Traffic4D: Single View Reconstruction of Repetitious Activity Using Longitudinal Self-Supervision

Fangyu Li[1], N. Dinesh Reddy[2], Xudong Chen[1], Srinivasa G. Narasimhan[2]

[1]NVIDIA [2]Carnegie Mellon University

http://www.cs.cmu.edu/~ILIM/projects/IM/TRAFFIC4D

Introduction

Goal: Traffic 4D = 3D Real World + Time at City Scale

Applications:

- Real world velocity estimation, traffic anomaly analysis, driving assistance systems, traffic monitoring for smart cities, *etc.*

Challenges:

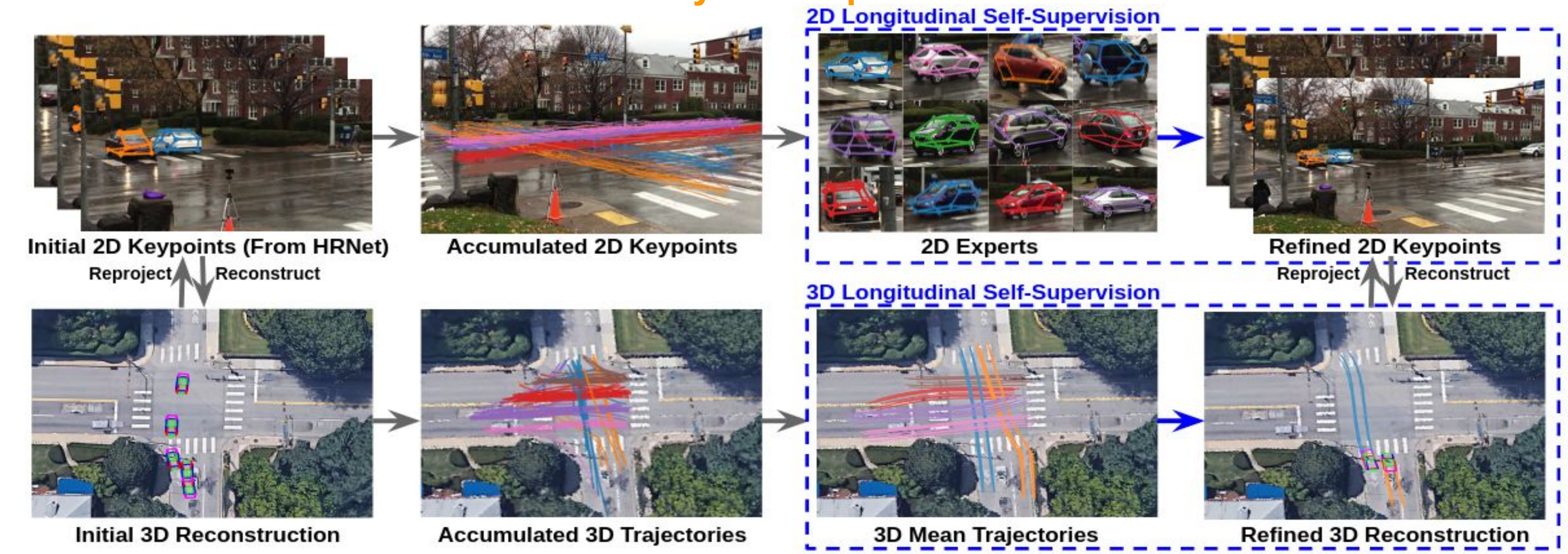
- *Data:* Large scale depth sensors are too expensive; Surveillance cameras only collect 2D information, requiring single view algorithm
- *Algorithm:* 2D detectors may fail due to occlusion or appearance; 3D direct reconstruction is spatially inaccurate and temporally inconsistent

Key Observation: Traffic activity is repetitious in long term; learned longitudinal information can be used to improve 2D/3D accuracy

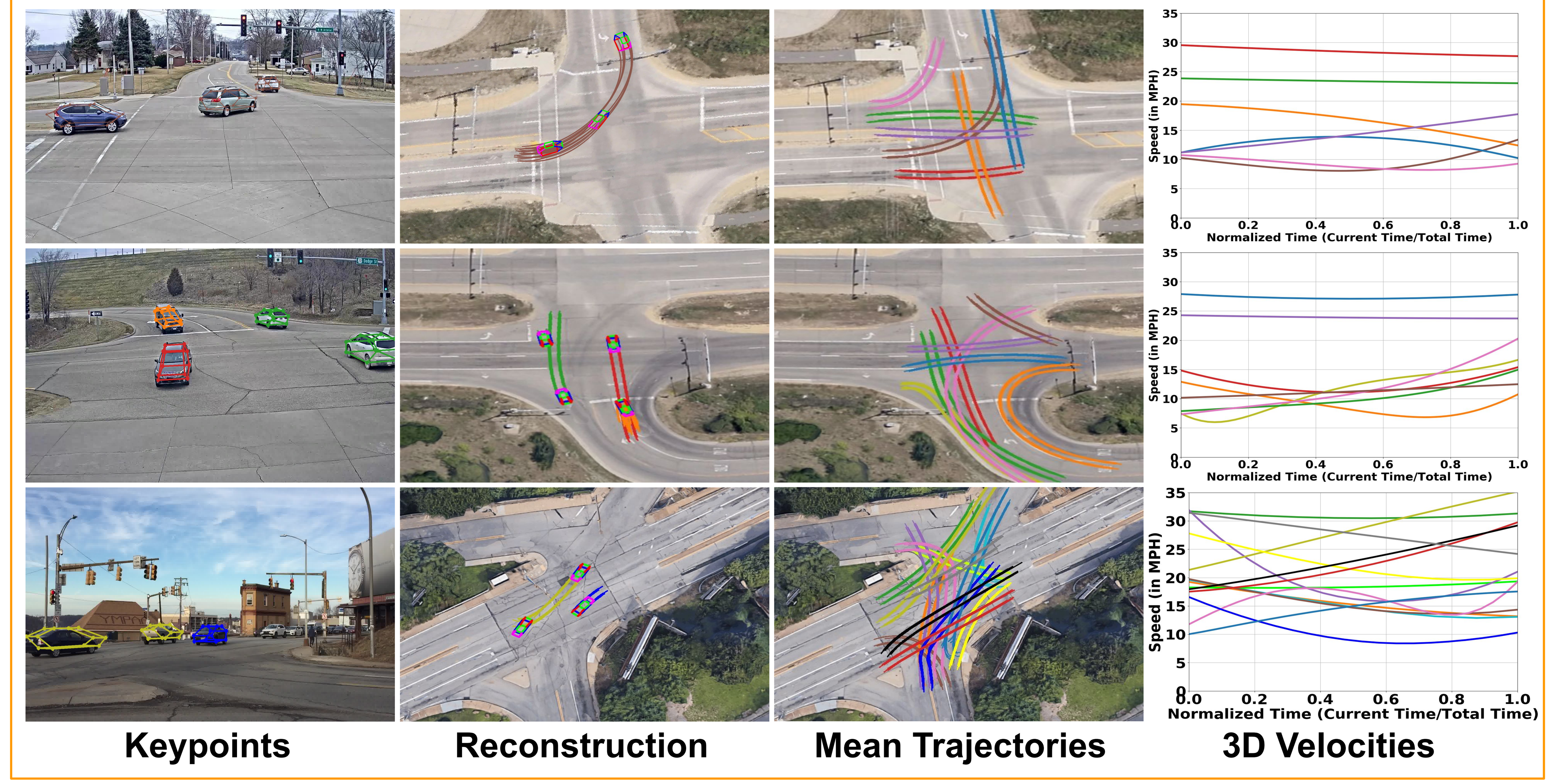
Contributions:

- Joint optimization for longitudinal reconstruction
- Scene-specific repetitious activity clustering outperforming SOTA
- 2D/3D longitudinal self-supervision to improve 2D/3D accuracy

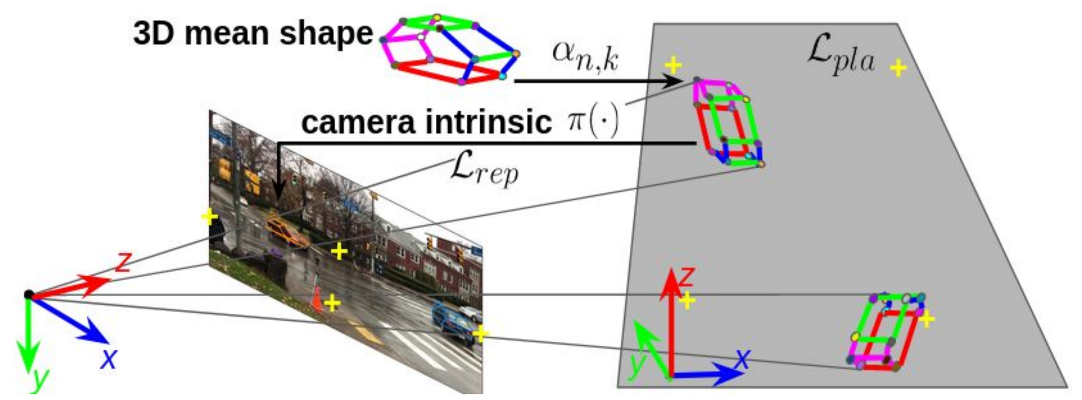
System Pipeline



4D Reconstruction



Initial Joint Optimization



Initialization:

Off-the-shelf 2D Keypoints: HRNet

Tracking: Visual Intersection Over Union Tracker

Vehicle Model: 3D Active Shape Model

Reprojection Loss:

$$\mathcal{L}_{rep} = \sum_{n,m,j} \|\boldsymbol{\pi}(\mathbf{P}^{(c)}_{n,m,j}) - \mathbf{p}_{n,m,j}\|^2$$

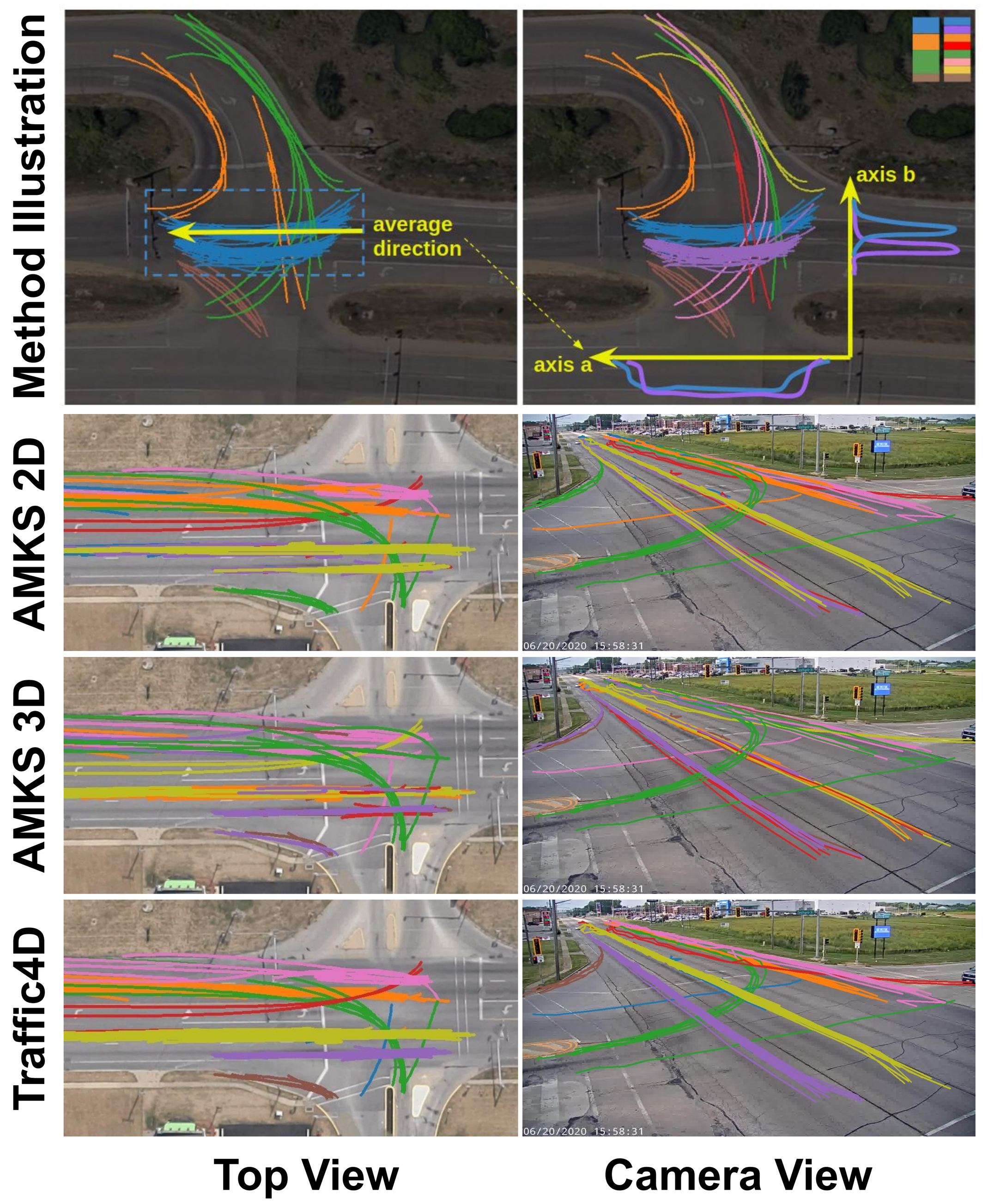
Joint Planar Loss:

$$\mathcal{L}_{pla} = \sum_{n,m} \frac{(\eta_1 P^{(c)}_{n,m,j_b,x} + \eta_2 P^{(c)}_{n,m,j_b,y} + \eta_3 P^{(c)}_{n,m,j_b,z} + \eta_4)^2}{\eta_1^2 + \eta_2^2 + \eta_3^2}$$

Total Loss:

$$\mathcal{L}_{rec} = \gamma_1 \mathcal{L}_{rep} + \gamma_2 \mathcal{L}_{pla}$$

Repetitious Activity Clustering



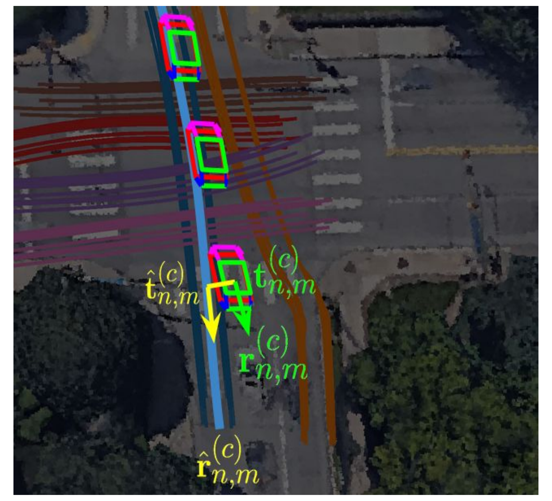
2D Longitudinal Self-Supervision

For each bad vehicle ...

- Retrieve 2D expert using 2D bounding boxes from its consecutive frames
- Align the expert's and the vehicle's keypoints according to bounding box scale and translation

$$\mathbf{b}_{n,m} = \mathbf{s}^{(b)}_{n,m}\hat{\mathbf{b}}_{n,m} + \mathbf{t}^{(b)}_{n,m}$$

3D Longitudinal Self-Supervision

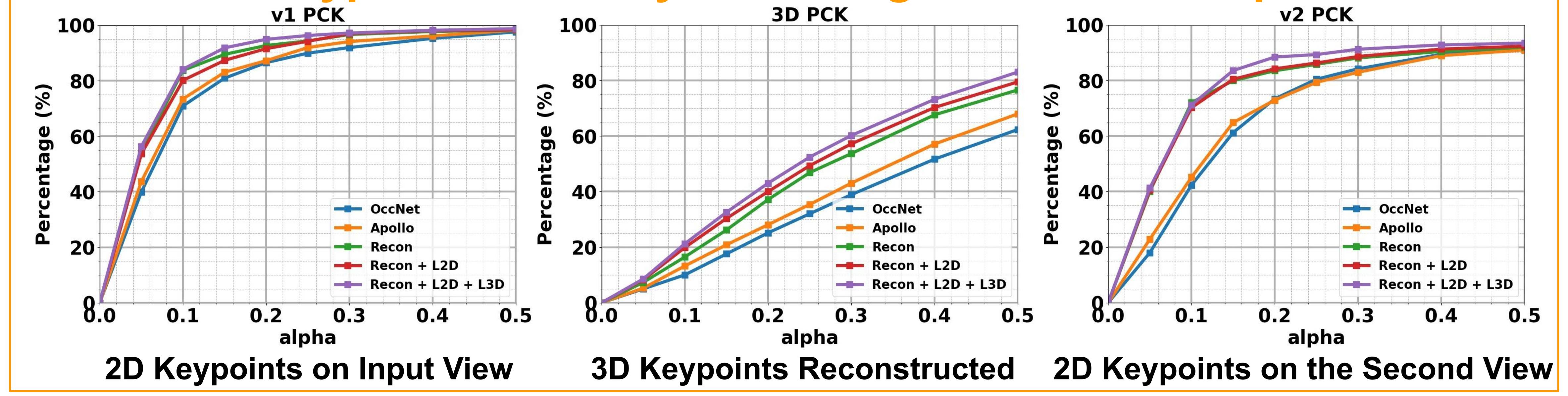


For each vehicle in consecutive frames ...

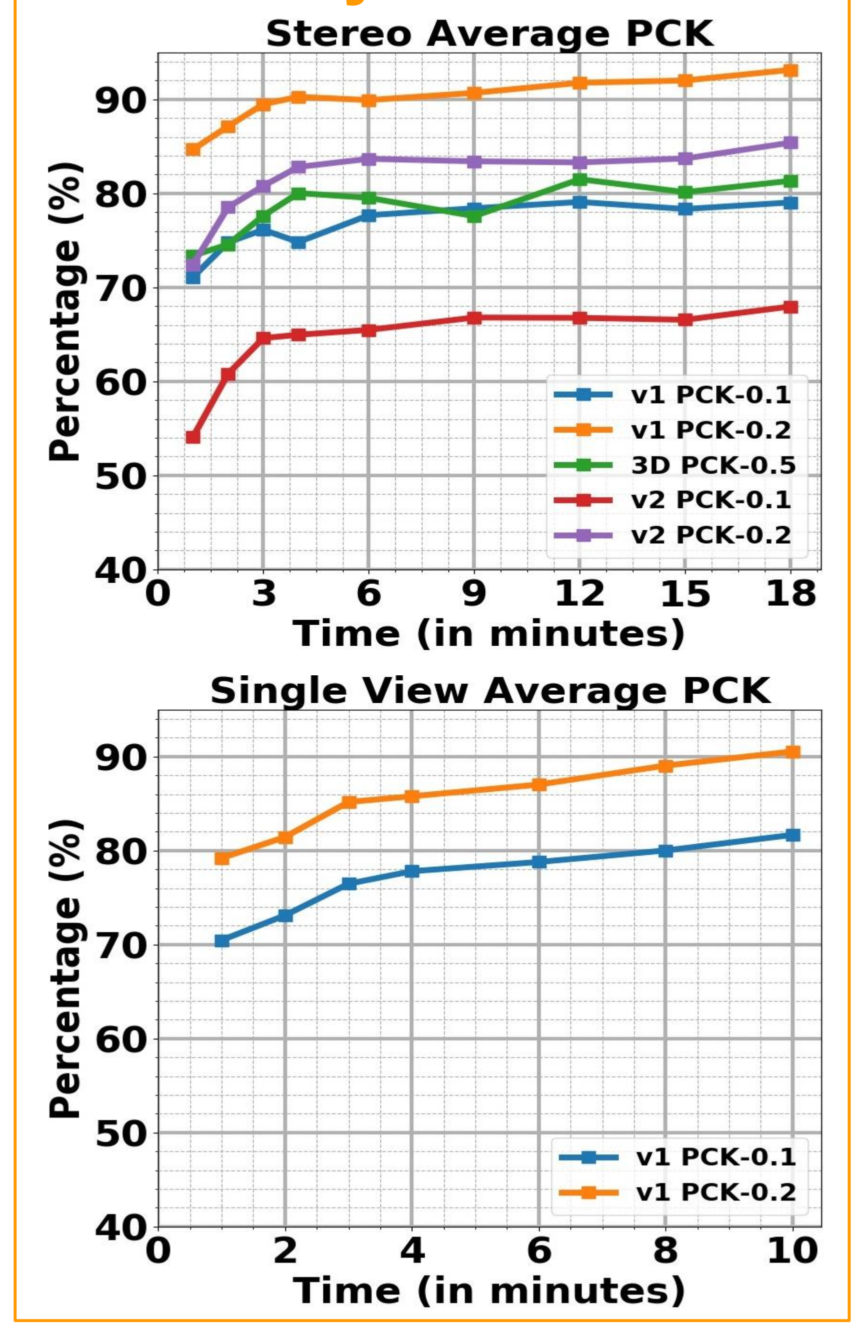
- Find the corresponding 3D expert with minimal Chamfer distance
- Refine 3D poses using **Longitudinal Loss** (w/ reprojection and planar)

$$\mathcal{L}_{long} = \beta_4 \|\mathbf{t}^{(c)}_{n,m} - \hat{\mathbf{t}}^{(c)}_{n,m}\|^2 + \beta_5 \|\mathbf{r}^{(c)}_{n,m} - \hat{\mathbf{r}}^{(c)}_{n,m}\|^2$$

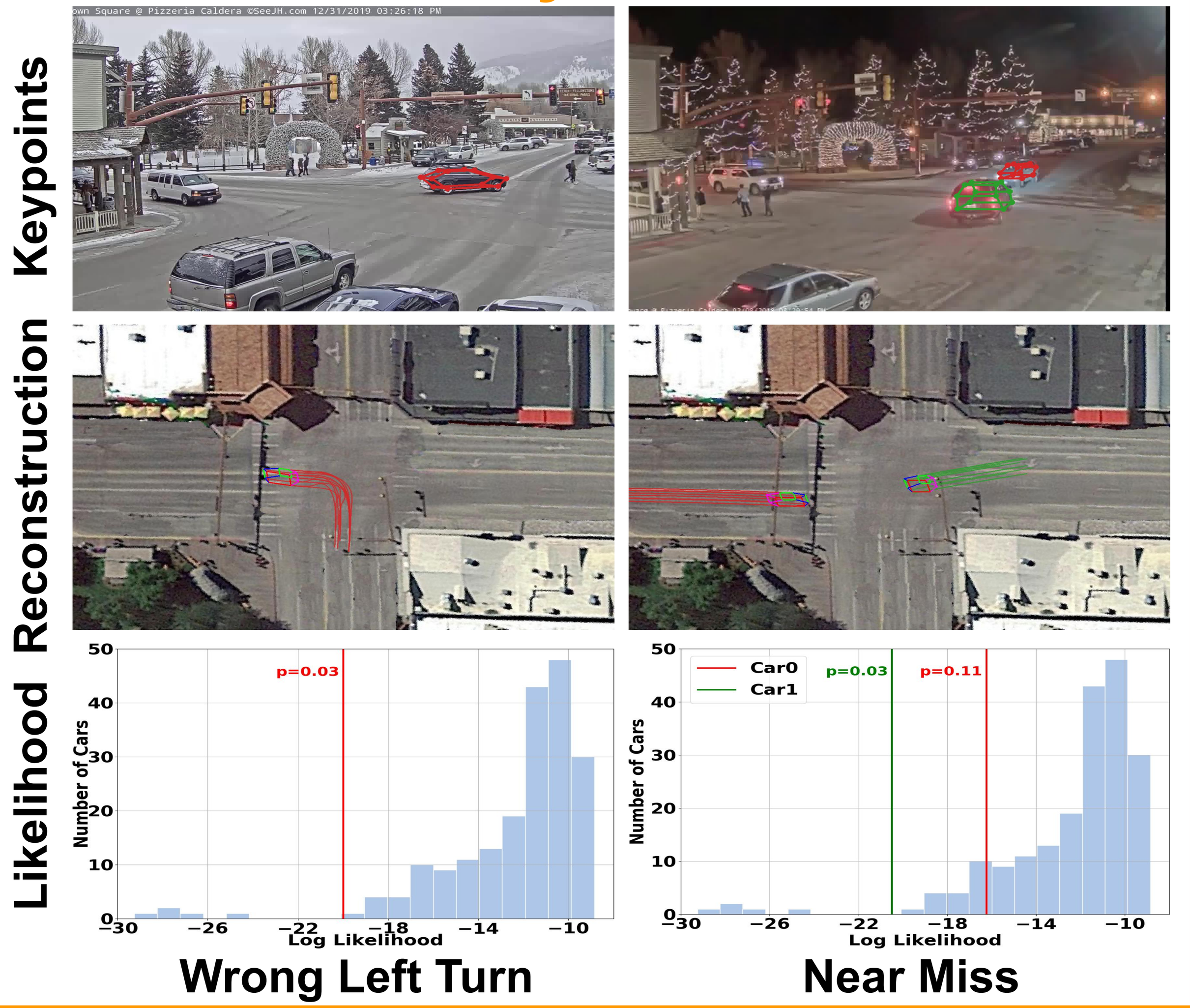
2D/3D Keypoint Accuracy with Longitudinal Self-Supervision



Accuracy vs Duration



Anomaly Detection



Acknowledgement

This paper was supported in parts by NSF Grants IIS-1900821 and CNS-2038612, DOT RITA Mobility-21 Grant 69A3551747111, and a PhD fellowship from Amazon Go.