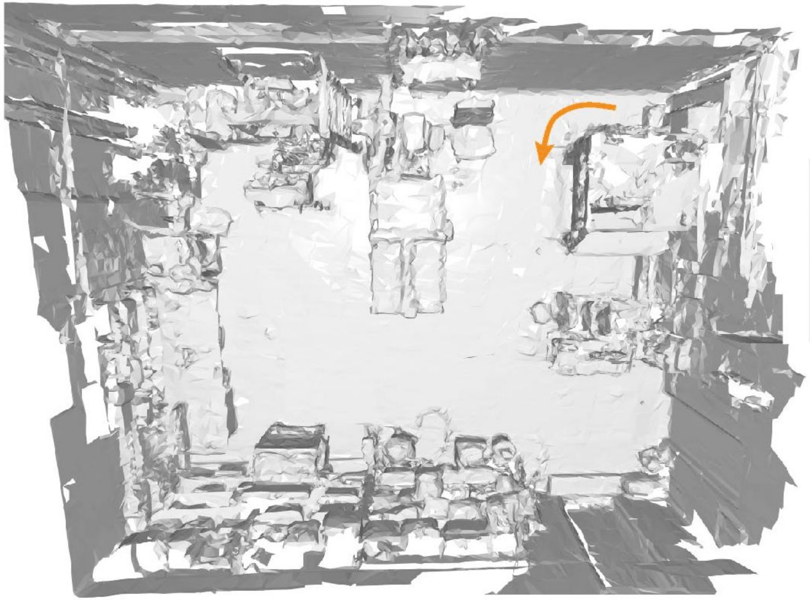
Instance Tracking in 3D Scenes from Egocentric Videos

Yunhan Zhao Haoyu Ma Shu Kong Charless Fowlkes

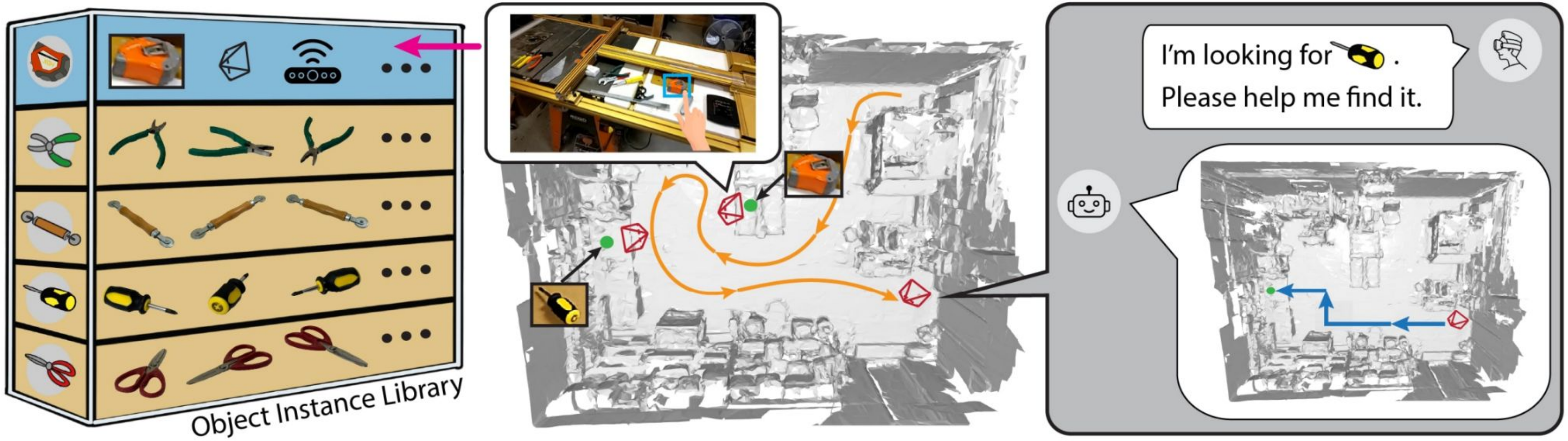
UC Irvine
Texas A&M University
澳門大學
Universidade de Macau
University of Macau

Motivation



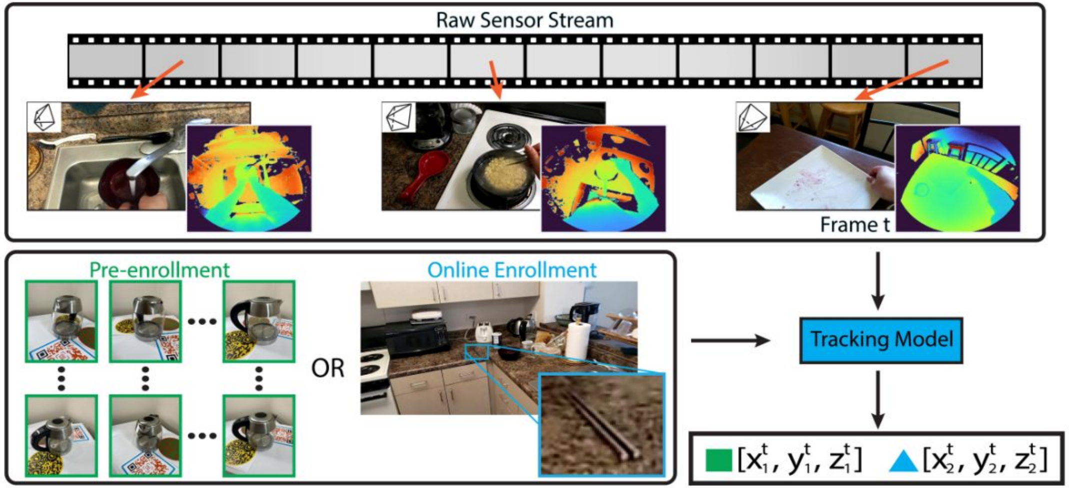
- Developing **task-aware assistive agents** running on AR/VR devices.
- Guiding users to recall the 3D locations of objects of interest (enrolled objects).

Contributions Overview



- A new benchmark protocol
    - Instance enrollments
    - Evaluation protocols
- Implement and evaluate SOTA methods
- Collect and annotate a new dataset
    - Raw videos
    - Object instance collections
    - Annotations

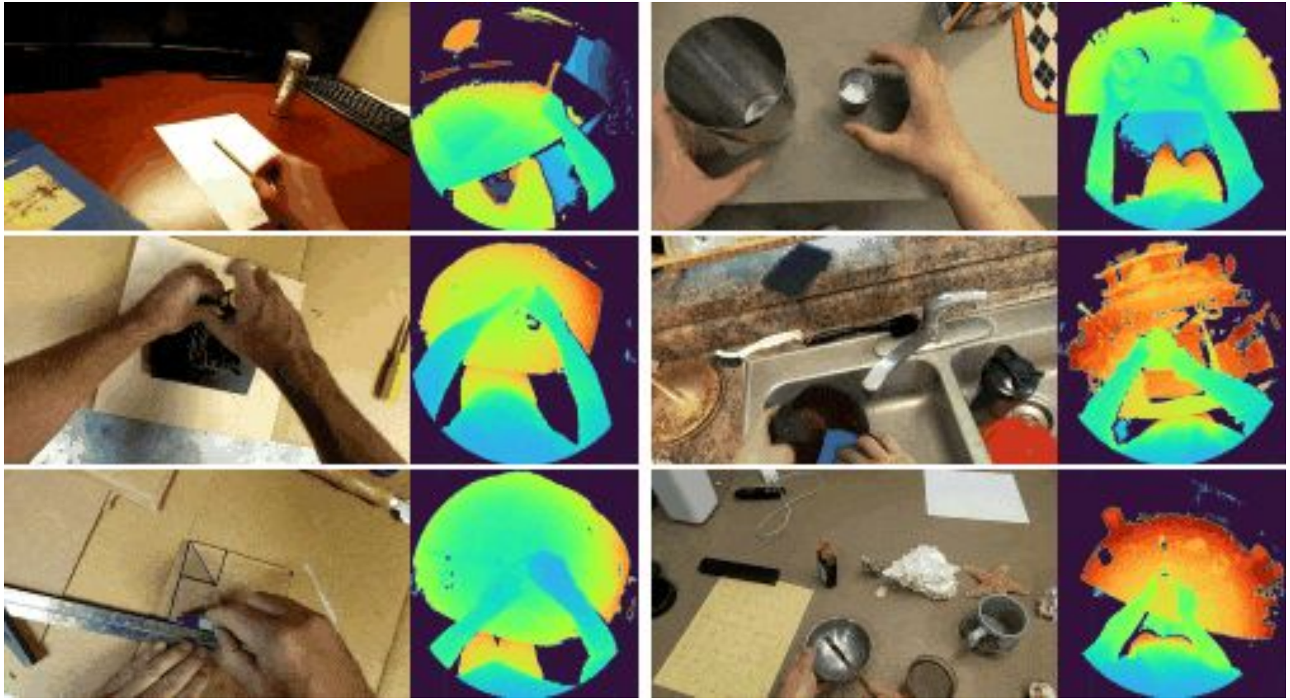
Visual Illustration of the Benchmark Task



- Given a video sequence and enrolled instances, the model keeps track of the 3D location of the object instance in a **predefined world coordinate system**.
- An important prior: an object should remain **stationary** unless being interacted with.

Benchmark Dataset – Raw Videos

- Performing daily activities captured with a Hololens 2.
- 50 videos (30 fps) with average length >= 5 min.
- 10 different indoor scenes with natural camera trajectories.



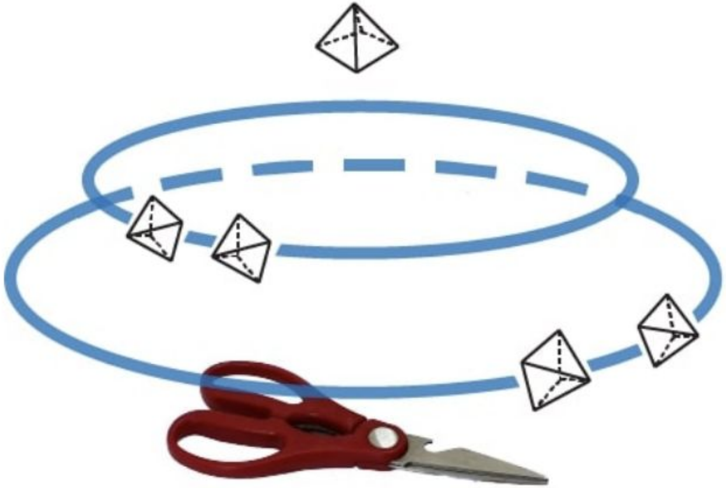
Benchmark Dataset – Instance Enrollment

Single-view online enrollment (SVOE)

- Enroll on-the-fly by the user.
- Comes with in-context information but lower visual quality (i.e. low resolution)

Multi-view pre-enrollment (MVPE)

- Pre-enroll with a collection of images for objects of interest.
- Rich visual information but not captured in the tracking environment.



Benchmark Dataset – Annotations

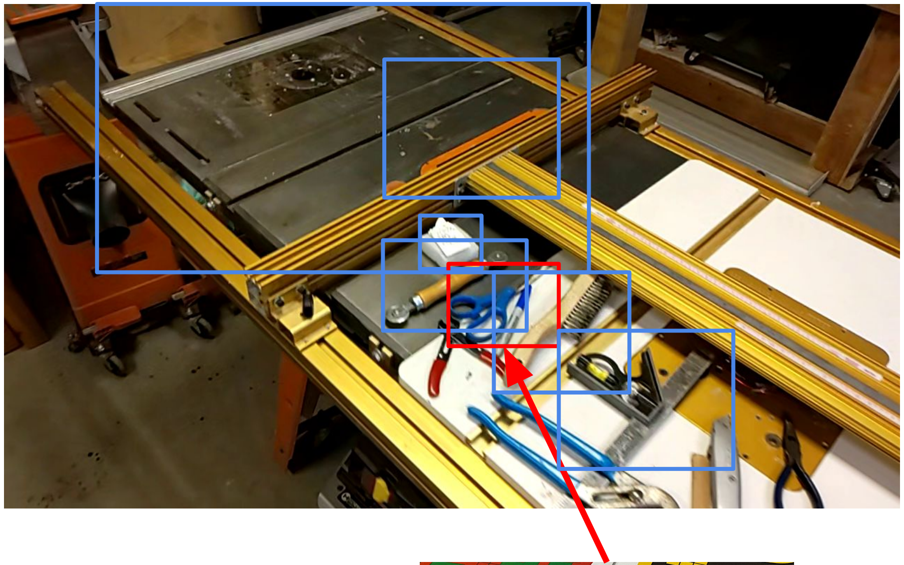
- Object instance 3D center
  - 3D positions of object instance center in the **world coordinate frame**.
- 2D bounding box annotations
  - Axis-aligned *amodal* 2D bounding boxes.
- Object motion state annotations
  - Binary annotation, either stationary or dynamic (being interacted with).

Motion state: dynamic

Proposed Approach

- Egocentric perspective:
  - Objects are mostly static while camera is moving.
  - Better to formulate the tracking problem in the **world coordinate frame**.

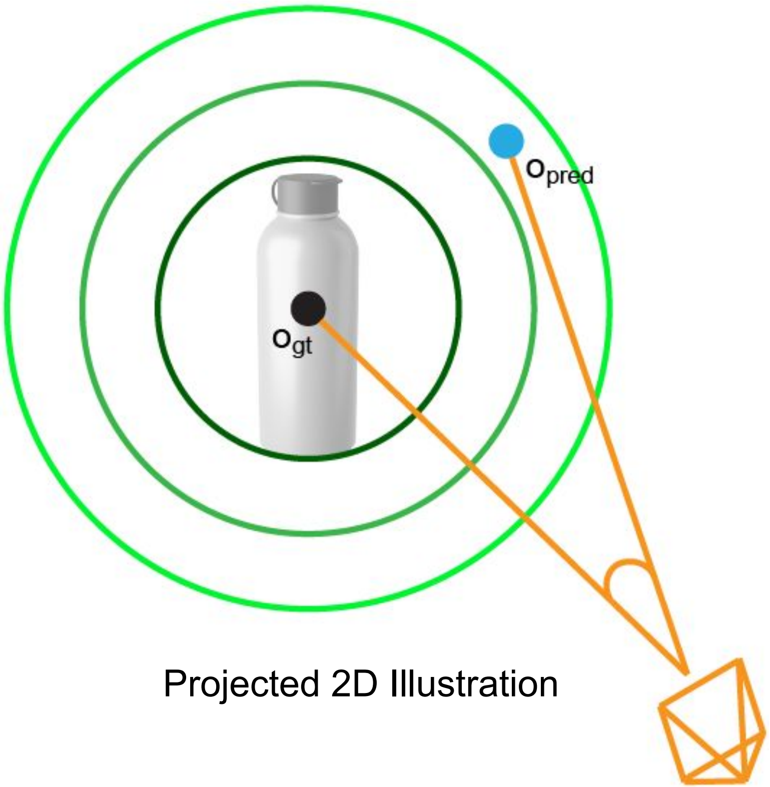
- Proposed approach:
  - Leverage recent foundation models: SAM [1], DINOv2 [2].
  - Match encoded feature cosine similarity for data association.



[1] Kirillov, Alexander, et al. "Segment anything." In ICCV 2023.
[2] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).

Evaluation Protocols

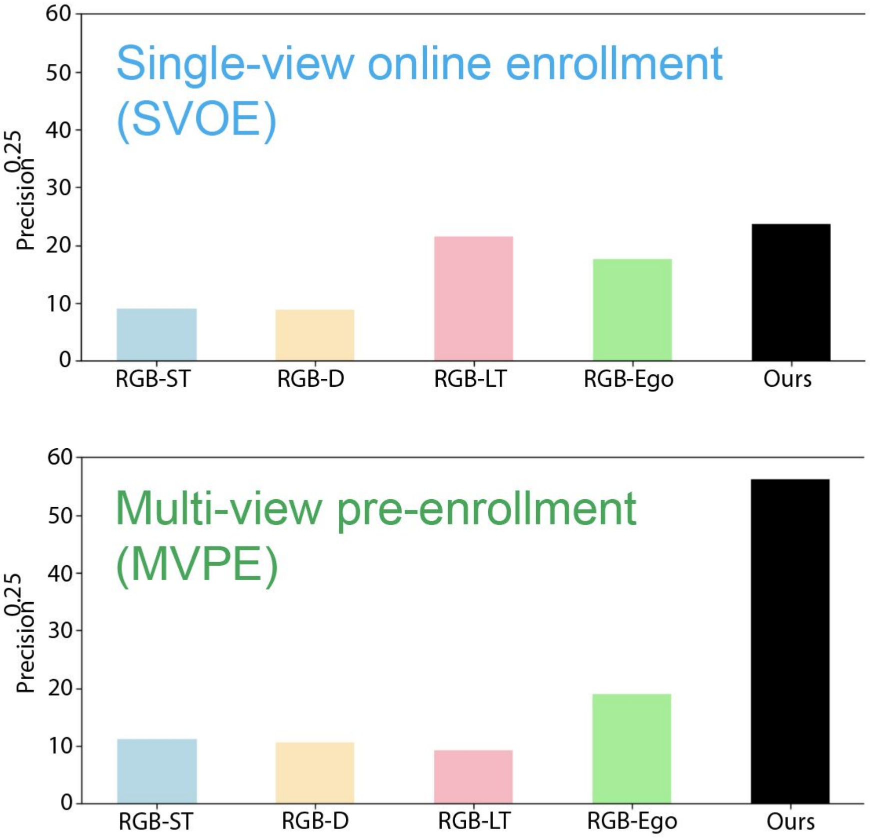
- 3D threshold-aware precision and recall
  - True positive (TP) is defined as: $\|\mathbf{O}_{pred} - \mathbf{O}_{gt}\|_2$ <= threshold.
  - Precision = TP / (TP+FP)
  - Recall = TP / (TP+FN)
- L2 and angular error
  - Require both GT and Pred to compute



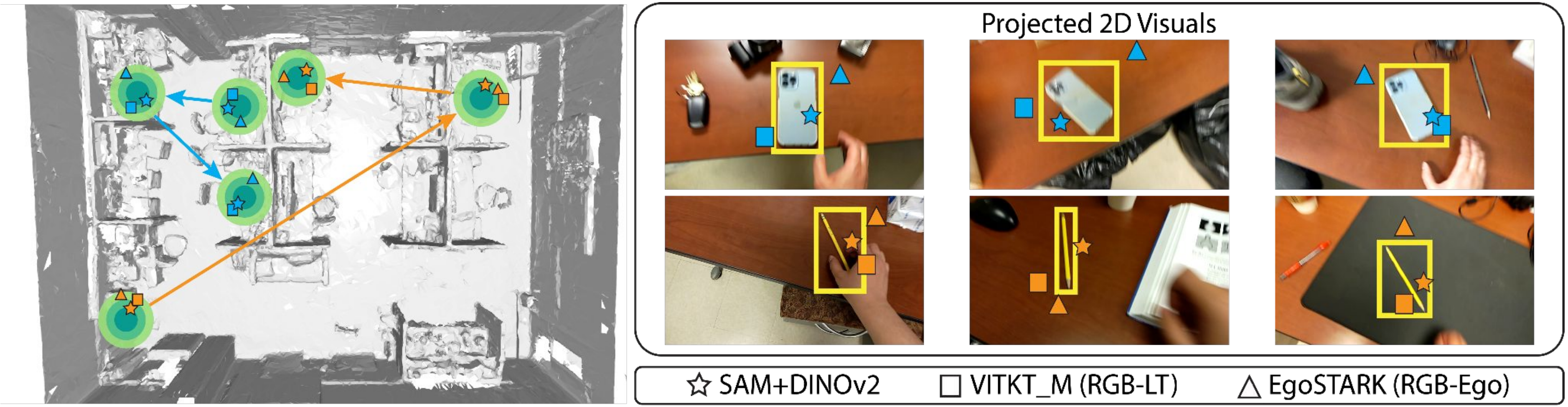
Projected 2D Illustration

Experimental Results

- We compared against state-of-the-art single object trackers.
- ST – short term; D – using depth map; LT – long term; Ego – fine-tuned egocentric.
- Our proposed approach (SAM+DINOv2) outperforms SOTA methods under both enrollment settings.



Qualitative Visualizations



- Concentric circles on the left indicate different 3D thresholds.
- The proposed model has predictions closer to the center of object.

Conclusions

- We propose **a novel benchmark problem** to study the problem of tracking object instances in 3D from egocentric videos.
- We **re-purpose and evaluate** state-of-the-art approaches and **develop a strong baseline** leveraging recent foundation models.
- Future work: (1) accurately detect object 3D motion changes; (2) better utilization of object instance information.

Github Page

Acknowledgements: This work was supported in part by the DARPA Perceptually enabled Task Guidance (PTG) Program under contract number HR00112220005.