MOD-UV: Learning Mobile Object Detectors from Unlabeled Videos

Yihong Sun, Bharath Hariharan

Cornell University

Project Page

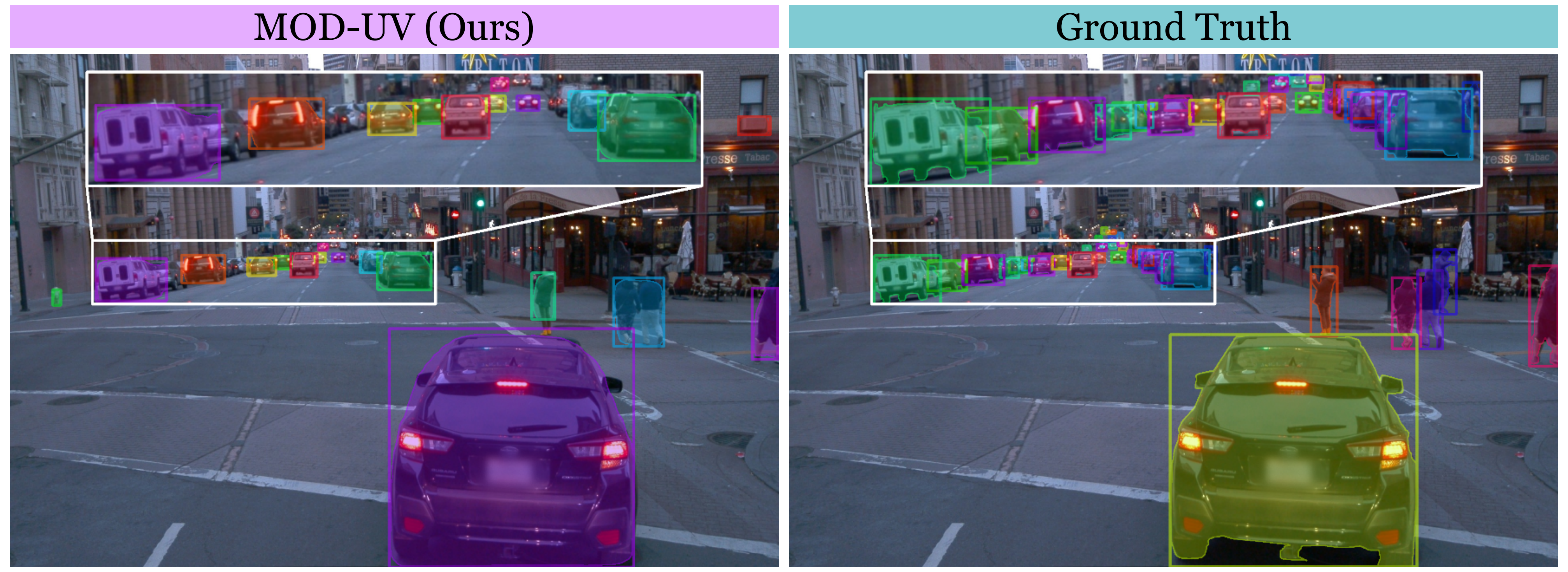
Introduction

Motivation

- Embodied agents must detect and localize **objects of interest**.
- To alleviate the burden of box-level annotations, prior works proposed **unsupervised instance detection** and **segmentation** via self-supervised features (e.g. DINO).
- However, it is unclear how pixels must be grouped into objects and which objects are of interest, which results in **over-/under-segmentation** and **irrelevant objects**.

Insight

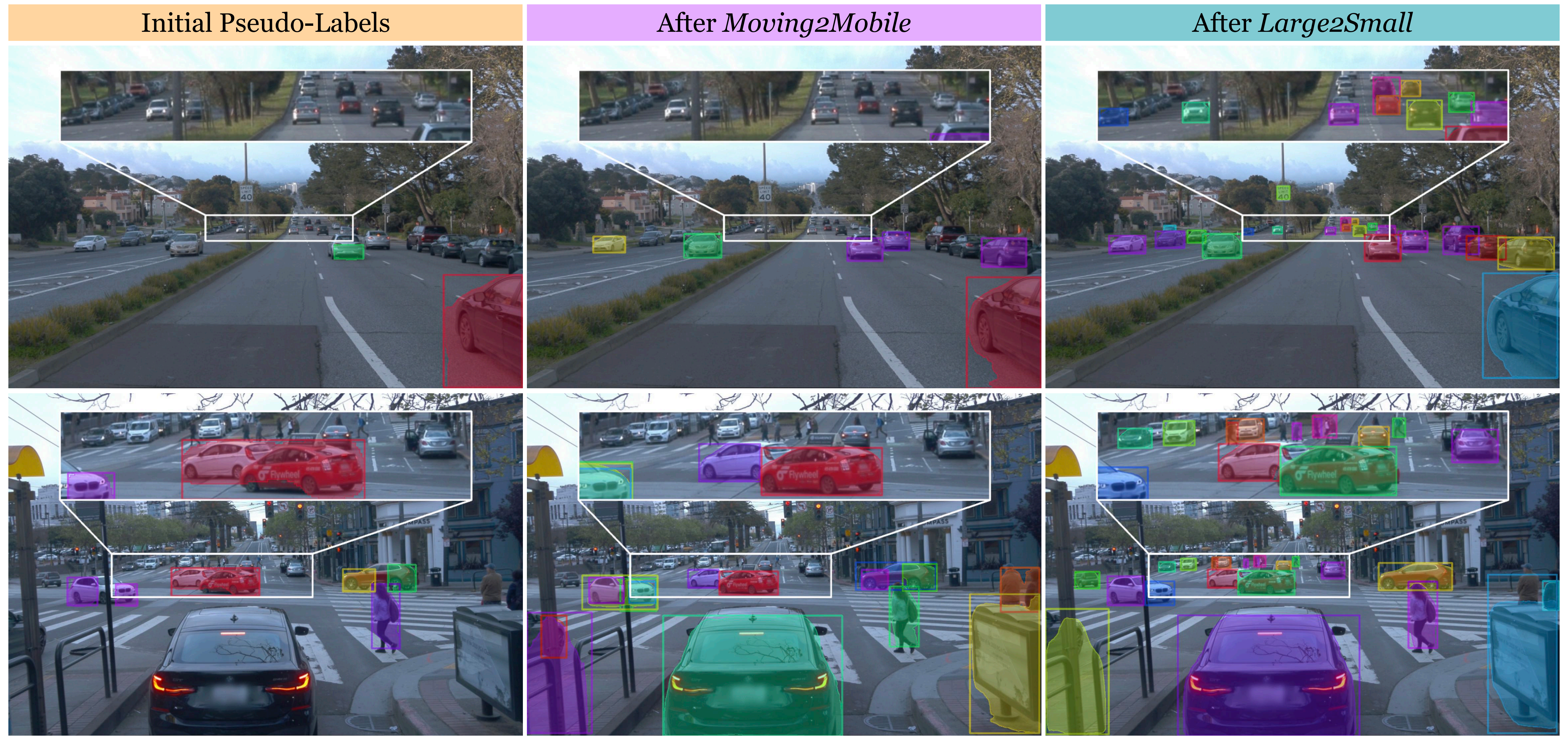
- A key missing cue is **motion**: objects of interest are typically **mobile objects that frequently move** and their motions can **specify separate instances**.



Contributions

- We propose **MOD-UV**, a **M**obile **O**bject **D**etector learned from **U**nlabeled **V**ideos only.
- We argue that **motion** can serve as an effective cue for unsupervised training of instance-level object detectors.
- We propose a **new training scheme** that trains on unlabeled videos to produce a mobile object detector that can run on **static images**.
- We demonstrate **marked improvements** over unsupervised object detection baselines across a range of datasets and metrics.

Methods



Initial Pseudo-Labels: We compute **initial object seeds** from self-supervised motion segmentation network.

- Due to bias in motion segmentation, static and small objects are missing in the initial seeds.

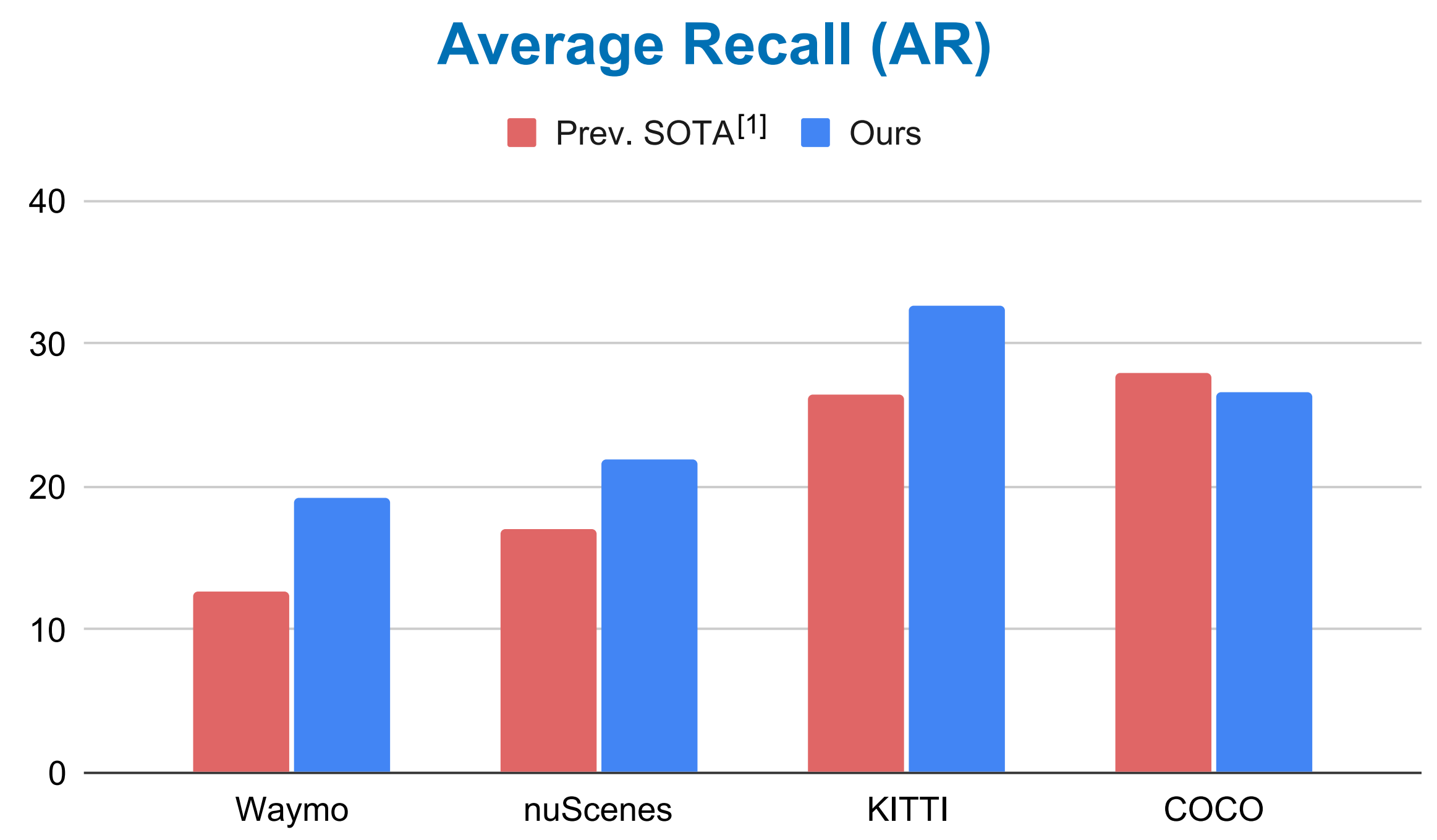
***Moving2Mobile*:** We first discover the **static** objects by training on *(single-frame, pseudo-label)* pairs.

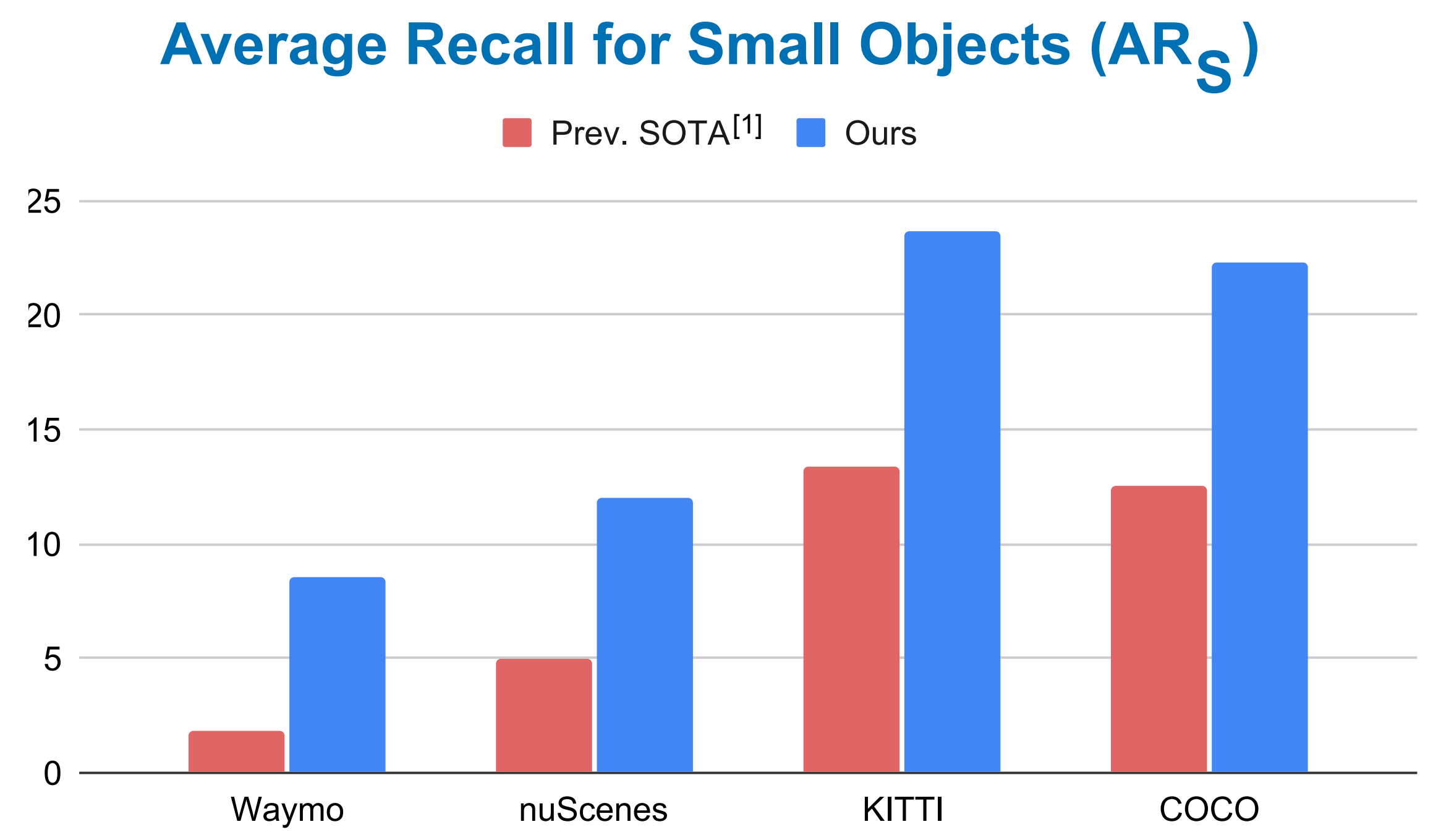
- The per-frame detector cannot distinguish static objects from moving when trained on the initial labels.

***Large2Small*:** We then discover **small** objects by training detectors at varying image scales.

- After training with the labels after *M2M*, we merge predictions at varying scales to retrieve small objects.

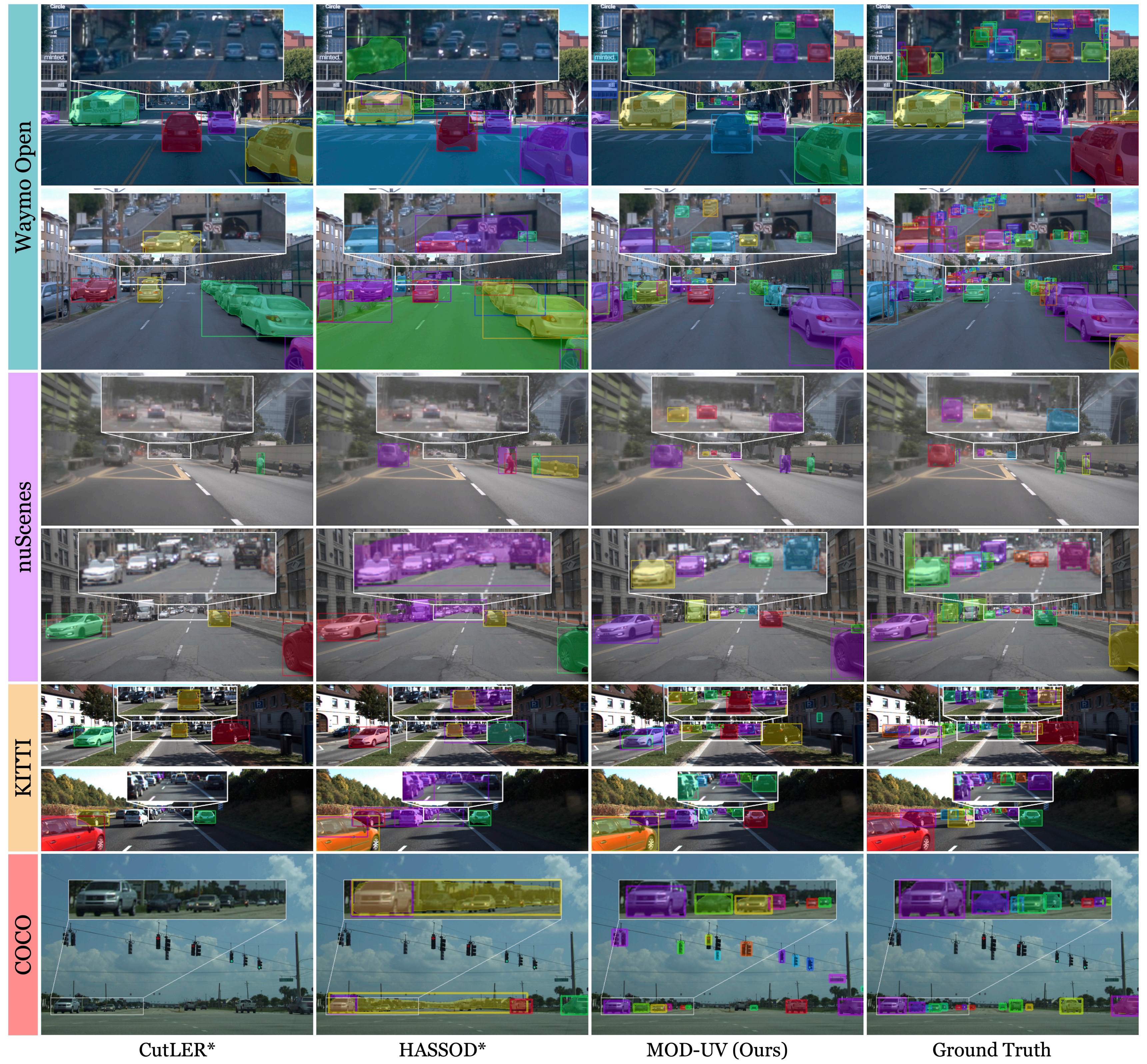
Quantitative Results





[1] Cao, Shengcao, et al. "HASSOD: Hierarchical adaptive self-supervised object detection." in *NeurIPS* 2023.

Qualitative Results



Conclusions

- We argue that **motion** is an important cue for unsupervised object detection.
- We propose a new training pipeline, **MOD-UV**, that bootstraps from **motion segmentation** but removes its bias by discovering **static and small objects**.
- MOD-UV achieves **significant** improvement over prior self-supervised detectors on multiple datasets.

Acknowledgements

- This research is based upon work supported in part by the National Science Foundation (IIS-2144117 and IIS-2107161).
- Yihong Sun is supported by an NSF graduate research fellowship.