

Visual-Inertial-Semantic Scene Representation for 3D Object Detection

Jingming Dong, Xiaohan Fei, Stefano Soatto

Introduction

Overview

- We describe a system to detect objects in three-dimensional space using visual and inertial sensors (accelerometer and gyroscope).
- The resulting system can process the video stream causally in real time, and provides a representation of objects in the scene that is persistent.

Motivation

Principles

- Objects exist in the scene, not in the image;
- They persist, so confidence on their presence should grow as more evidence is accrued from multiple (test) images;
- Once seen, the system should be aware of their presence even when temporarily not visible;
- Such awareness should allow it to predict when they will return into view, based on scene geometry and topology;
- Objects have characteristic shape and size in 3D, and vestibular (inertial) sensors provide a global scale and orientation reference that the system should leverage on.

Methods

Problem Formalization

Given measurements up to time t y^t

Estimate scene ξ and objects z^{j} with geometric (shape & pose) s_j and semantic (class) l_j attributes

Quantity of interest: posterior of objects in the scene

$$p(\xi, z^j|y^t) = p(\xi|z^j)p(z^j|y^t)$$

-- can be learned from data

which is a *minimal sufficient* representation.

Methods (cont'd)

Solution

I. marginalize over viewpoint and point-cloud

$$p(z^{j}|y^{t}) = \int p(z^{j}|g_{t}, x, y^{t})dP(g_{t}, x|y^{t})$$

the measure is given by SLAM

$$p(g_t, x|y^t) \simeq \mathcal{N}(\hat{g}_{t|t}, \hat{x}_{t|t}; P_{t|t})$$

when covariance P is small

$$\hat{p}_{g,x}(z^j|y^t) \doteq p(z^j|g_t = \hat{g}_{t|t}, x = \hat{x}_{t|t}, y^t) \simeq p(z^j|y^t)$$

update is simplified to

$$\hat{p}(z^{j}|y^{t+1}) \propto \underbrace{p(y_{t+1}|z^{j}, \hat{g}_{t|t}u_{t}, \hat{x}_{t|t})}_{\text{CNN}} \underbrace{\hat{p}(z^{j}|y^{t})}_{\text{BF}}$$

II. CNN as an implicit likelihood query function

$$\phi_{\text{CNN}}(l|I_{t_{|b_j}})_{[k]} \simeq p(I_t|l_k,b_j)$$

$$p(y_t|z^j, g_t, \hat{x}) \simeq \phi_{\text{CNN}}(l|I_{t|_{\pi(g_t s_i)}})_{[k]} \mathcal{N}(\bar{u}; Q)$$

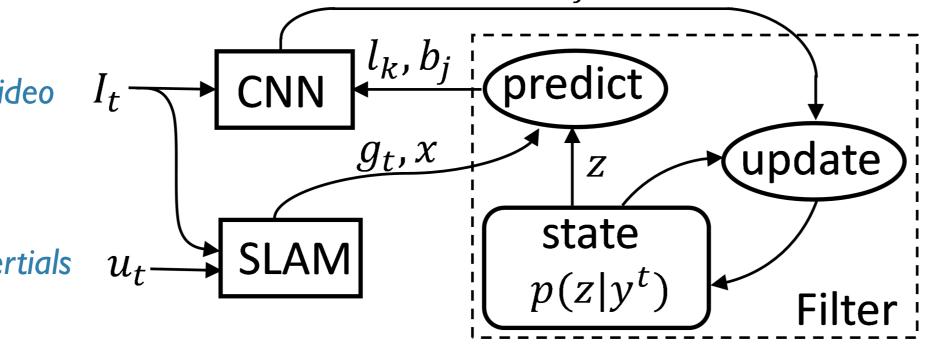
III. further assumptions between objects

$$\hat{p}_{g,x}(z^{j}|y^{t}) \doteq p(z^{j}|y^{t}, g_{t}, x) \simeq \prod_{j} p(z_{j}|y^{t}, g_{t}, x, z^{-j})$$

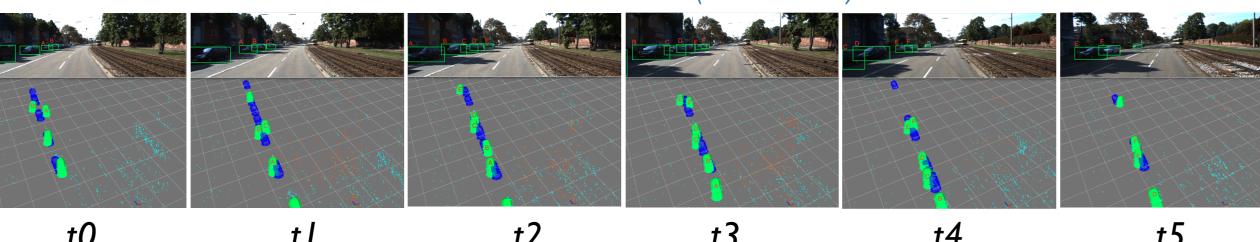
$$p(s_{j}, l_{j}|y^{t}, g_{t}, x, z^{-j}) = \underbrace{p(s_{j}|l_{j}, y^{t}, g_{t}, x, s^{-j})}_{\text{EKF}} \underbrace{P(l_{j}|y^{t}, g_{t}, x, l^{-j})}_{\text{PMF}}$$

System Flow Chart

 $p(I_t|l_k,b_i)$ Likelihood by CNN



Geometry (shape & pose) Semantics (class label)



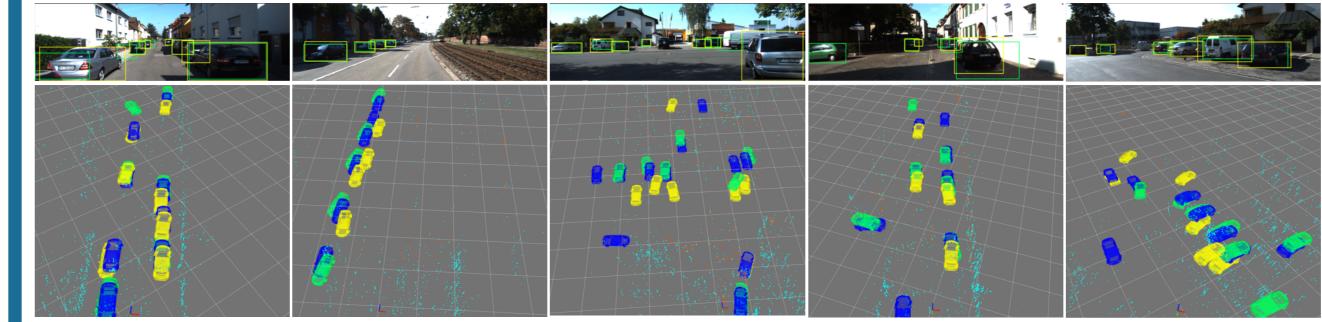
State Update Illustration. The location and orientation of the cars are refined causally over time.

Performance

- Geometry: Intel Core i7 Overall ~300 FPS
- Semantics: Nvidia GTX 760 Overall ~17 FPS
- Bottleneck: image-based object detectors
- Current implementation runs CNN every 3 frames

Comparison and Evaluation

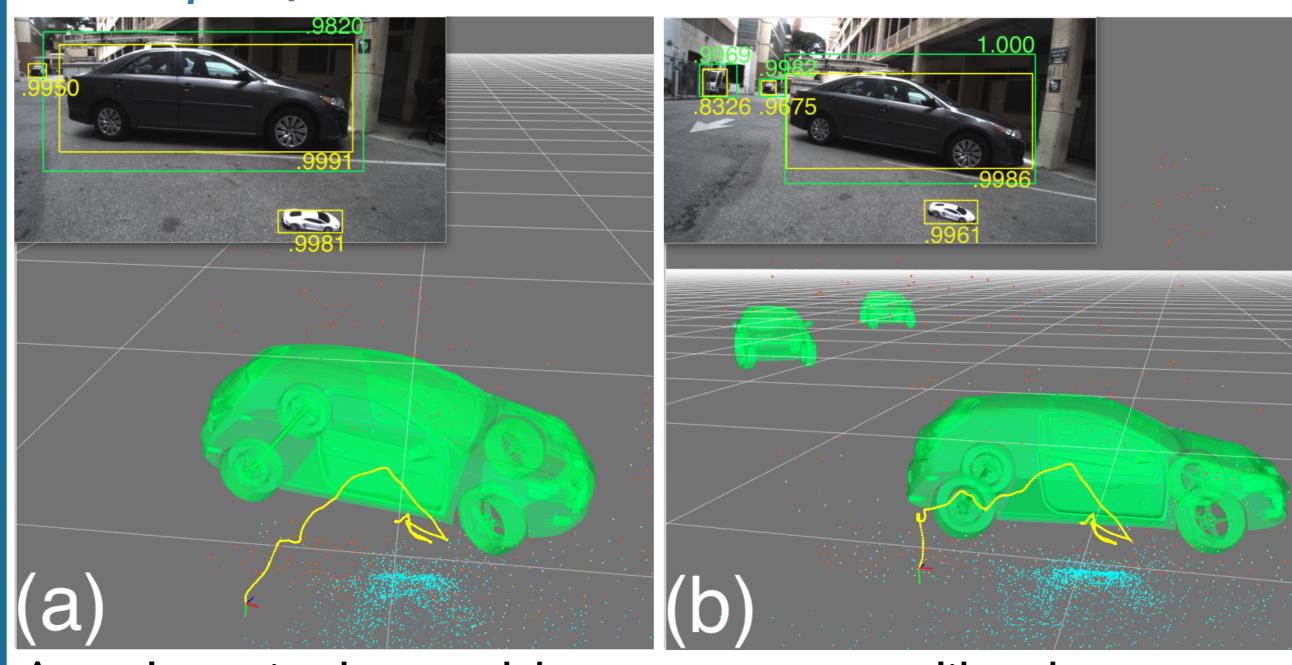
1 OSITION CITO	√ 0.0 m			\ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \			₹ 1.0 III		
method	#TP	Precision	Recall	#TP	Precision	Recall	#TP	Precision	Recall
Ours-FNL	150	0.14	0.10	355	0.34	0.24	513	0.49	0.35
Ours-INST	135	0.13	0.09	270	0.26	0.18	368	0.35	0.25
SubCNN	99	0.10	0.07	254	0.26	0.17	376	0.38	0.26
Ours-FNL	157	0.15	0.11	367	0.35	0.25	533	0.50	0.36
Ours-INST	141	0.13	0.10	283	0.27	0.19	388	0.37	0.26
SubCNN	99	0.10	0.07	257	0.26	0.17	383	0.38	0.26
Ours-FNL	169	0.16	0.11	425	0.40	0.29	618	0.58	0.42
Ours-INST	149	0.14	0.10	320	0.30	0.22	450	0.43	0.31
SubCNN	104	0.10	0.07	272	0.27	0.18	409	0.41	0.28
	method Ours-FNL Ours-INST SubCNN Ours-FNL Ours-INST SubCNN Ours-FNL Ours-INST	method #TP Ours-FNL 150 Ours-INST 135 SubCNN 99 Ours-FNL 157 Ours-INST 141 SubCNN 99 Ours-FNL 169 Ours-INST 149	method #TP Precision Ours-FNL 150 0.14 Ours-INST 135 0.13 SubCNN 99 0.10 Ours-FNL 157 0.15 Ours-INST 141 0.13 SubCNN 99 0.10 Ours-FNL 169 0.16 Ours-INST 149 0.14	method #TP Precision Recall Ours-FNL 150 0.14 0.10 Ours-INST 135 0.13 0.09 SubCNN 99 0.10 0.07 Ours-FNL 157 0.15 0.11 Ours-INST 141 0.13 0.10 SubCNN 99 0.10 0.07 Ours-FNL 169 0.16 0.11 Ours-INST 149 0.14 0.10	method #TP Precision Recall #TP Ours-FNL 150 0.14 0.10 355 Ours-INST 135 0.13 0.09 270 SubCNN 99 0.10 0.07 254 Ours-FNL 157 0.15 0.11 367 Ours-INST 141 0.13 0.10 283 SubCNN 99 0.10 0.07 257 Ours-FNL 169 0.16 0.11 425 Ours-INST 149 0.14 0.10 320	method #TP Precision Recall #TP Precision Ours-FNL 150 0.14 0.10 355 0.34 Ours-INST 135 0.13 0.09 270 0.26 SubCNN 99 0.10 0.07 254 0.26 Ours-FNL 157 0.15 0.11 367 0.35 Ours-INST 141 0.13 0.10 283 0.27 SubCNN 99 0.10 0.07 257 0.26 Ours-FNL 169 0.16 0.11 425 0.40 Ours-INST 149 0.14 0.10 320 0.30	method #TP Precision Recall #TP Precision Recall Ours-FNL 150 0.14 0.10 355 0.34 0.24 Ours-INST 135 0.13 0.09 270 0.26 0.18 SubCNN 99 0.10 0.07 254 0.26 0.17 Ours-FNL 157 0.15 0.11 367 0.35 0.25 Ours-INST 141 0.13 0.10 283 0.27 0.19 SubCNN 99 0.10 0.07 257 0.26 0.17 Ours-FNL 169 0.16 0.11 425 0.40 0.29 Ours-INST 149 0.14 0.10 320 0.30 0.22	method #TP Precision Recall #TP Precision Recall #TP Ours-FNL 150 0.14 0.10 355 0.34 0.24 513 Ours-INST 135 0.13 0.09 270 0.26 0.18 368 SubCNN 99 0.10 0.07 254 0.26 0.17 376 Ours-FNL 157 0.15 0.11 367 0.35 0.25 533 Ours-INST 141 0.13 0.10 283 0.27 0.19 388 SubCNN 99 0.10 0.07 257 0.26 0.17 383 Ours-FNL 169 0.16 0.11 425 0.40 0.29 618 Ours-INST 149 0.14 0.10 320 0.30 0.22 450	method #TP Precision Recall #TP Precision Recall #TP Precision Ours-FNL 150 0.14 0.10 355 0.34 0.24 513 0.49 Ours-INST 135 0.13 0.09 270 0.26 0.18 368 0.35 SubCNN 99 0.10 0.07 254 0.26 0.17 376 0.38 Ours-FNL 157 0.15 0.11 367 0.35 0.25 533 0.50 Ours-INST 141 0.13 0.10 283 0.27 0.19 388 0.37 SubCNN 99 0.10 0.07 257 0.26 0.17 383 0.38 Ours-FNL 169 0.16 0.11 425 0.40 0.29 618 0.58 Ours-INST 149 0.14 0.10 320 0.30 0.22 450 0.43



Blue: GT Green: Our results Yellow: SubCNN

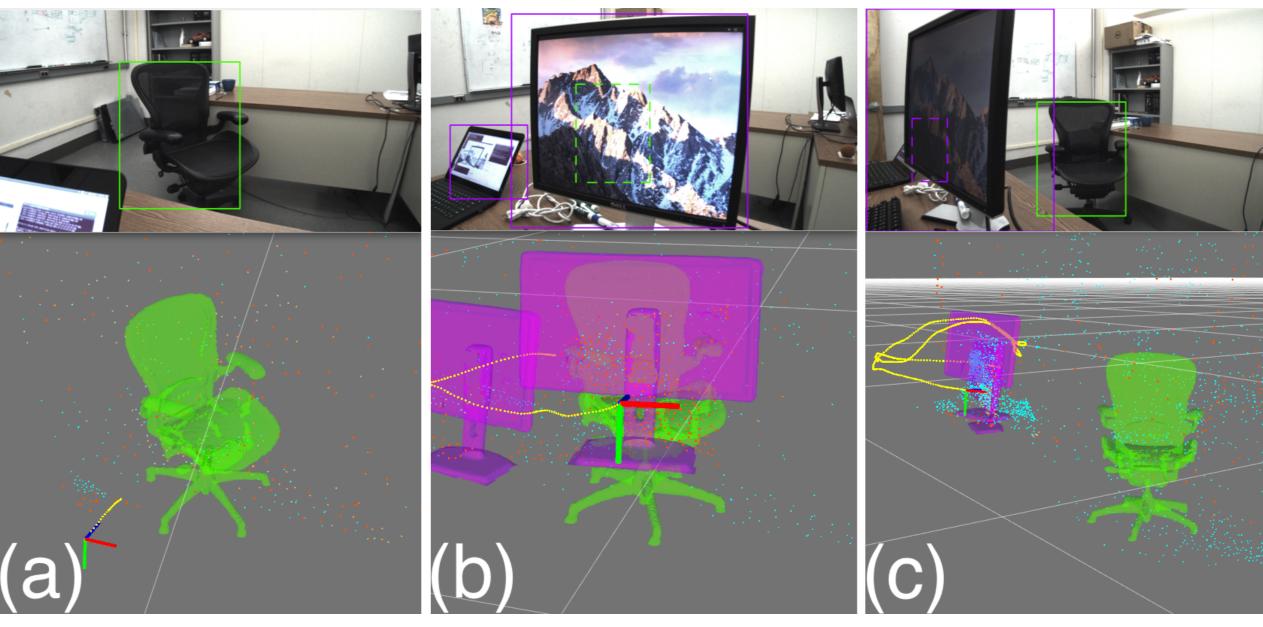
System Highlights

Class-specific Priors --- Characteristic Scales



A real car is detected by our system, unlike the toy car, Indoor Sequences despite both scoring high likelihood and therefore being detected by an image-based system.

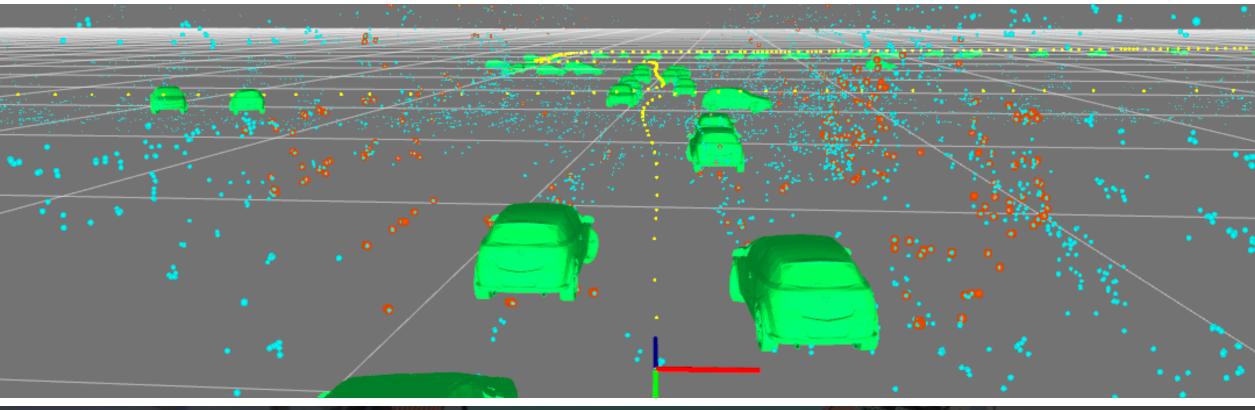
Occlusion Management & Short-term Memory



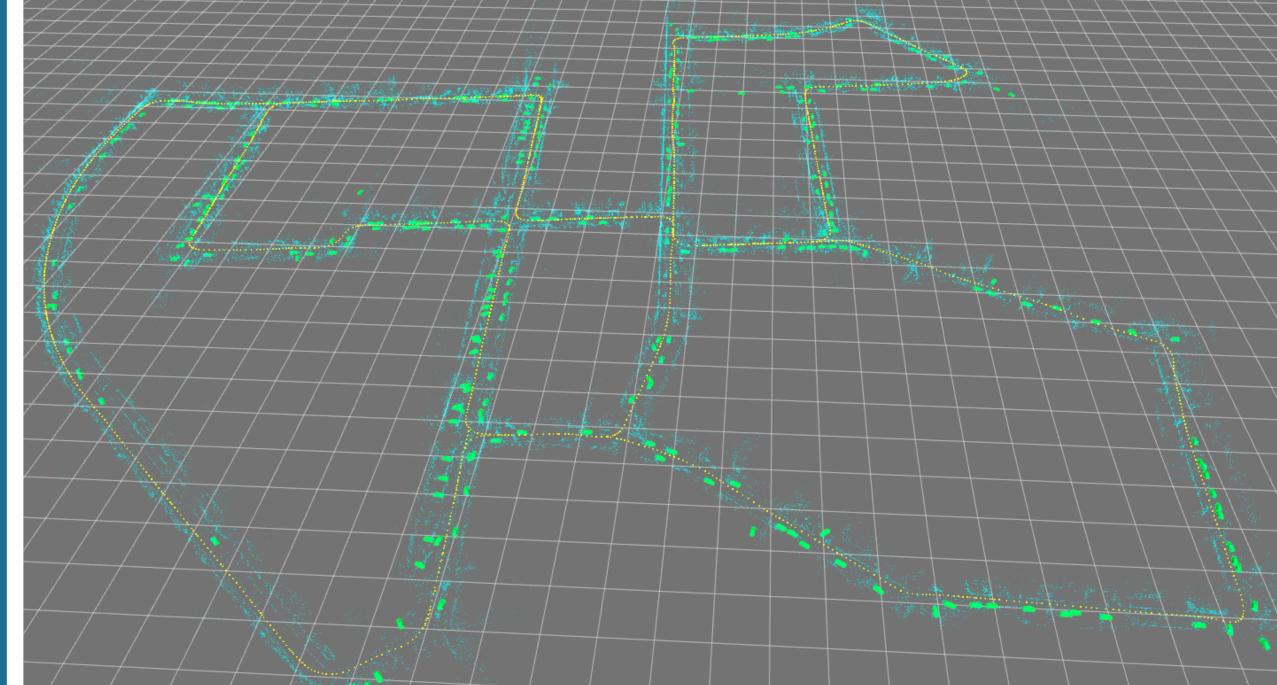
A chair is detected (a), and later becomes occluded (b, shown in dashed lines). Our system predicts its reappearance and resumes update (c).

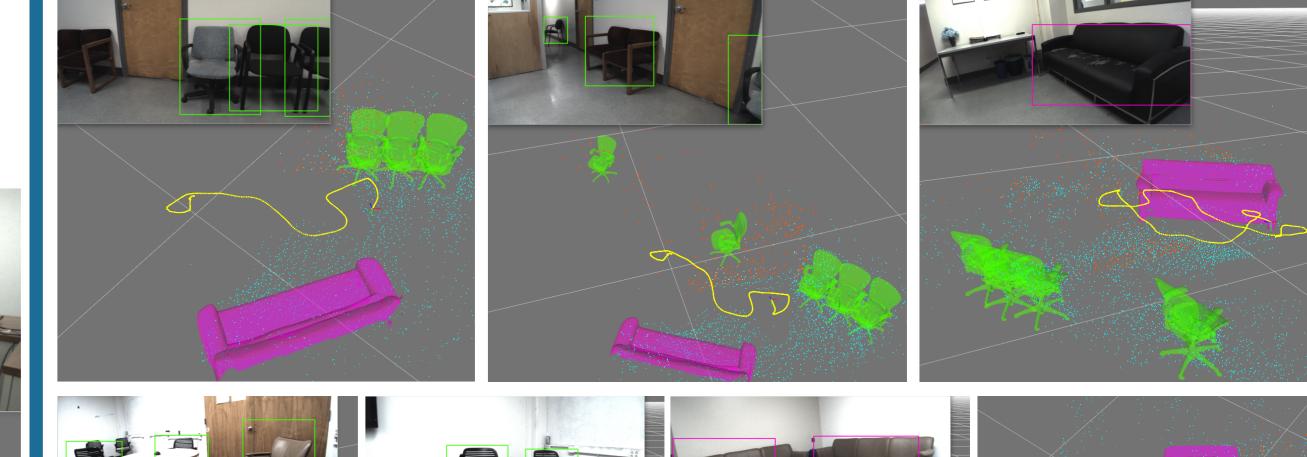
Real Scene Results

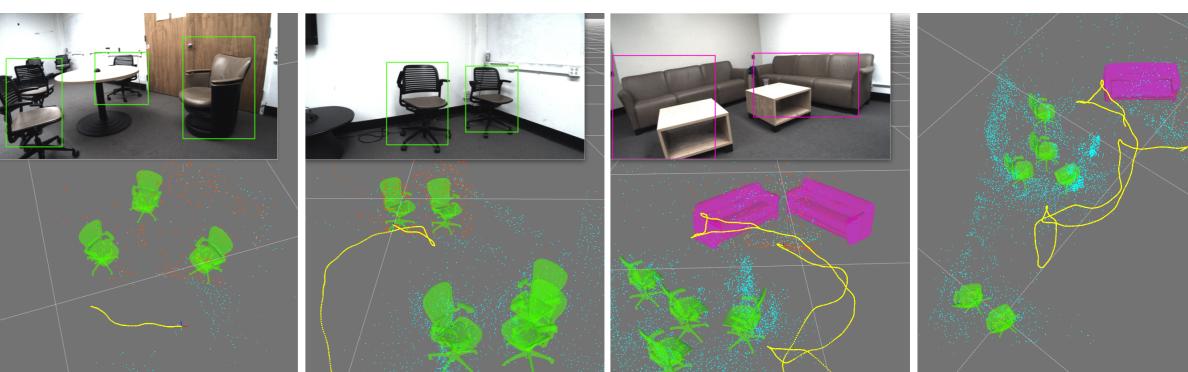
Outdoor Driving Sequence (3.7km)











Videos & Code available at http://vision.ucla.edu/vis

Research sponsored by ARO W911NF-15-1-0564/66731-CS, ONR N00014-17-1-2072, AFOSR FA9550-15-1-0229.