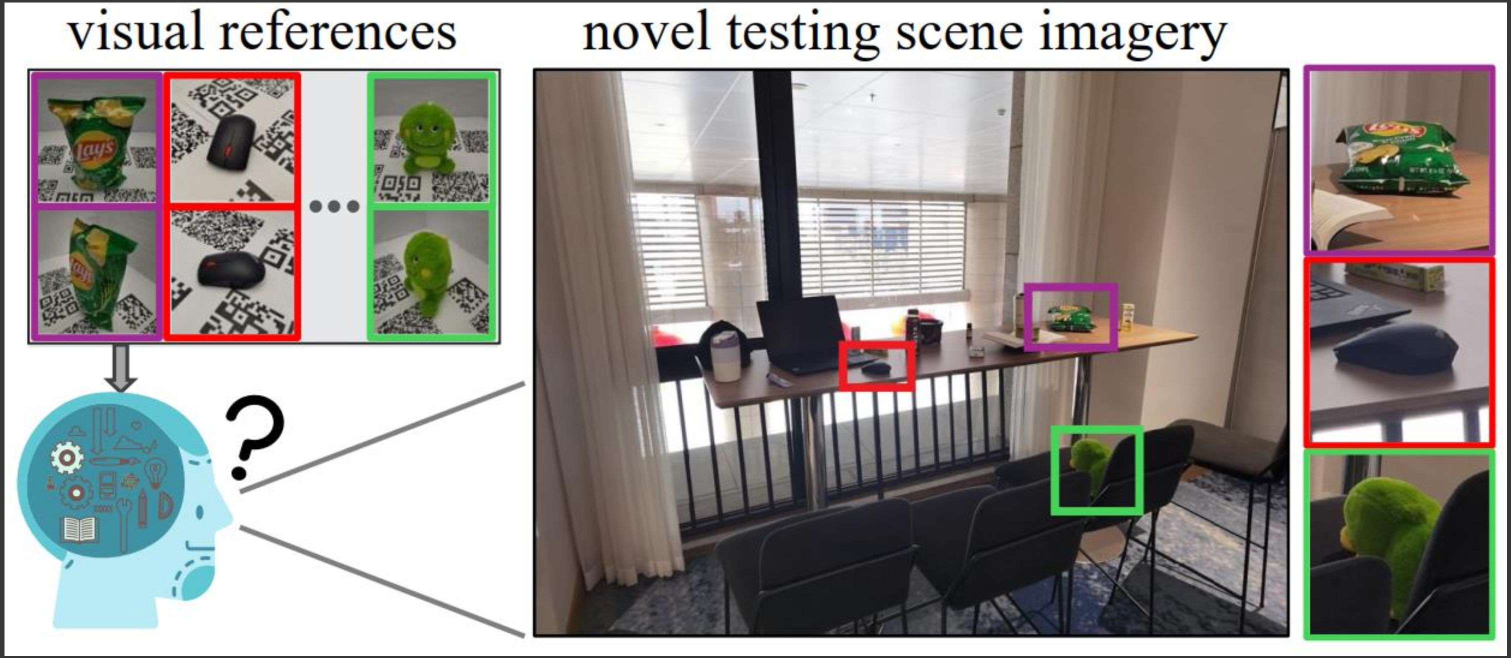


testing



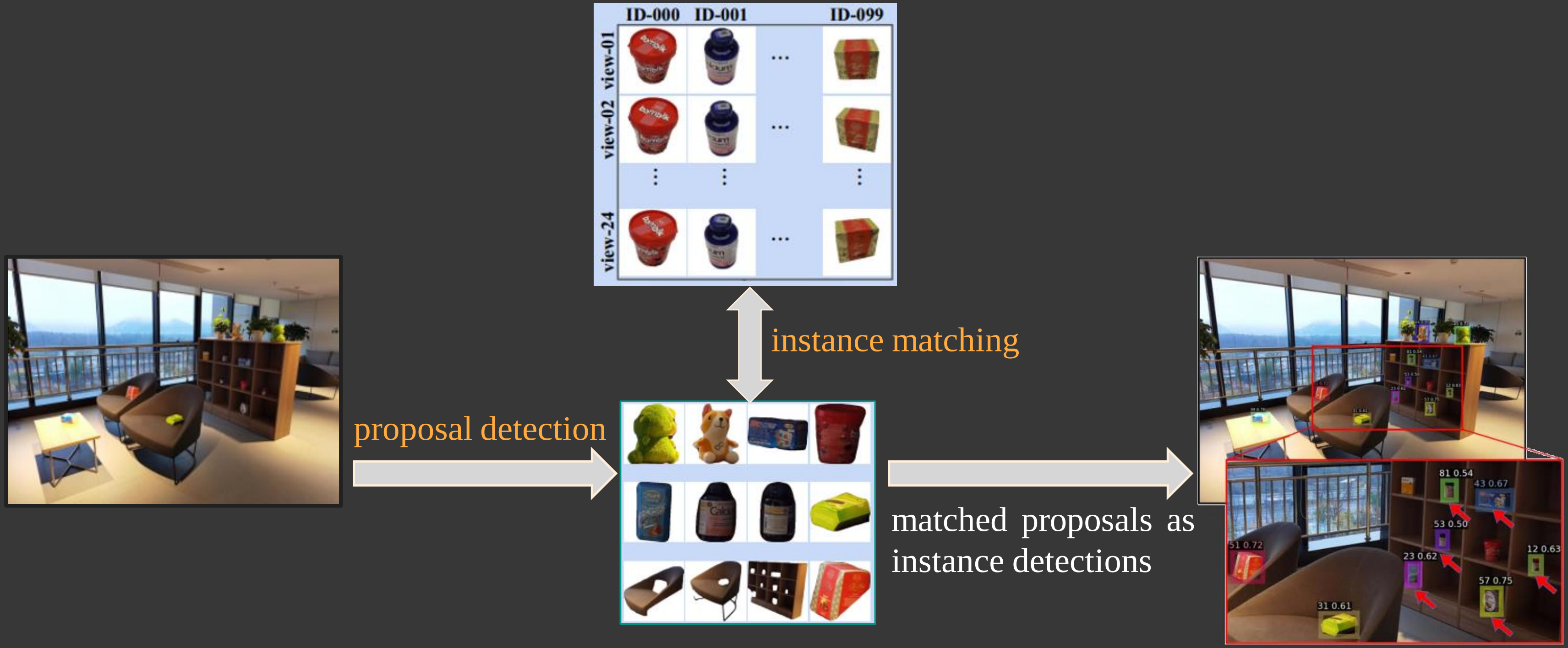
The Open-World Nature of Instance Detection

- Open-set testing imagery is never-before-seen and hence unknown to an instance detector.
- Domain gaps exist between visual references and instance proposals (due to occlusions, lighting variations, etc.).
- Robustness and generalization are desperately needed to detect diverse instances.



A General Framework: Proposal Detection + Instance Matching

- Open-set testing imagery is never-before-seen and hence unknown to an instance detector.
- Domain gaps exist between visual references and instance proposals (due to occlusions, lighting variations, etc.).
- Robustness and generalization are desperately needed to detect diverse instances.

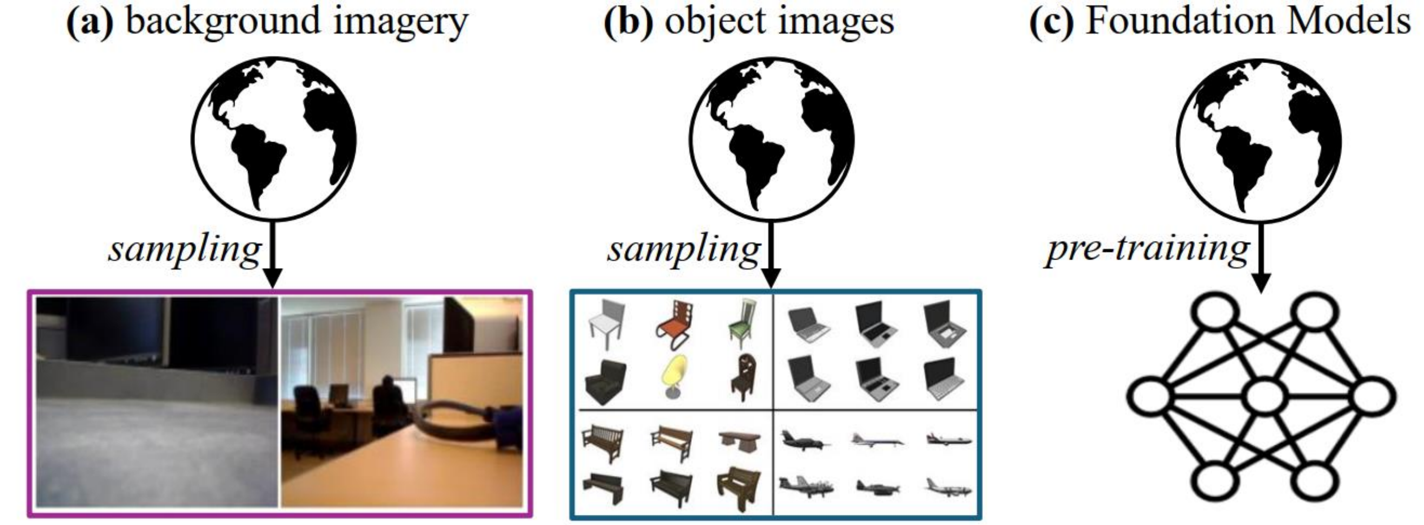


Outline

1. InsDet: problem definition and settings
2. InsDet: the state of the art
3. InsDet in the open world
4. InsTrack in 3D scenes from egocentric videos
5. Remarks

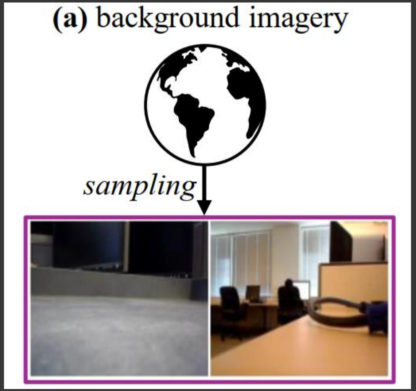
Leveraging the Open World

- Open-set testing imagery is never-before-seen and hence unknown to an instance detector.
- Domain gaps exist between visual references and instance proposals (due to occlusions, lighting variations, etc.).
- Robustness and generalization are desperately needed to detect diverse instances.



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

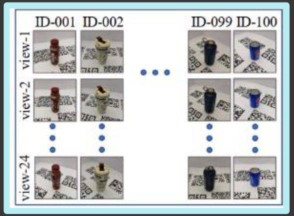
Method 1: Background Sampling



1. Sample background images from the open world
2. Cut the objects from visual references
3. Paste on the sampled background images to generate “free” bounding boxes
4. Learn a detector for the instances of interest

Cut-Paste Learn is a simple and strong baseline of instance detection in the CID setting.

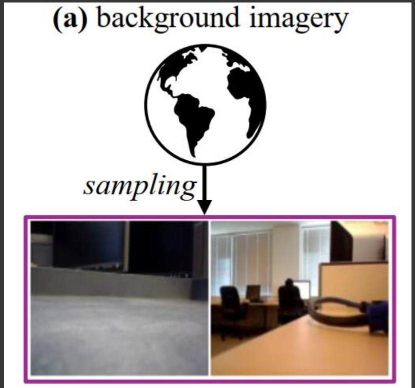
training



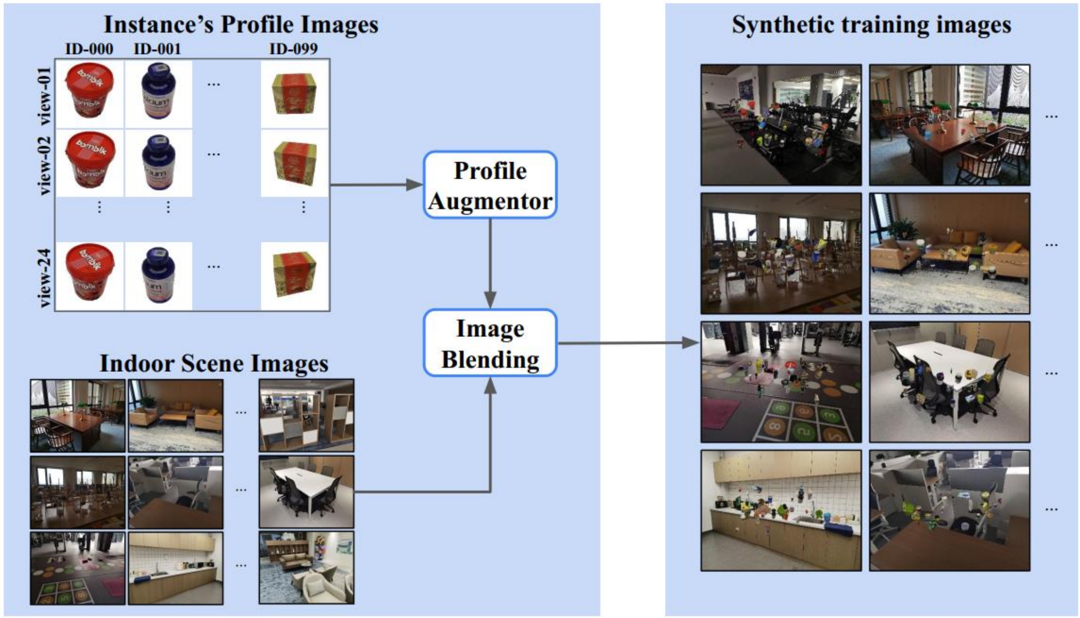
testing

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

Method 1: Background Sampling

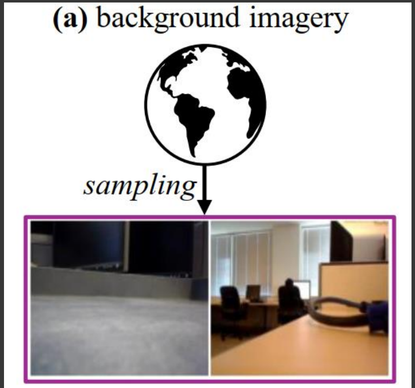


1. Sample background images from the open world
2. Cut the objects from visual references
3. Paste on the sampled background images to generate "free" bounding boxes
4. Learn a detector for the instances of interest

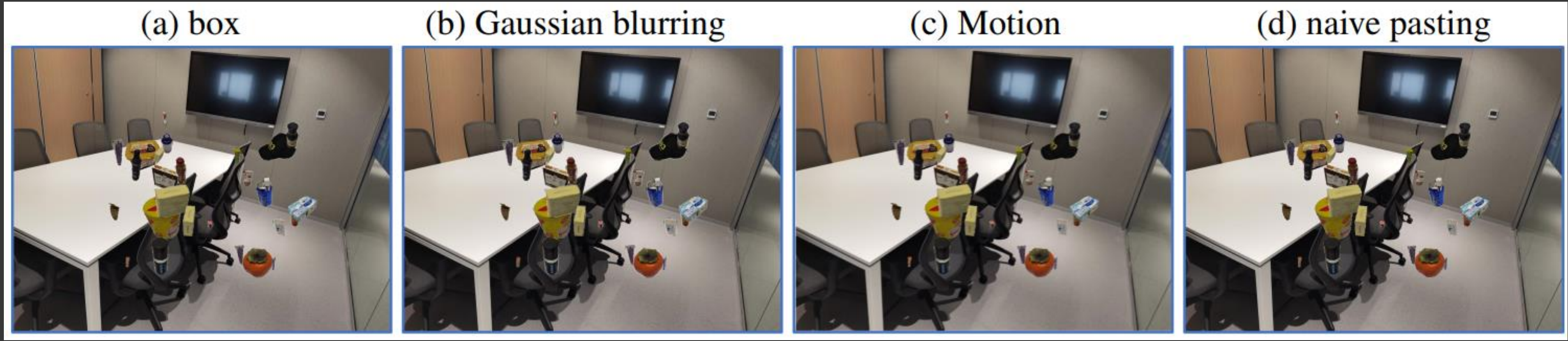


Dwibed & Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection", ICCV, 2017

Method 1: Background Sampling

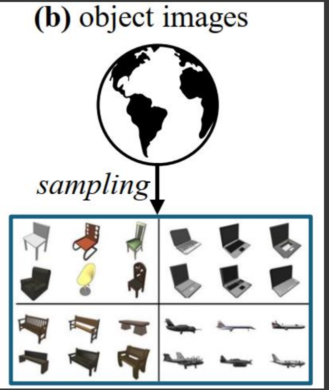


1. Sample background images from the open world
2. Cut the objects from visual references
3. Paste on the sampled background images to generate "free" bounding boxes
4. Learn a detector for the instances of interest

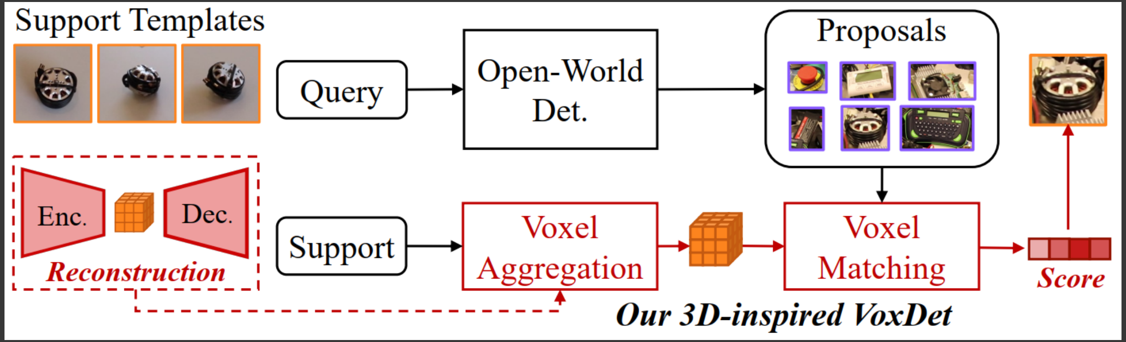


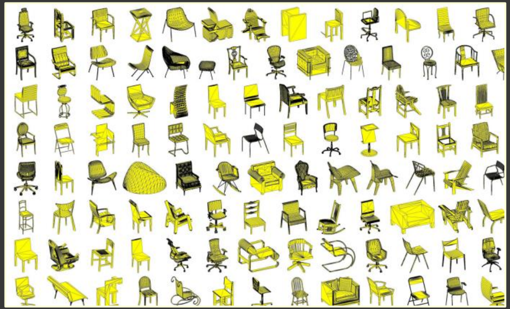
Dwibed & Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection", ICCV, 2017

Method 2: Object Sampling



1. Sample multi-view object instances images from the open world
2. Learn a function for reference-proposal matching
3. Use an open-world detector to detect proposals (i.e., all possible instances)



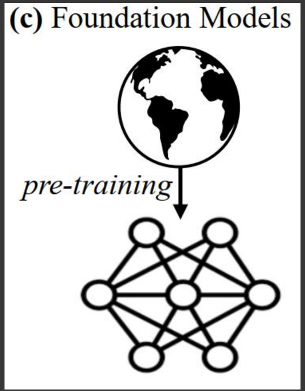


ShapeNet dataset

ABO dataset

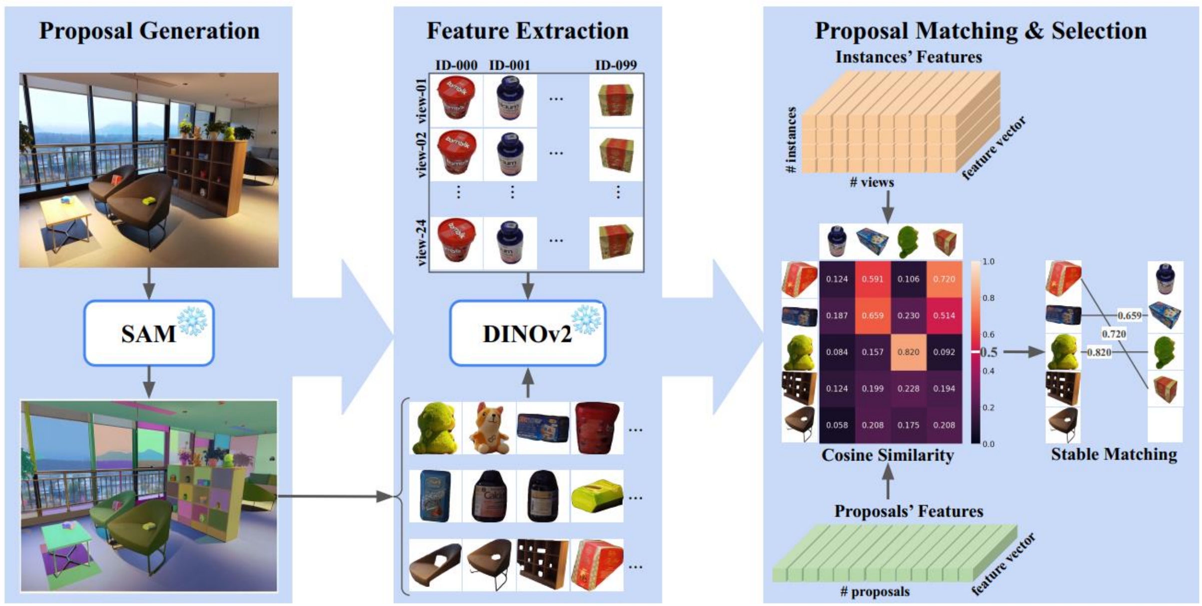
Li et al. "VoxDet: Voxel Learning for Novel Instance Detection", NeurIPS, 2023

Method 3: Using Foundation Models



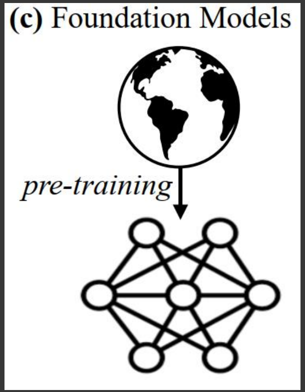
1. Utilize foundation models pretrained in the open world for proposal detection and reference-proposal matching.

2. This is a non-learned method!

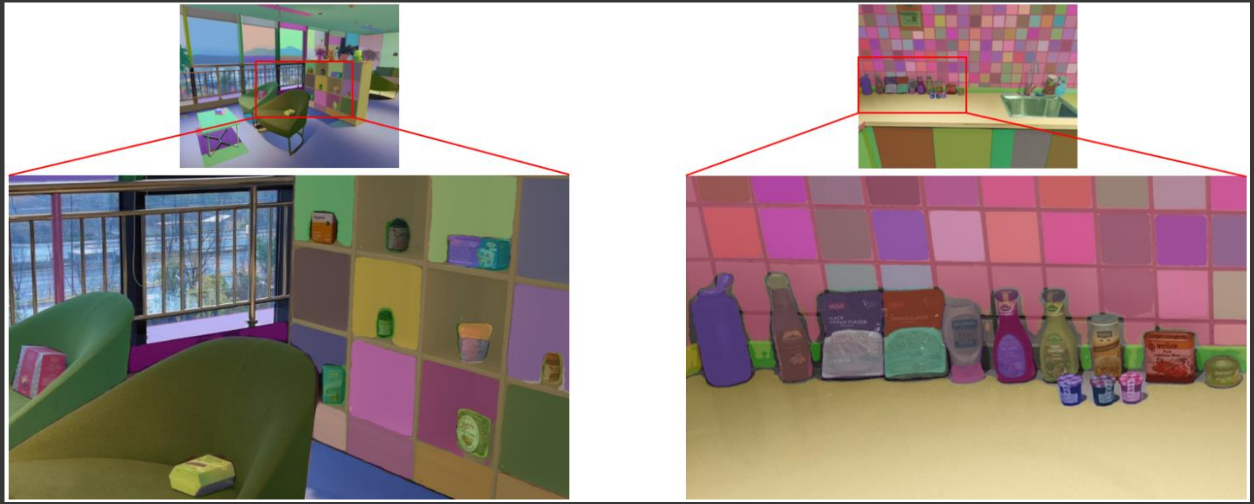


Shen et al. "A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture", NeurIPS, 2023
Kirillov, et al. "Segment anything." ICCV 2023.
Oquab, et al. "Dinov2: Learning robust visual features without supervision." TMLR, 2024.

Method 3: Using Foundation Models



1. Utilize foundation models pretrained in the open world for proposal detection and reference-proposal matching.
2. This is a non-learned method!

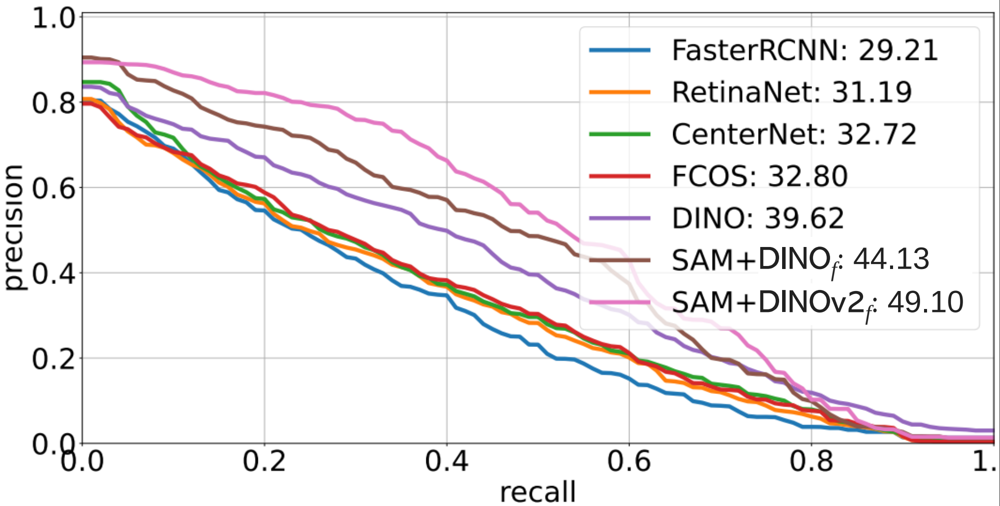


Proposal detection/segmentation by Segment Anything Model (SAM)

Kirillov, et al. "Segment anything." ICCV 2023.
Shen et al. “A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture”, NeurIPS, 2023

Method 3: Using Foundation Models

1. The foundation model SAM yields sufficiently high recall.
2. By using foundation models, the non-learned method significantly outperforms Cut-Paste-Learn (built on FasterRCNN and DINO).
3. Using better features (DINOv2$_f$ vs. DINO$_f$) improves instance detection.

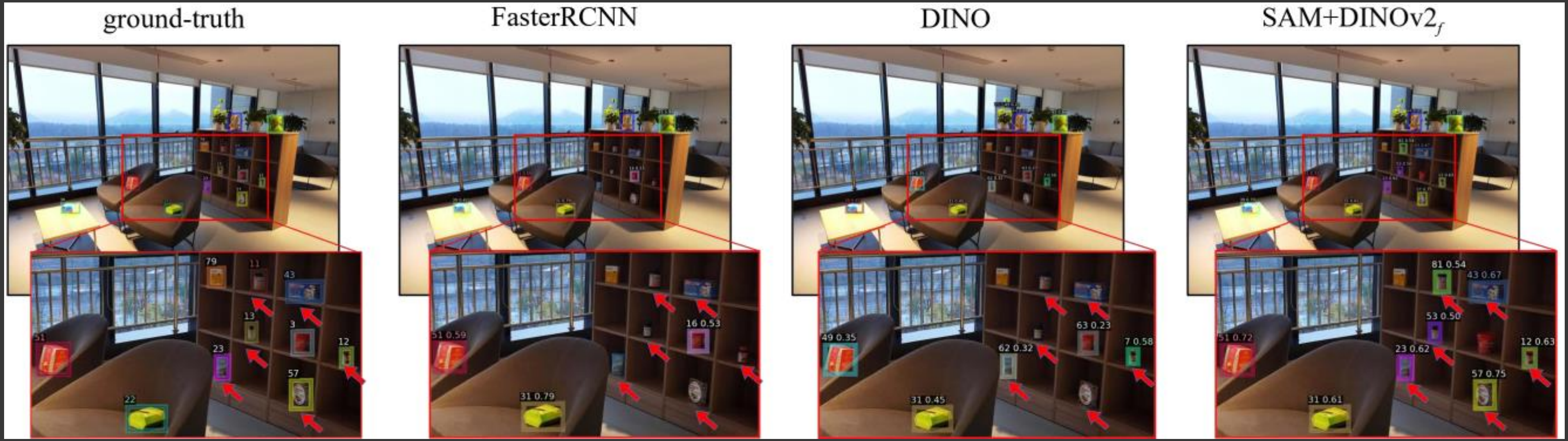


Benchmarking results on the HR-Insdet benchmark dataset in the CID setting.

Shen et al. “A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture”, NeurIPS, 2023

Method 3: Using Foundation Models

1. The foundation model SAM yields sufficiently high recall.
2. By using foundation models, the non-learned method significantly outperforms Cut-Paste-Learn (built on FasterRCNN and DINO).
3. Using better features (DINOv2_f vs. DINO_f) improves instance detection.



Visual results by Cut-Paste-Learn and our non-learned SAM+DINOv2_f

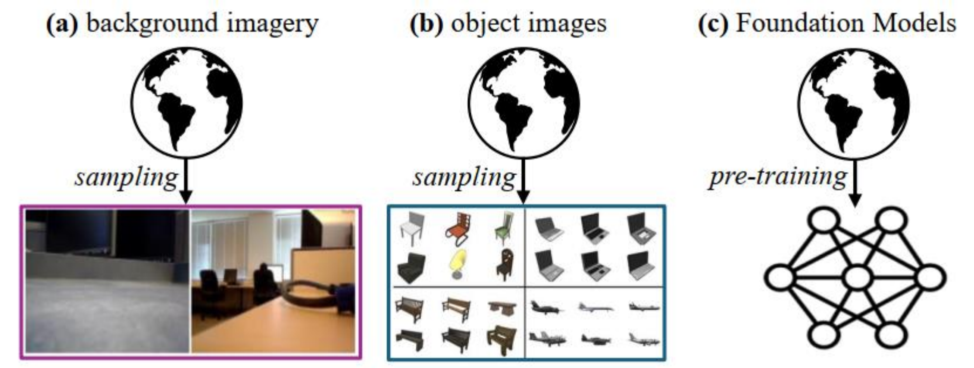
Shen et al. "A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture", NeurIPS, 2023

Outline

1. InsDet: problem definition and settings
2. InsDet: the state of the art
3. InsDet in the open world
4. InsTrack in 3D scenes from egocentric videos
5. Remarks

Solving Instance Detection Fully from an Open-World Perspective

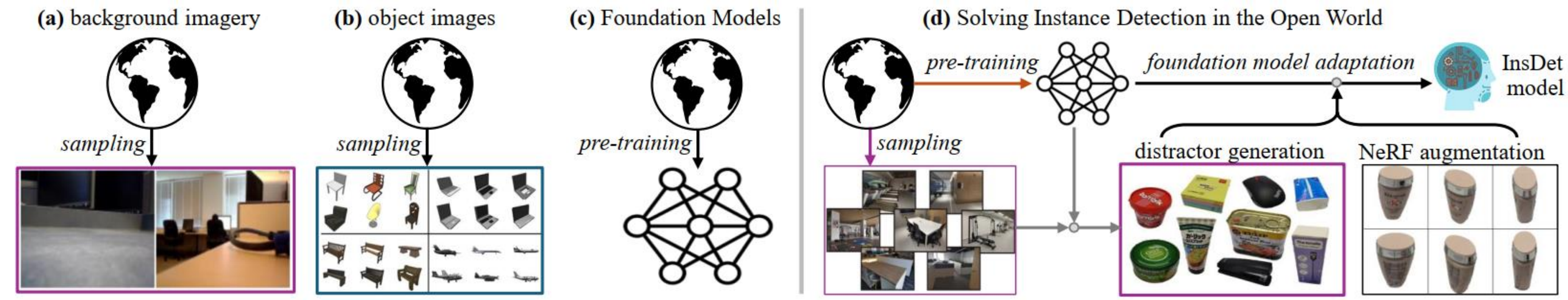
- Open-set testing imagery is never-before-seen and hence unknown to an instance detector.
- Domain gaps exist between visual references and instance proposals (due to occlusions, lighting variations, etc.).
- Robustness and generalization are desperately needed to detect diverse instances.



Shen, et al., "Solving Instance Detection from an Open-World Perspective", arxiv'ing, 2024

Solving Instance Detection Fully from an Open-World Perspective

- Open-set testing imagery is never-before-seen and hence unknown to an instance detector.
- Domain gaps exist between visual references and instance proposals (due to occlusions, lighting variations, etc.).
- Robustness and generalization are desperately needed to detect diverse instances.

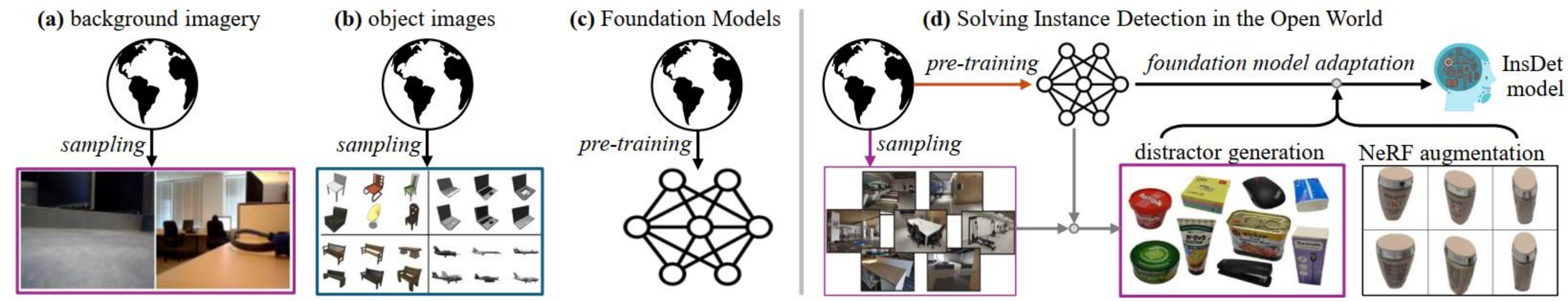


Shen, et al., “Solving Instance Detection from an Open-World Perspective”, arxiv’ing, 2024

Solving Instance Detection Fully from an Open-World Perspective

Thoughts:

- A foundational detector yields high recall, i.e., 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.

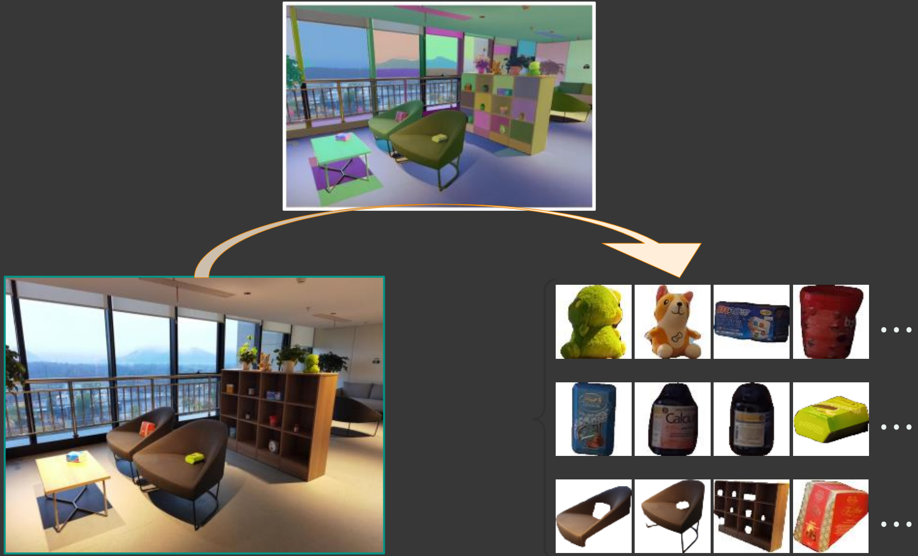


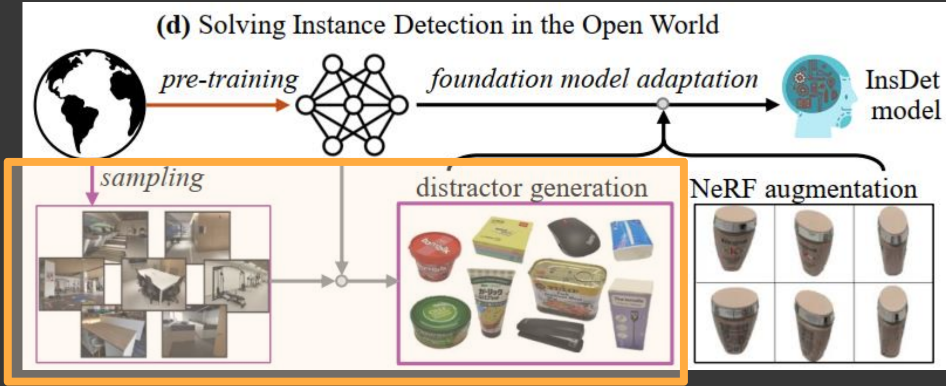
Shen, et al., "Solving Instance Detection from an Open-World Perspective", arxiv'ing, 2024

Sampling Distractor Instance from Real Imagery

Thoughts:

- A foundational detector yields high recall, i.e., 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.



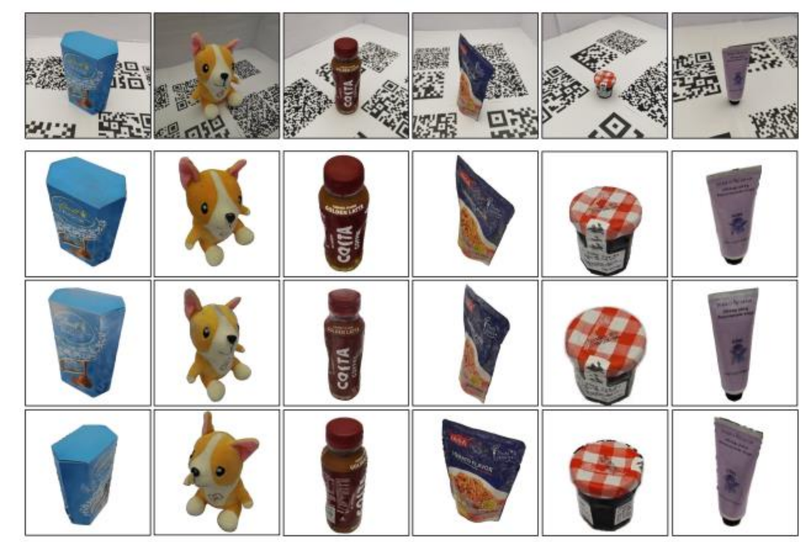


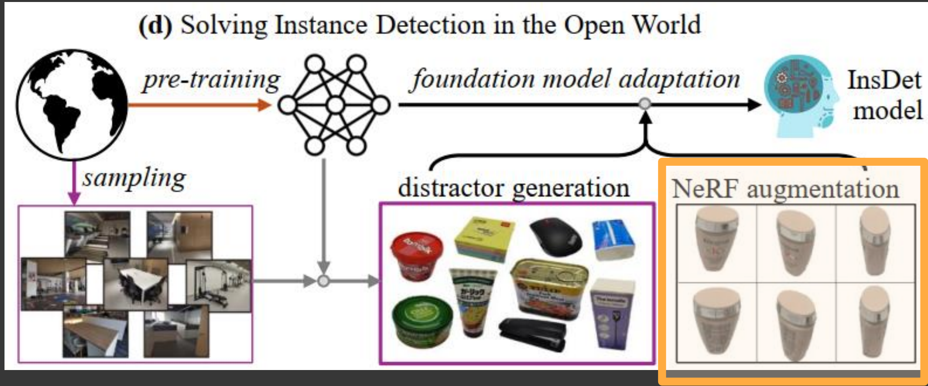
Shen, et al., "Solving Instance Detection from an Open-World Perspective", arxiv'ing, 2024

Sampling More Positive Instances using NeRF

Thoughts:

- A foundational detector yields high recall, i.e., 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.





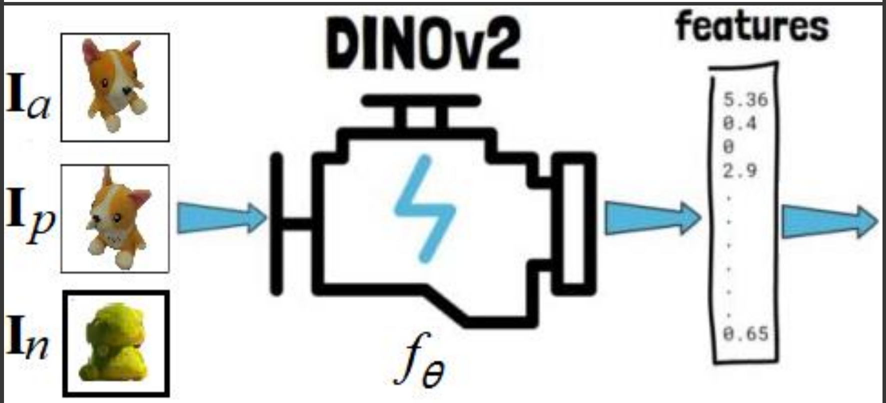
Ben et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". ECCV 2020
Shen, et al., "Solving Instance Detection from an Open-World Perspective", arxiv'ing, 2024

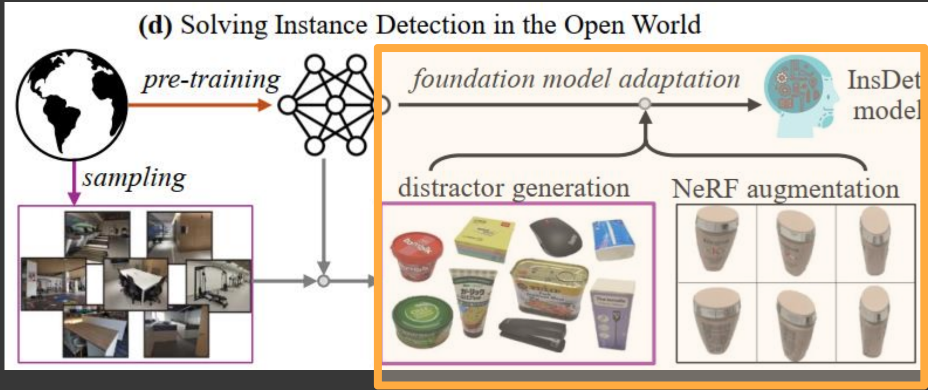
Adapting DINOv2 using Metric Learning

Thoughts:

- A foundational detector yields high recall, i.e., 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.

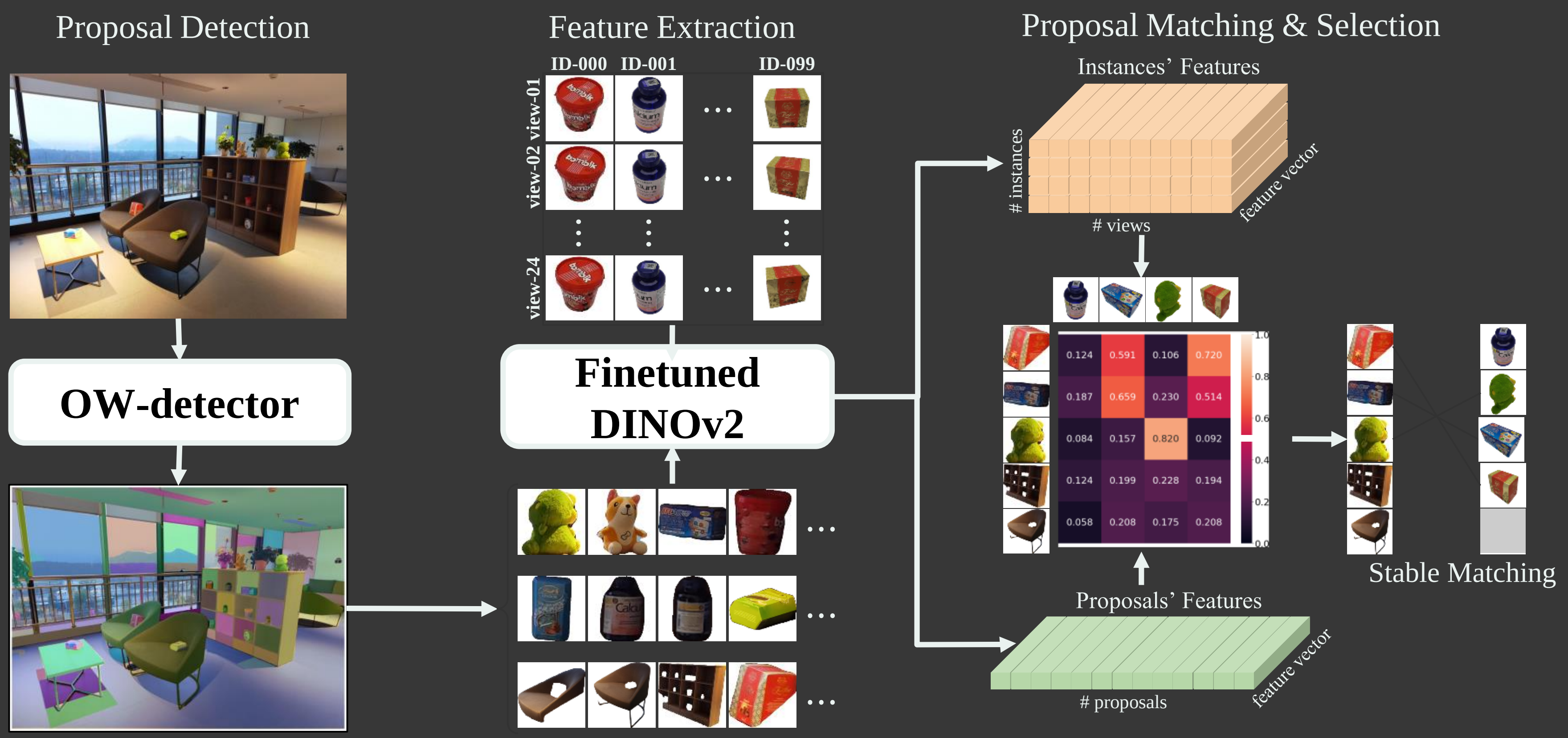
$$\ell = \left[d(f_\theta(\mathbf{I}_a), f_\theta(\mathbf{I}_p)) - d(f_\theta(\mathbf{I}_a), f_\theta(\mathbf{I}_n)) + \alpha \right]_+$$





Shen, et al., "Solving Instance Detection from an Open-World Perspective", arxiv'ing, 2024

IDOW: Solving InsDet from an Open-World Perspective

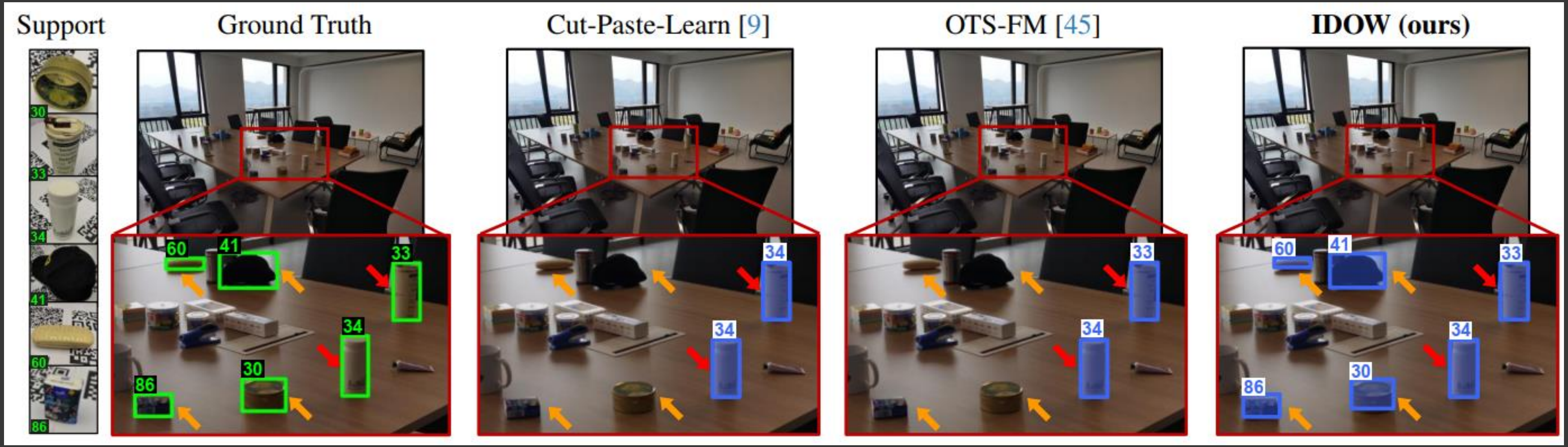


Shen, et al., "Solving Instance Detection from an Open-World Perspective", arxiv'ing, 2024

Results

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

Results on HR-InsDet in the CID setting

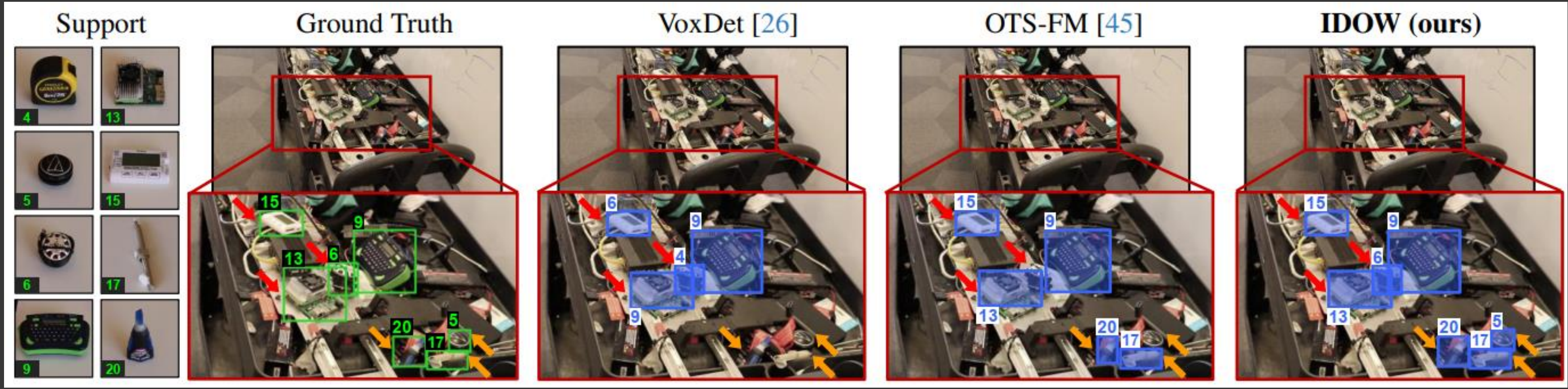


Dwibed & Hebert, “Cut, paste and learn: Surprisingly easy synthesis for instance detection”, ICCV, 2017
Shen et al. “A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture”, NeurIPS, 2023
Shen, et al., “Solving Instance Detection from an Open-World Perspective”, arxiv’ing, 2024

Results

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

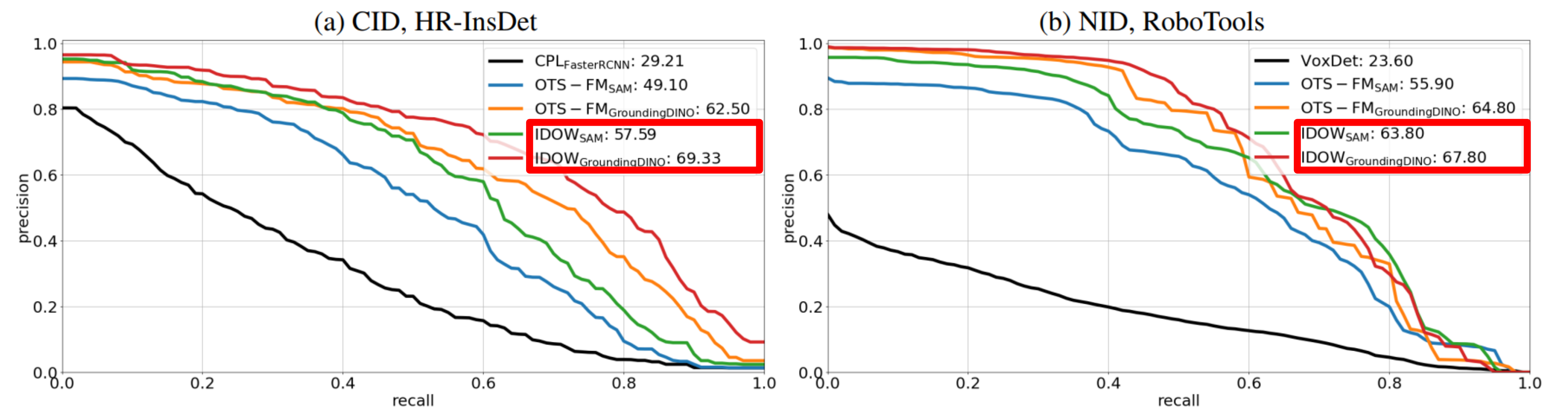
Results on RoboTools in the NID setting



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
Shen, et al., "Solving Instance Detection from an Open-World Perspective", arxiv'ing, 2024

Results

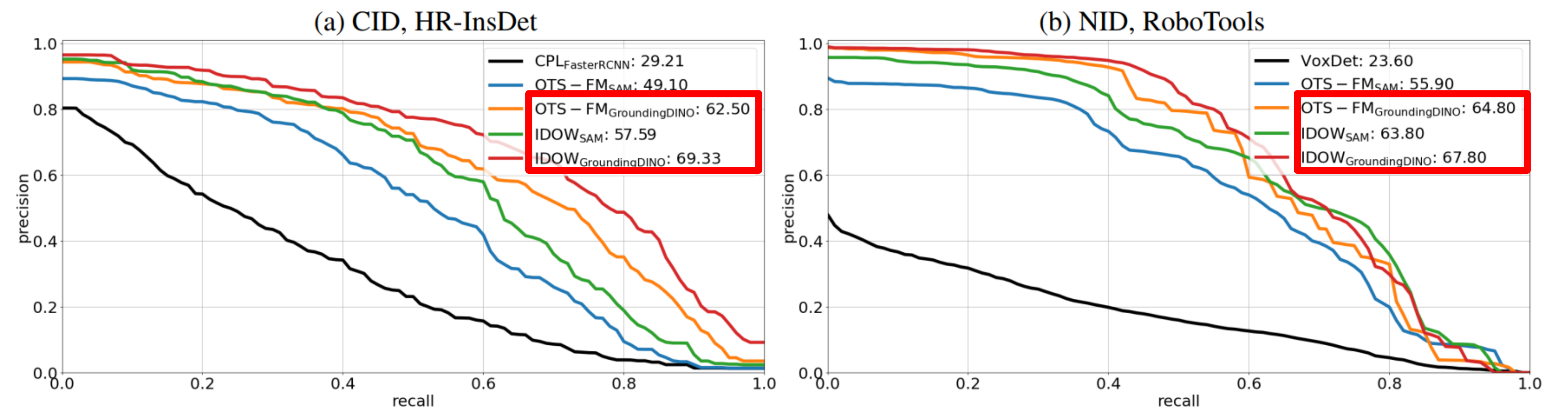
- Our IDOW significantly outperforms the compared methods in both CID and NID settings.
- Using stronger open-world detector improves InsDet performance, cf. GroundingDINO vs. SAM.



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
Shen, et al., "Solving Instance Detection from an Open-World Perspective", arxiv'ing, 2024

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.



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
Shen, et al., “Solving Instance Detection from an Open-World Perspective”, arxiv’ing, 2024

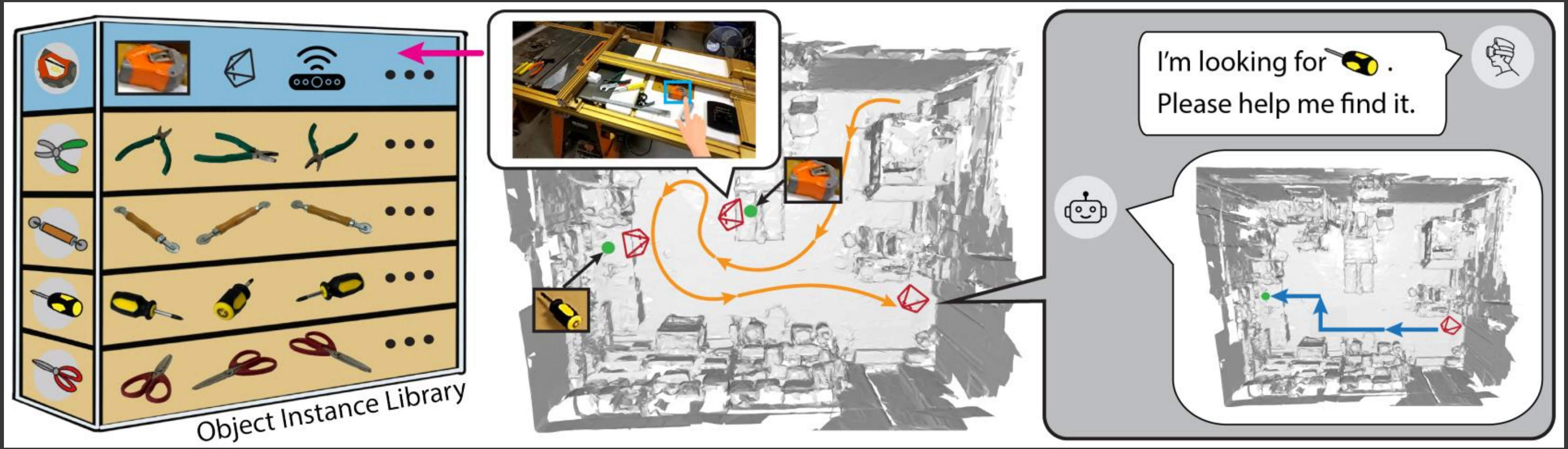
Outline

1. InsDet: problem definition and settings
2. InsDet: the state of the art
3. InsDet in the open world
4. InsTrack in 3D scenes from egocentric videos
5. Remarks

Instance Tracking in 3D from Egocentric Video

Motivation

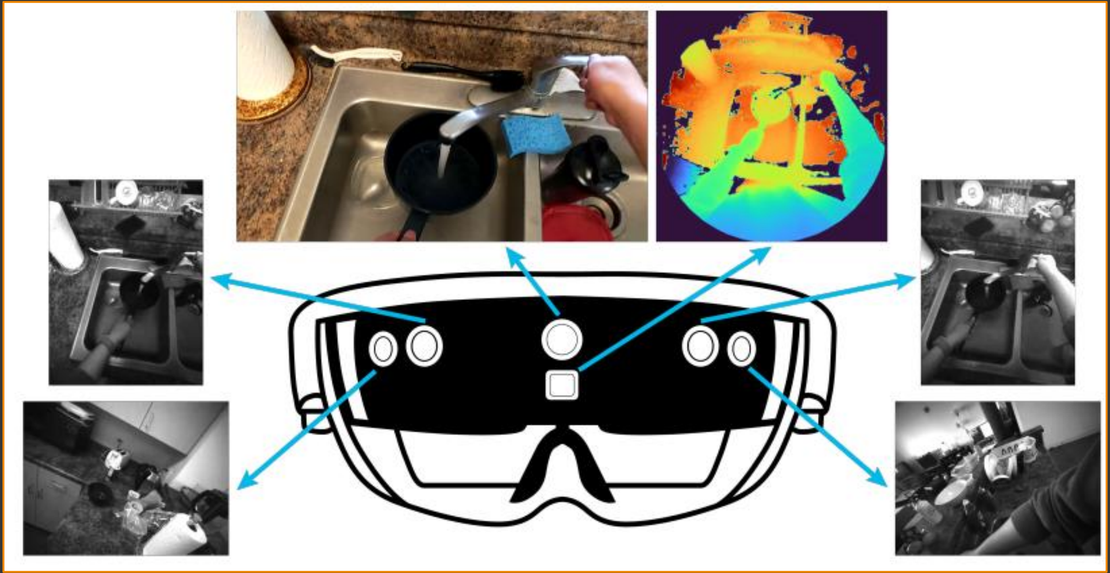
- Developing an AI assistant running on AR/VR devices.
- Guiding users to recall the 3D locations of objects of interest (“where is my key?”).

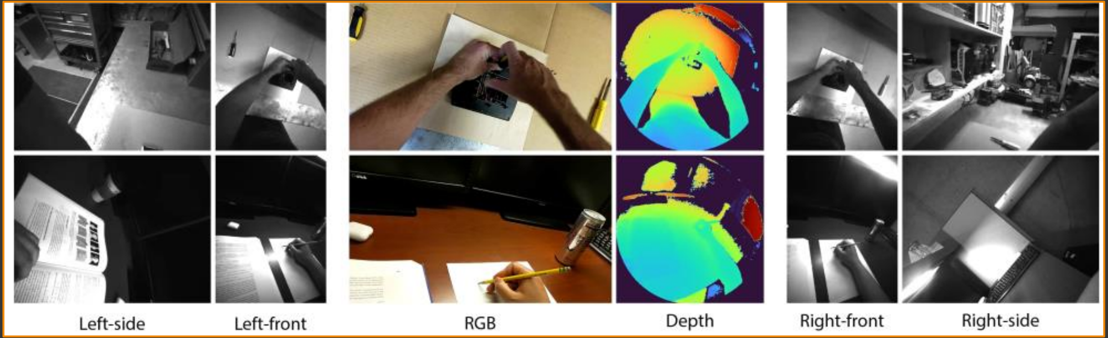


Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024.

Instance Tracking in 3D from Egocentric Video

Device: HoloLens 2

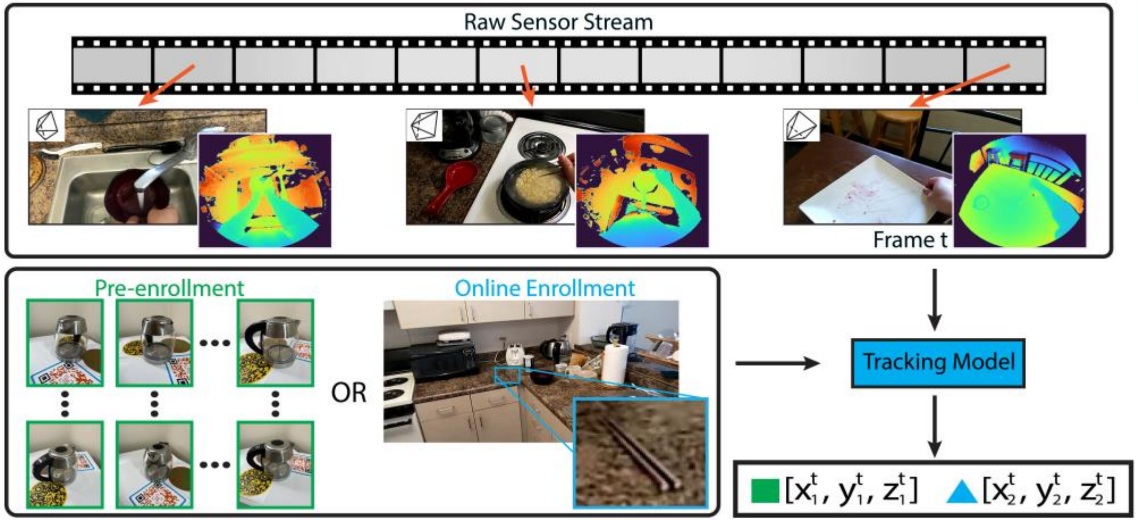




Instance Tracking in 3D from Egocentric Video

Problem definition

- Given a video sequence, tracking instances of interest (i.e., being *enrolled*) in the 3D world coordinate system.
- Assumption: objects remain stationary unless being interacted with.



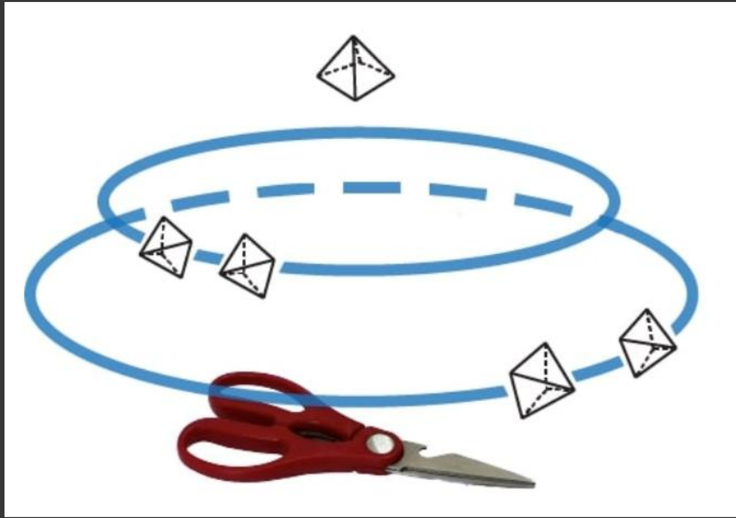
Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024.

Instance Tracking in 3D from Egocentric Video

Settings

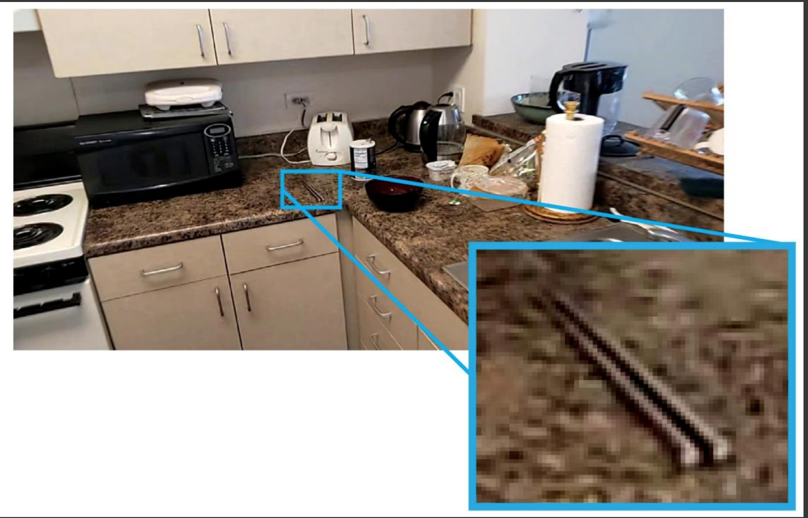
Multi-View Pre-Enrollment (MVPE)

Pre-enroll objects with multiple visual references.



Single-View Online Enrollment (SVOE)

Enroll on-the-fly by the user, e.g., pointing to specify

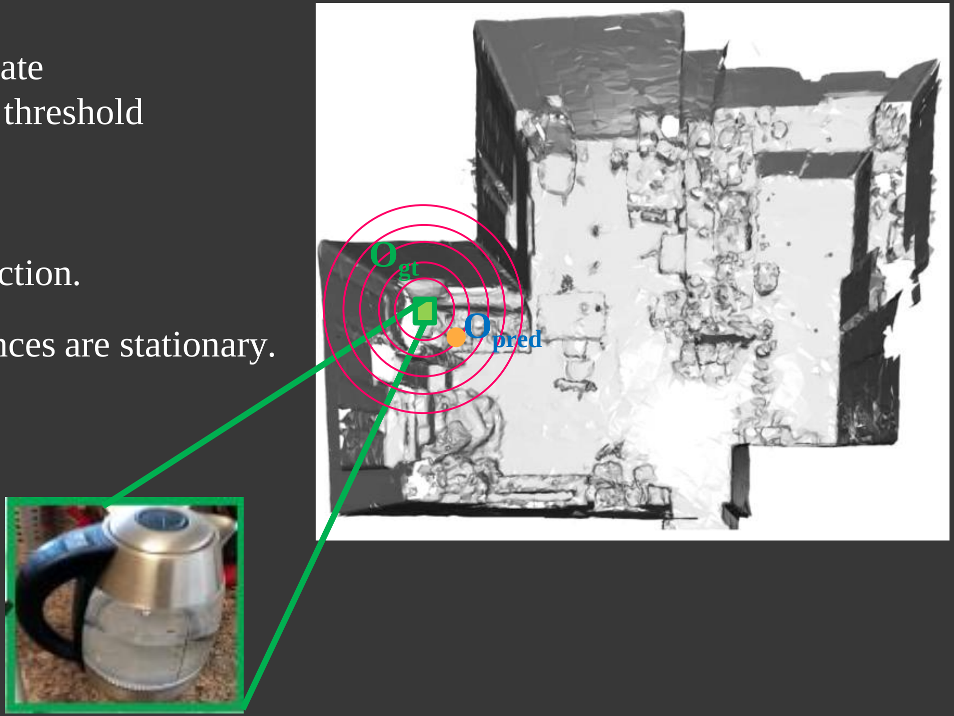


Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024.

Instance Tracking in 3D from Egocentric Video

Evaluation metrics

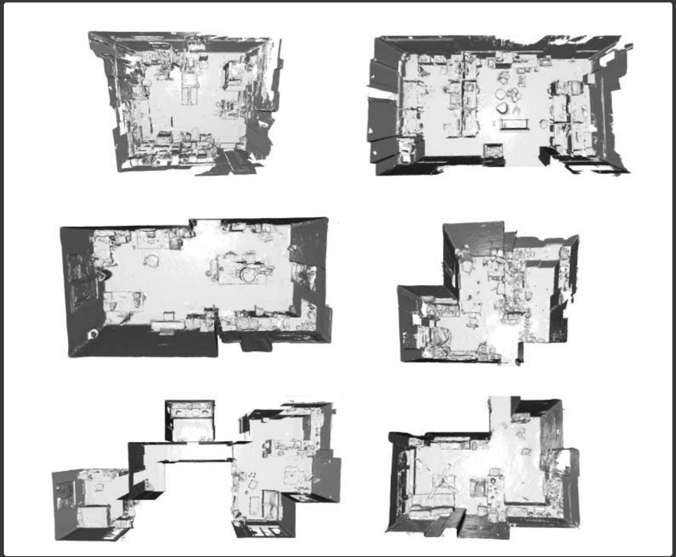
- precision and recall of detection in 3D world coordinate
  - True positive (TP) is defined as: $|O_{pred} - O_{gt}| \leq$ threshold
  - Precision = TP / (TP+FP)
  - Recall = TP / (TP+FN)
- L2 and angular error between ground-truth and prediction.
- We evaluate in time intervals where concerned instances are stationary.

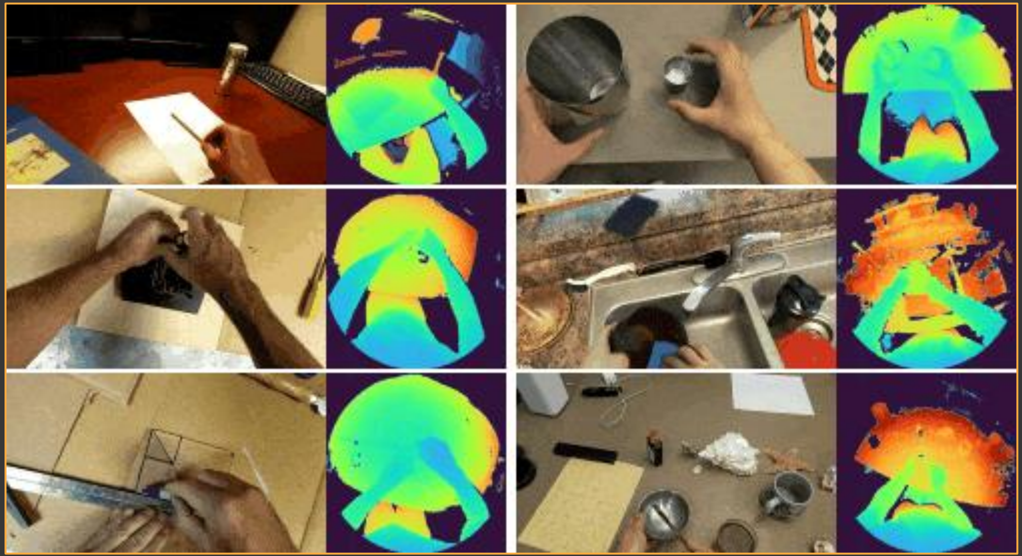


Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024.

Data Collection

- Videos of daily activities captured with a Hololens 2.
- 50 videos (30 fps, ≥5min).
- 10 different indoor scenes with natural camera trajectories.

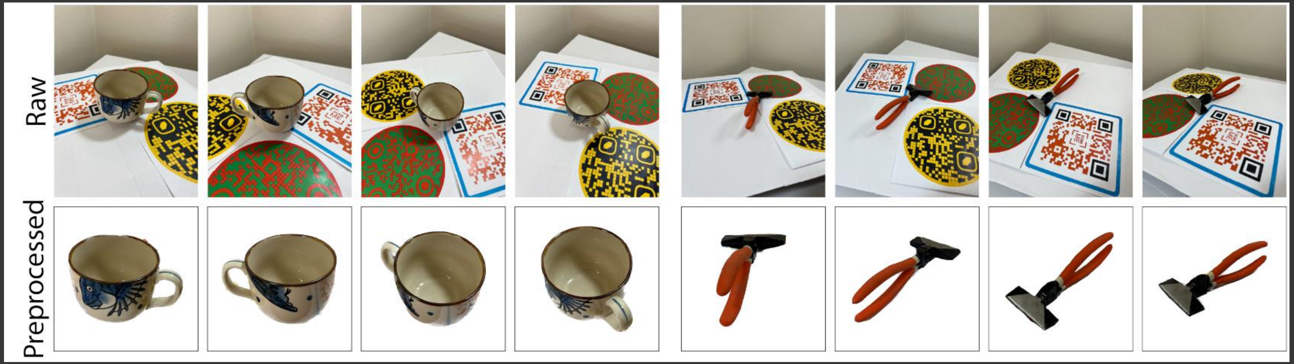




Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024.

Data Annotation

- Object instance 3D center
  - 3D positions of object instance center in the 3D world coordinate frame.
- 2D bounding box annotations
  - Axis-aligned *amodal* 2D bounding boxes.
- Object motion state annotations
  - Binary annotation, either stationary or dynamic (being interacted with).



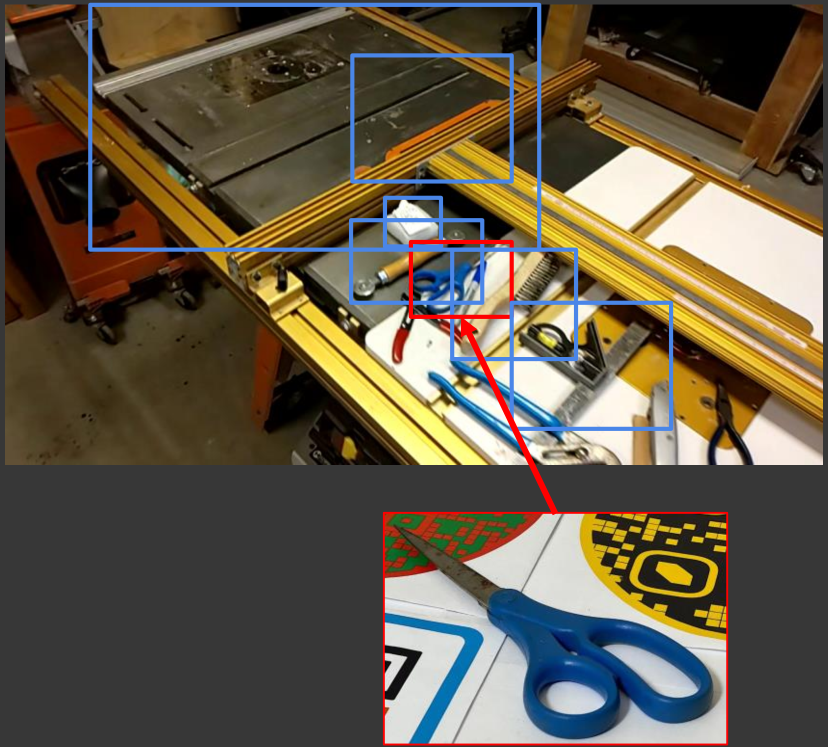
Motion state: dynamic

Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024.

Method

Our method is similar to instance detection

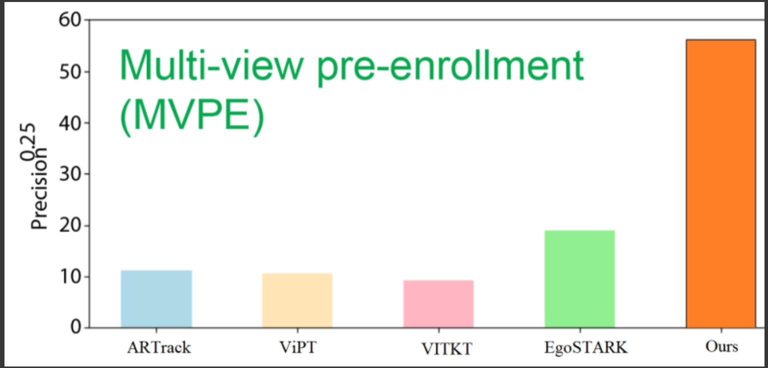
- Leverage foundation models SAM and DINOv2 for proposal detection and instance matching, respectively;
- Using depth camera to project 2D detections to 3D world;
- Record 3D coordinates of detected instances (when confidence is high and they are static).

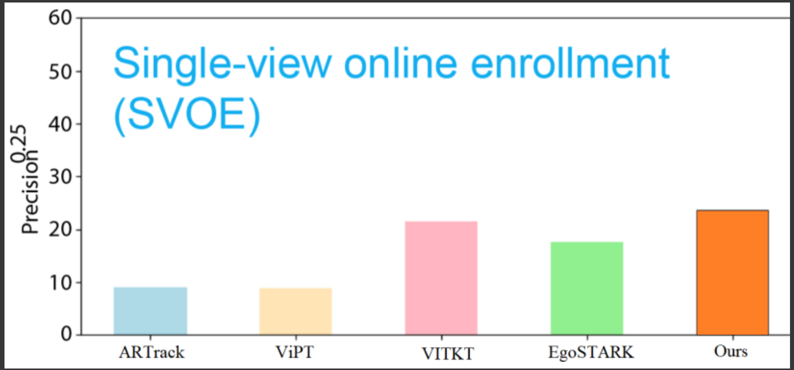


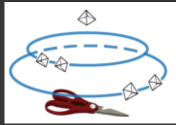
[SAM] Kirillov, et al. "Segment anything." ICCV 2023.
[DINOv2] Oquab, et al. "Dinov2: Learning robust visual features without supervision." TMLR, 2024.

Results

- Compared against state-of-the-art single object trackers, our non-learned method "SAM+DINOv2" performs the best.
- The problem of Instance Tracking in 3D from Egocentric Video is made much easier by leveraging camera pose and using a 3D allocentric (world) coordinate representation.







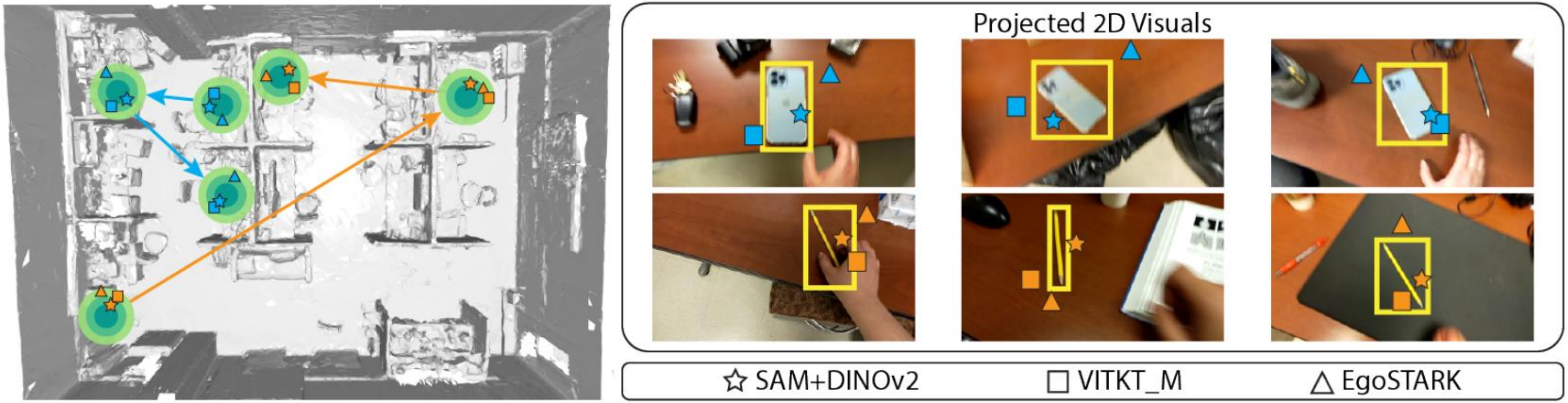
Multi-view pre-enrollment (MVPE)
Pre-enroll objects with multiple visual references.

Single-view online enrollment (SVOE)
Enroll on-the-fly by the user, e.g., pointing to specify.

[ARTrack] Wei, et al., "Autoregressive visual tracking". CVPR, 2023
[ViPT] Zhu, et al., "Visual prompt multi-modal tracking". CVPR, 2023
[VITKT] Kristan, et al., "Visual prompt multi-modal tracking". The tenth visual object tracking vot2022 challenge results, 2022
[EgoSTARK] Tang, et al., "Egotracks: A long-term egocentric visual object tracking dataset". NeurIPS, 2024
[Ours] Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024

Results

- Compared against state-of-the-art single object trackers, our non-learned method “SAM+DINOv2” performs the best.
- The problem of Instance Tracking in 3D from Egocentric Video is made much easier by leveraging camera pose and using a 3D allocentric (world) coordinate representation.



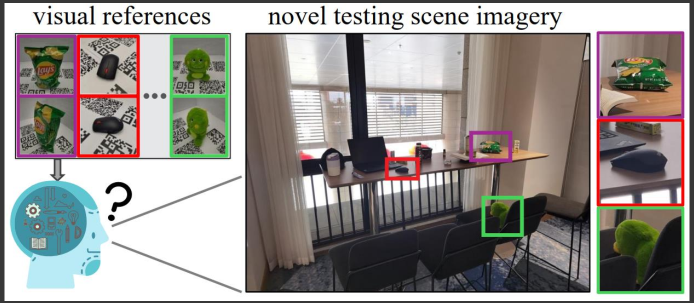
Concentric circles on the left indicate different 3D thresholds.

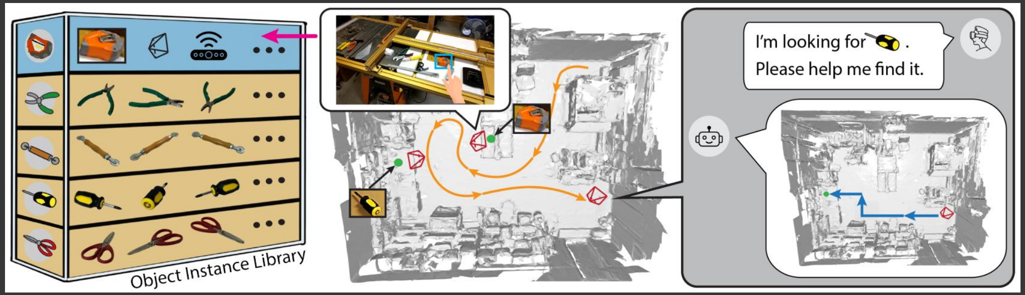
[VITKT] Kristan, et al., "Visual prompt multi-modal tracking". The tenth visual object tracking vot2022 challenge results, 2022
[EgoSTARK] Tang, et al., "Egotracks: A long-term egocentric visual object tracking dataset". NeurIPS, 2024
[Ours] Y. Zhao, H. Ma, S. Kong, C. Fowlkes. "Instance tracking in 3D scenes from egocentric videos." CVPR, 2024

Outline

Remarks

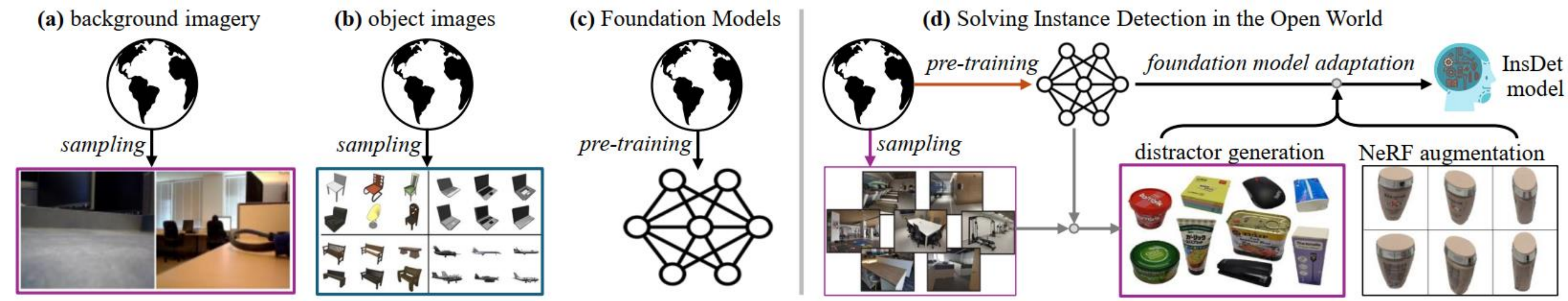
- Instance-level perception is a challenging problem even by using foundation models; it supports research in multiple fields, e.g., CV, ML, Robotics, AR/VR, and HCI.





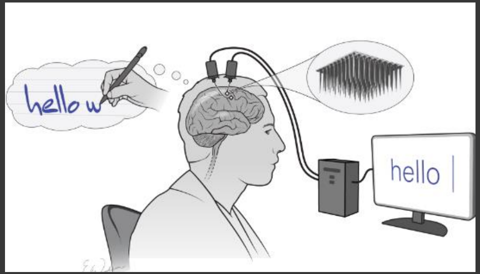
Remarks

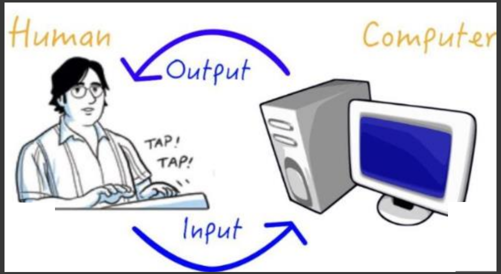
- Instance-level perception is a challenging problem even by using foundation models; it supports research in multiple fields, e.g., CV, ML, Robotics, AR/VR, and HCI.
- Open-world training (via foundation models) significantly improves robustness and generalization of models in the open world.



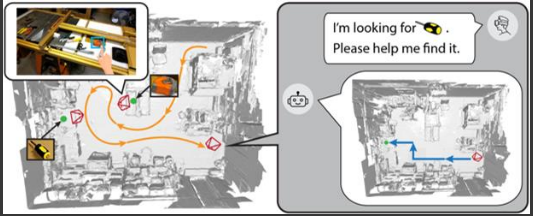
Remarks

- Instance-level perception is a challenging problem even by using foundation models; it supports research in multiple fields, e.g., CV, ML, Robotics, AR/VR, and HCI.
- Open-world training (via foundation models) significantly improves robustness and generalization of models in the open world.
- How to specify instances in a user-friendly manner? Using language?
    - Hi, Robot, please take my coffee mug to me
    - Who are your? Which coffee mug?
    - I am your master! My coffee mug is like this  !!!

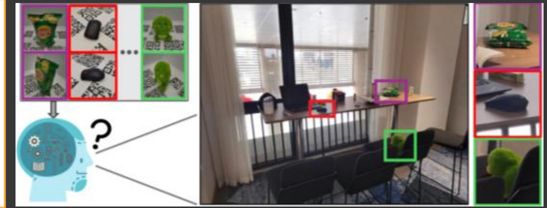




Thanks! Q&A



Instance Detection



Instance Tracking

Qianqian Shen

Nahyun Kwon

Yanan Li

Jeeeun Kim

Yunhan Zhao

Haoyu Ma

Charless Fowlkes

Shu Kong