MonoSOWA: Scalable monocular 3D Object detector Without human Annotations

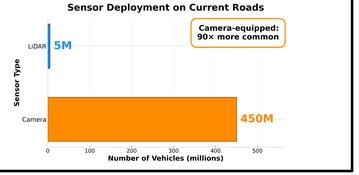
Jan Škvrna, Lukáš Neumann
Visual Recognition Group, CTU in Prague

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE
ICCV OCT 19-23, 2025 HONOLULU HAWAII

MonoSOWA automatically labels vast amounts of unlabelled videos to train monocular 3D object detectors

Motivation

- Human labelling is expensive
- Training data for 3D object detectors are scarce, because they rely on LiDAR
- There are significantly more cars with calibrated cameras than with LiDAR:



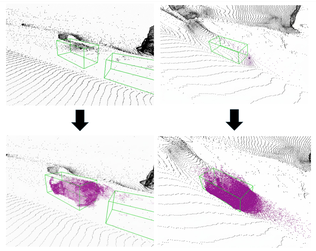
Requirements

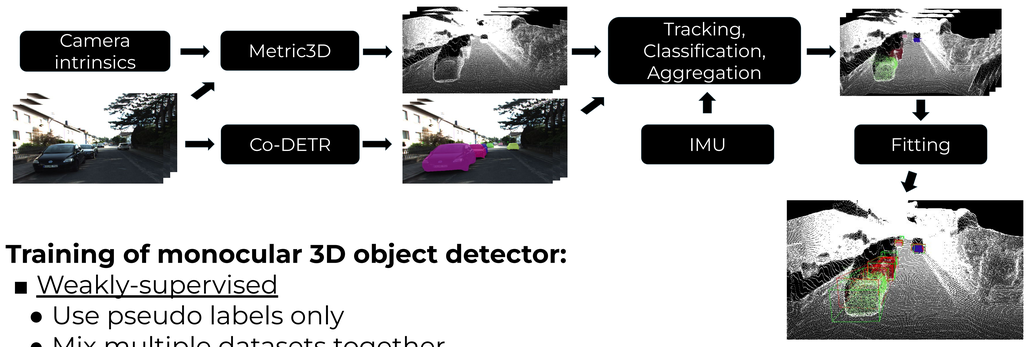
Inputs:
- Videos from calibrated cameras
- IMU data
- Template models
- Compute (~1.5 sec for 1 frame)

Outputs:
- Precise 3D object bounding boxes of vehicles
- Trained monocular detector without human annotations

Method

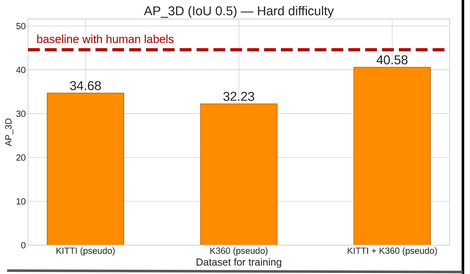
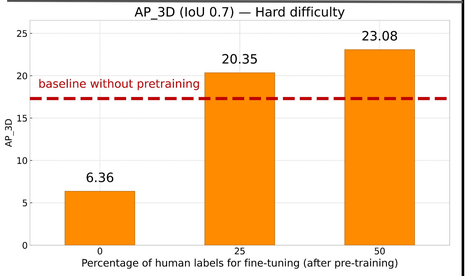
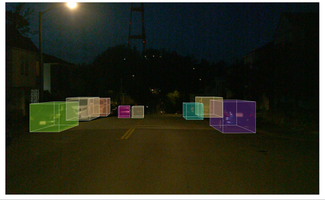
Pseudo label generation:
- Predict pseudo-LiDAR from a single frame
- Detect all vehicles
- Track and classify all vehicles
- Aggregate information over adjacent frames
- Fit templates to vehicles





Training of monocular 3D object detector:
- Weakly-supervised
  - Use pseudo labels only
  - Mix multiple datasets together
  - Use canonical object space to compensate for varying focal length
- Semi-supervised
  - Pre-train on pseudo labels
  - Fine tune on target human labels
  - Use canonical object space for faster fine tuning

Results

Training with multiple datasets yields significantly better results!
- Canonical object spaces enables to train on multiple datasets at the same time (KITTI + KITTI-360)
- Combining datasets achieves close to human performance

Pre-training on pseudo labels achieves significantly better results than training with human labels from scratch!
- Better performance is achieved with only 25% of human labels