Semantic Perception under Uncertainty with Viewpoint-Dependent Models

Yuri Feldman

Under the supervision of Assoc. Prof. **Vadim Indelman**

Ph.D. Seminar, March 2022

Work partially supported by

THE ISRAELI SMART TRANSPORTATION RESEARCH CENTER

ANPL Autonomous Navigation and Perception Lab

TECHNION | The Henry and Marilyn Taub Faculty of Computer Science

Intro – Robot Autonomy

Key components:

Perception (Situational Awareness, Data Fusion)

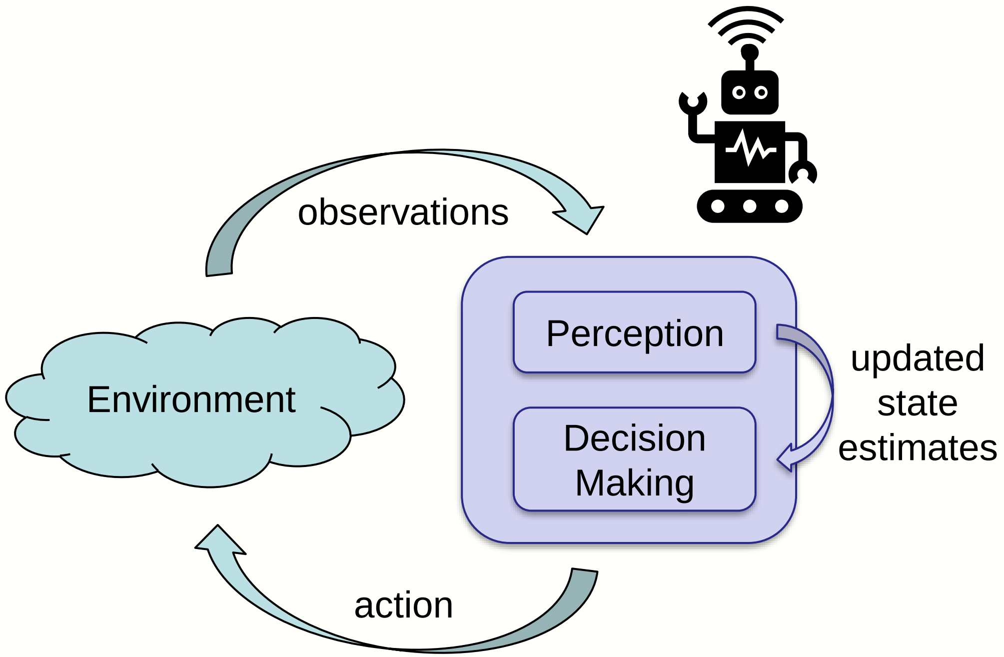
- Understanding of the environment and robot state within it

Decision Making

- Plan actions (towards task of interest)

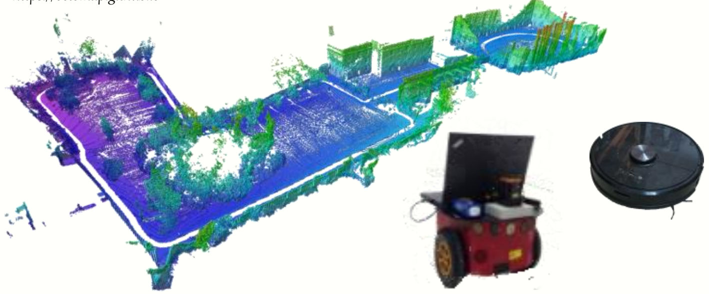
Need to deal with **uncertainty**

- Due to: noisy and aliased measurements, partial information, *imperfect models*…



Intro 2 – Semantic Perception

https://octomap.github.io



Geometric perception

⇨ established methods exist

SLAM – Simultaneous Localization and Mapping

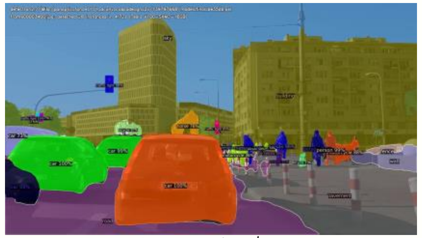
Semantic perception

⇨ required for less-structured tasks

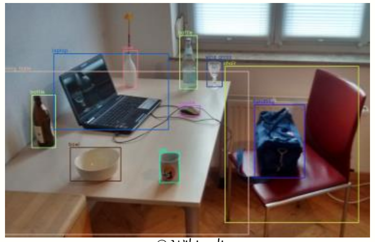
⇨ need resilience to per-frame errors for safe and reliable operation.

https://www.drive.com.au

https://www.cc.gatech.edu/projects/sting-racing



Detectron 2, Wu et al. 2019



© Wikipedia

Outline

Viewpoint-Dependent models for Semantic Perception under Uncertainty

1. The semantic perception problem (Object-Level SLAM)
2. Viewpoint-dependent semantic measurement models
3. <u>Contributions</u>:
    I. Classification under Model and Localization Uncertainty
    II. Data Association-Aware Semantic Mapping and Localization
    III. Semantic Perception with a Continuous Learned Representation

Intro 3 – (Geometric) Simultaneous Localization and Mapping (SLAM)

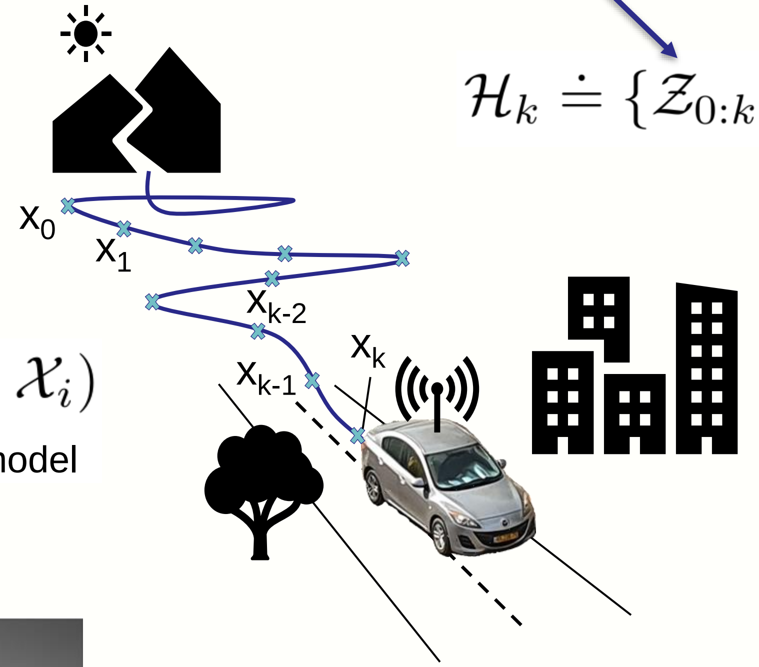
robot track landmarks

$$\mathcal{X}^*_{0:k}, \mathcal{L}^* = \underset{\mathcal{X},\mathcal{L}}{\arg\max}\ \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{L} \mid \mathcal{H}_k)$$

$$= \underset{\mathcal{X},\mathcal{L}}{\arg\max}\ \eta \cdot \underbrace{\mathbb{P}(\mathcal{X}_0)}_{\text{prior}} \prod_i \underbrace{\mathbb{P}(\mathcal{X}_{i+1} \mid \mathcal{U}_i, \mathcal{X}_i)}_{\text{motion model}} \prod_{j \in \mathcal{M}_i} \underbrace{\mathbb{P}(\mathcal{Z}_j \mid \mathcal{L}_j, \mathcal{X}_i)}_{\text{observation model}}$$

observations user controls

$$\mathcal{H}_k \doteq \{\mathcal{Z}_{0:k}, \mathcal{U}_{0:k-1}\}$$



See for example:
Kaess et al. 2008 TRO (iSAM)
Kaess et al. 2012 IJRR (iSAM2)
And related publications

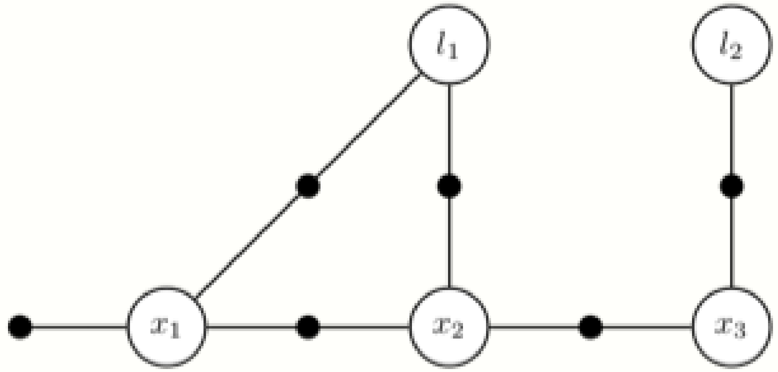
Intro 3 – (Geometric) Simultaneous Localization and Mapping (SLAM)

robot track landmarks

$$\mathcal{X}^*_{0:k}, \mathcal{L}^* = \underset{\mathcal{X},\mathcal{L}}{\arg\max}\ \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{L} \mid \mathcal{H}_k)$$

$$= \underset{\mathcal{X},\mathcal{L}}{\arg\max}\ \eta \cdot \underbrace{\mathbb{P}(\mathcal{X}_0)}_{\text{prior}} \prod_i \underbrace{\mathbb{P}(\mathcal{X}_{i+1} \mid \mathcal{U}_i, \mathcal{X}_i)}_{\text{motion model}} \prod_{j \in \mathcal{M}_i} \underbrace{\mathbb{P}(\mathcal{Z}_j \mid \mathcal{L}_j, \mathcal{X}_i)}_{\text{observation model}}$$

Factor graph:



© gtsam tutorial

Semantic Perception: Object-Level SLAM

(Salas-Moreno et al. 13’ CVPR, Choudhary et al. 14’ IROS, Bowman et al. 17’ ICRA, McCormack et al. 18 3dv, Nicholson 19’ ral, Yang 19’ TRO, …)

robot track

object categories (discrete!)

measurement and control history

$$\underset{\mathcal{X},\mathcal{C},\mathcal{O}}{\arg\max}\ \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{C}, \mathcal{O} \mid \mathcal{H}_k)$$

object geometry

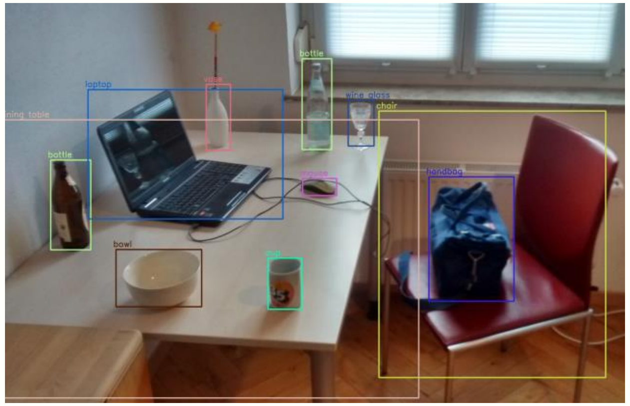
$$= \underset{\mathcal{X},\mathcal{C},\mathcal{O}}{\arg\max}\ \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O} \mid \mathcal{C}, \mathcal{H}_k) \cdot \mathbb{P}(\mathcal{C} \mid \mathcal{H}_k)$$

(continuous) hypothesis

hypothesis weight

$$c \in \{1, \ldots, N\}$$

$$|\{(c_1, \ldots, c_m)\}| = N^m$$



© Wikipedia

Semantic Perception: Object-Level SLAM

(Salas-Moreno et al. 13' CVPR, Choudhary et al. 14' IROS, Bowman et al. 17' ICRA, McCormack et al. 18 3dv, Nicholson 19' ral, Yang 19' TRO, …)

robot track

object categories (discrete!)

measurement and control history

$$\arg\max_{\mathcal{X},\mathcal{C},\mathcal{O}} \; \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{C}, \mathcal{O} \mid \mathcal{H}_k)$$

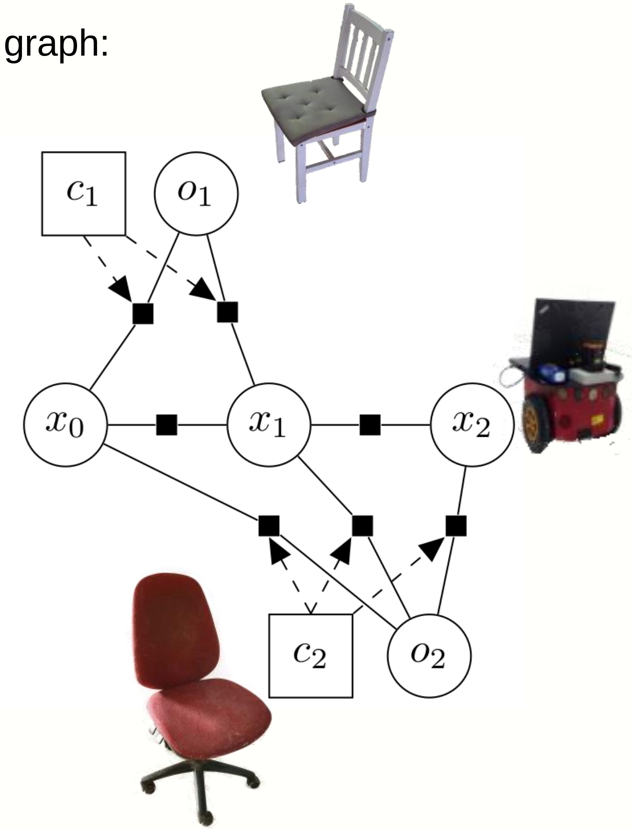
object geometry

$$= \arg\max_{\mathcal{X},\mathcal{C},\mathcal{O}} \; \underbrace{\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O} \mid \mathcal{C}, \mathcal{H}_k)}_{\text{(continuous) hypothesis}} \cdot \underbrace{\mathbb{P}(\mathcal{C} \mid \mathcal{H}_k)}_{\text{hypothesis weight}}$$

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O} \mid \mathcal{C}, \mathcal{H}_k) = \eta \cdot \mathbb{P}(\mathcal{X}_0) \prod_i \mathbb{P}(\mathcal{X}_{i+1} \mid \mathcal{U}_i, \mathcal{X}_i) \prod_{j \in \mathcal{M}_i} \mathbb{P}(\mathcal{Z}_j \mid \mathcal{O}_j, \mathcal{X}_i, \mathcal{C}_j)$$

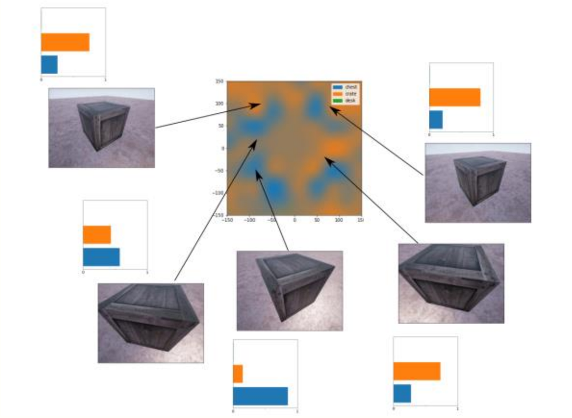
observation model (per-class)

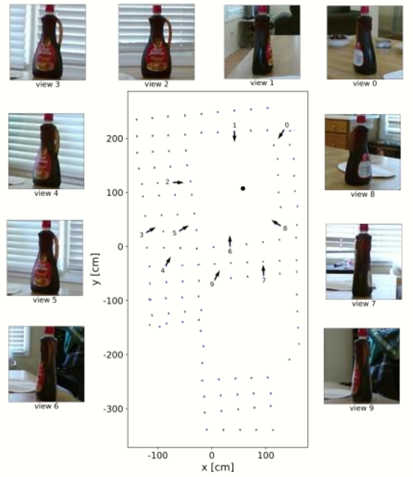
Factor graph:



Viewpoint-Dependent Models

- Viewpoint dependency ⇒ viewpoint-dependent models
  - Allow to couple semantics and geometry





current detection

object class

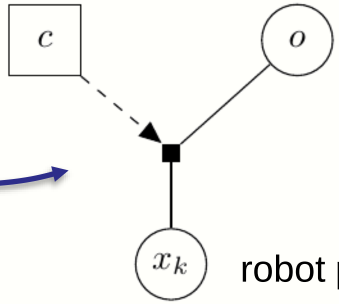
robot viewpoint relative to object

$$\mathcal{X}_k^{(rel)} \doteq \mathcal{X}_k \ominus o$$

$$\mathbb{P}(z_k \mid c, \mathcal{X}_k^{(rel)})$$

semantic factor

discrete category



6-dof object pose

robot pose

Yuri Feldman and Vadim Indelman. "Spatially-dependent Bayesian semantic perception under model and localization uncertainty." Autonomous Robots (2020): 1-29.

Outline

Viewpoint-Dependent models for Semantic Perception under Uncertainty

✓ ~~1. The semantic perception problem (Object-Level SLAM)~~

✓ ~~2. Viewpoint-dependent semantic measurement models~~

3. Contributions:
   I. Spatially-Dependent Classification under Model and Localization Uncertainty
   II. Data Association-Aware Semantic Mapping and Localization
   III. Semantic Perception with a Continuous Learned Representation

Spatially-Dependent Uncertainty-Aware Classification

"Bayesian viewpoint-dependent robust classification under model and localization uncertainty", Feldman & Indelman 18' ICRA
"Spatially-dependent Bayesian semantic perception under model and localization uncertainty", Feldman & Indelman 20' ARJ

object categories (discrete!)

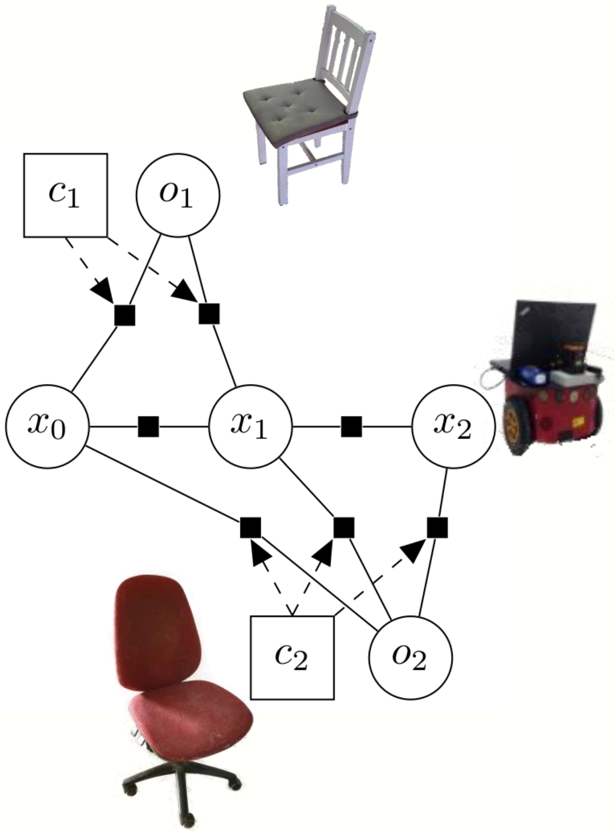
robot track

measurement and control history

$$\arg\max_{\mathcal{X},\mathcal{C},\mathcal{O}} \mathbb{P}(\mathcal{X}_{0:k},\mathcal{C},\mathcal{O}\mid\mathcal{H}_k)$$

object geometry

$$= \arg\max_{\mathcal{X},\mathcal{C},\mathcal{O}} \underbrace{\mathbb{P}(\mathcal{X}_{0:k},\mathcal{O}\mid\mathcal{C},\mathcal{H}_k)}_{\text{(continuous) hypothesis}}\cdot\underbrace{\mathbb{P}(\mathcal{C}\mid\mathcal{H}_k)}_{\text{hypothesis weight}}$$



Semantic Measurements – Spatial Correlation

Semantic measurements spatially correlated ⇒ not i.i.d.

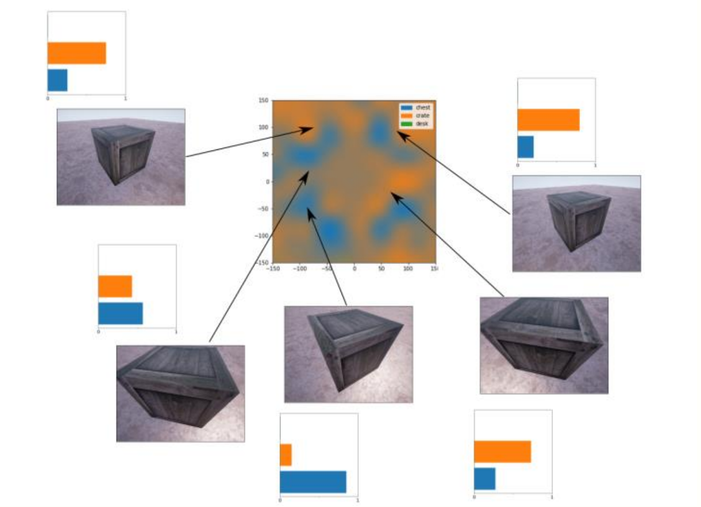
Would like to model the joint likelihood

object class

classifier responses

track of relative poses

$$\mathbb{P}(\mathcal{S}_{0:k} \mid c, X_{0:k}^{(rel)})$$



Can be done by fitting a Gaussian Process to classifier responses (offline training step)

$$s = f_c(x^{(rel)}) + \epsilon$$

$$f_c(x^{(rel)}) \sim \mathcal{GP}\left(\mu_c(x^{(rel)}), k_c(\cdot, \cdot)\right)$$

Semantic Measurements – Model Uncertainty

Problem: DNN output away from training set unstable

Solution (Gal & Ghahramani, 16' and others):

- ❖ marginalize over weight (model) uncertainty

Formally

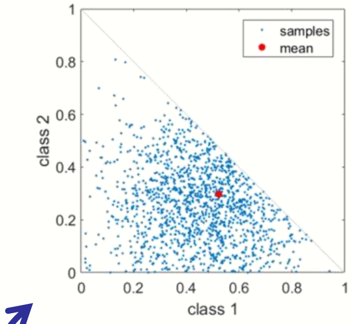
network output sample model uncertainty

training data

$$\mathbb{P}(c \mid \mathcal{D}) = \int \mathbb{P}(c \mid \mathcal{W}) \cdot \mathbb{P}(\mathcal{W} \mid \mathcal{D}) \, d\mathcal{W}$$

G&G 16': $\mathbb{P}(\mathcal{W} \mid \mathcal{D})\text{“} \approx \text{”}\, \mathrm{Bernoulli}\left(\frac{1}{2}\right) \cdot \hat{\mathcal{W}}$

Marginalization approximated via importance sampling.



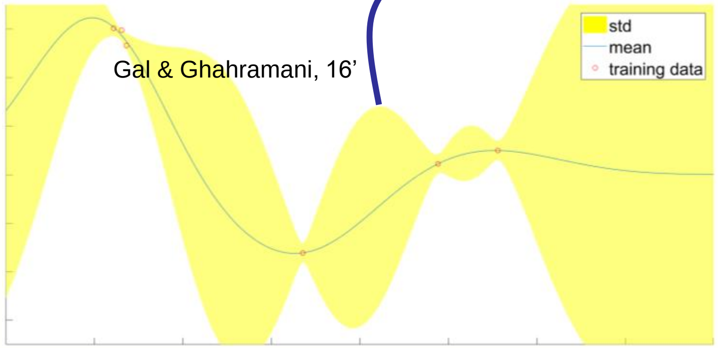


Image source: wikipedia

Approach – Uncertainty-Aware Classification

$$\mathbb{P}(c \mid \mathcal{H}_k) = \int_{\mathcal{X}_{0:k},o} \underbrace{\mathbb{P}(c \mid \mathcal{X}_{0:k}, o, \mathcal{H}_k)}_{(a)} \underbrace{\mathbb{P}(\mathcal{X}_{0:k}, o \mid \mathcal{H}_k)}_{(b)} \, d\mathcal{X}_{0:k} do$$

Marginalize over last classification

$$\frac{1}{n_k} \sum_{s_k \in \mathcal{S}_k} \mathbb{P}(c \mid s_k, H_k \setminus \{z_k\})$$

Bayes

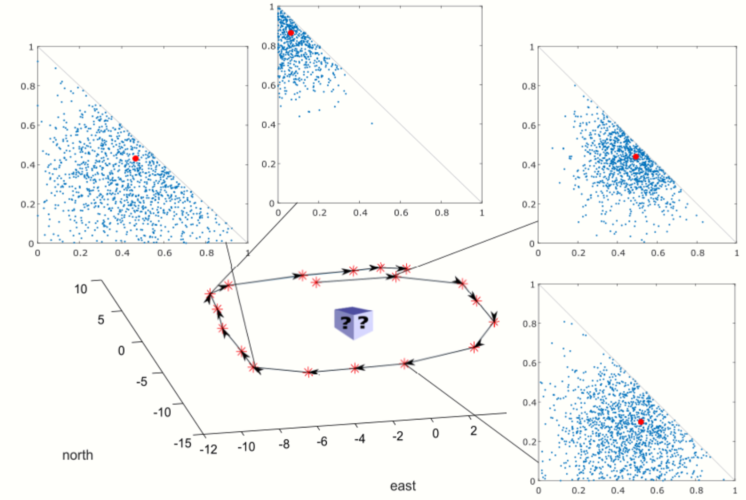
$$\frac{\boxed{\mathbb{P}(s_k \mid c, H_k \setminus \{z_k\})} \cdot \mathbb{P}(c \mid H_{k-1})}{\sum_{c \in \mathcal{C}} \mathbb{P}(s_k \mid c, H_k \setminus \{z_k\}) \cdot \mathbb{P}(c \mid H_{k-1})}$$

Marginalize over past classifications

$$\int_{\mathcal{S}_{0:k-1}} \underbrace{\mathbb{P}(s_k \mid c, \mathcal{S}_{0:k-1}, \mathcal{X}_{0:k}^{(rel)})}_{\text{Class model}} \cdot \prod_{i=0}^{k-1} \underbrace{\mathbb{P}(s_i \mid z_i)}_{\text{Model uncertainty}} \, d\mathcal{S}_{0:k-1}$$

Marginalize over Landmarks

$$\int_{\mathcal{L}} \mathbb{P}(\mathcal{X}_{0:k}, o, \mathcal{L} \mid \mathcal{H}_k) \, d\mathcal{L}$$



Results

Evaluated in synthetic simulation, 3D simulation, real-world data (BigBIRD, AVD).

Single object classification from measurements over a track. Localization uncertain but estimate available.

Evaluation criteria:

1. Probability of ground-truth class (higher is better) $\mathbb{P}(c^{GT} \mid \mathcal{H})$
2. Most-likely-to-ground-truth ratio (lower is better) $MGR \doteq \frac{\arg\max_c \; \mathbb{P}(c \mid \mathcal{H})}{\mathbb{P}(c^{GT} \mid \mathcal{H})}$

   ⇨ sensitive to confident misclassifications

Baselines:

- ❖ "Model Based" Teacy et al. 15' AAMAS – GP class model, assumes known localization
- ❖ Naïve Bayes (synthetic simulation) – directly fuses class predictions, no class model

Results – Synthetic Simulation

Statistics for several hand-specified scenarios over realizations of simulated classification.
3 candidate classes with hand-specified GP models, measurements from ground truth GP.

Model uncertainty benchmark

⇨ Simulated classification is randomly offset at each step.
Our method is input with uncertainty.

Localization uncertainty benchmark

⇨ Localization biased in a way that creates aliasing.
Our method is input with uncertainty.

Results – Synthetic Simulation – Model Uncertainty

$$\mathbb{P}(c_{GT} \mid \mathcal{H}) \qquad \log\left(\frac{\arg\max_c \ \mathbb{P}(c \mid \mathcal{H})}{\mathbb{P}(c_{GT} \mid \mathcal{H})}\right)$$

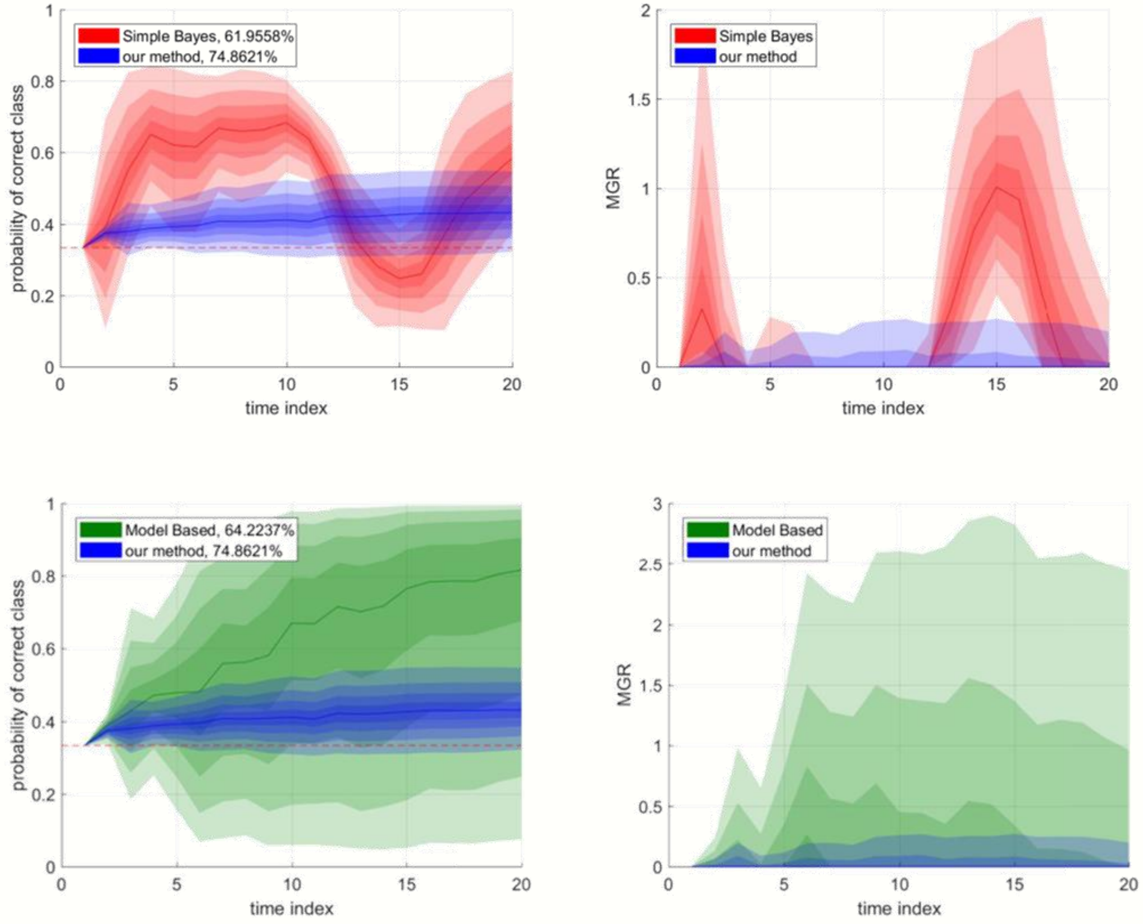
Color patches are equal-step (10%) percentiles.
Saturated line is median.
Legend lists % of steps with GT class most likely.

Results:

❖ Naïve Bayes arbitrarily off due to erroneous input.

❖ Model Based gives high scores to GT, but misclassifies in nearly 30% of the cases (MGR).

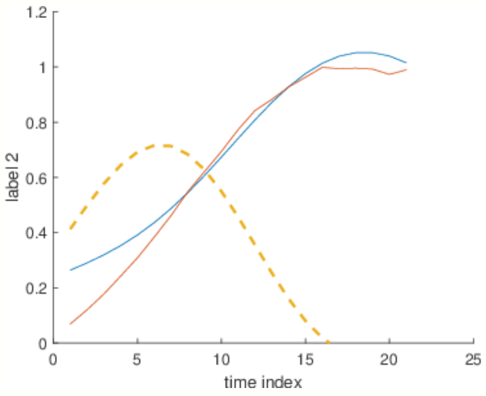
❖ Our method gracefully accumulates information, dense performance percentiles.



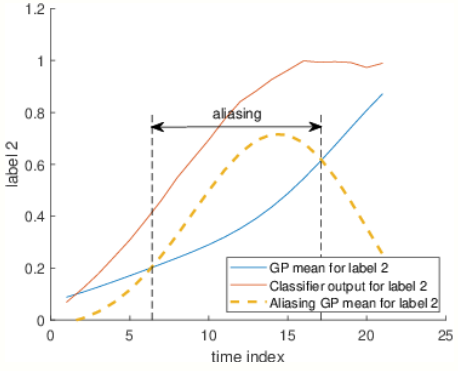
Results – Synthetic Simulation – Localization Uncertainty

- Model Based arbitrarily off due to aliasing.
- Naïve Bayes unaffected (classification measurements are correct).
- Our method gracefully accumulates information.

no localization bias

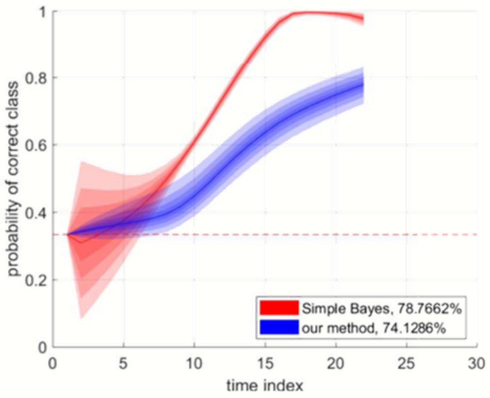


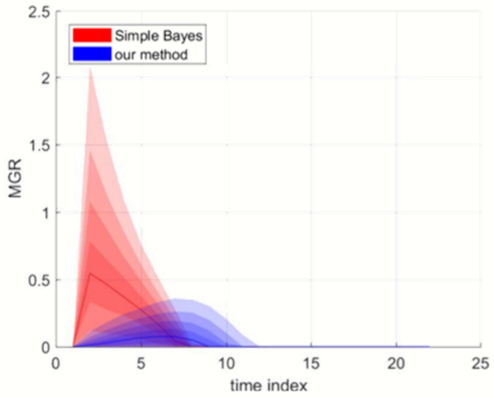
with localization bias

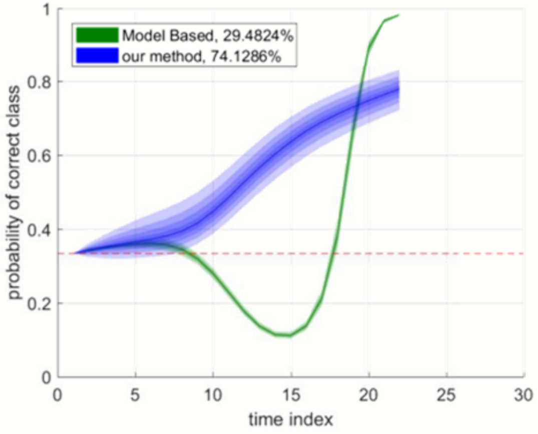


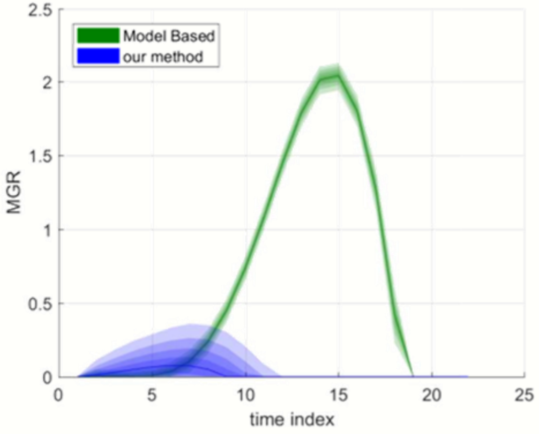
$$\mathbb{P}(c_{GT} \mid \mathcal{H})$$

$$\log\left(\frac{\arg\max_c \ \mathbb{P}(c \mid \mathcal{H})}{\mathbb{P}(c_{GT} \mid \mathcal{H})}\right)$$







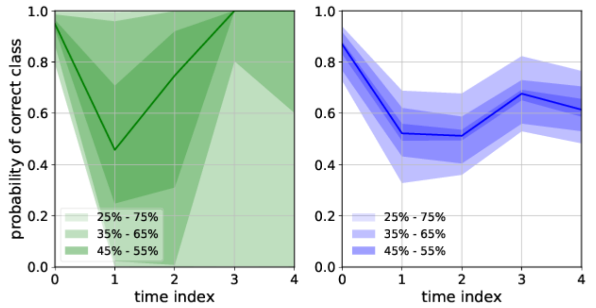


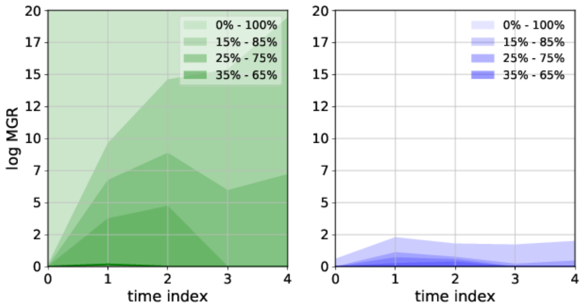
Results – 3D Simulation

Localization uncertainty:

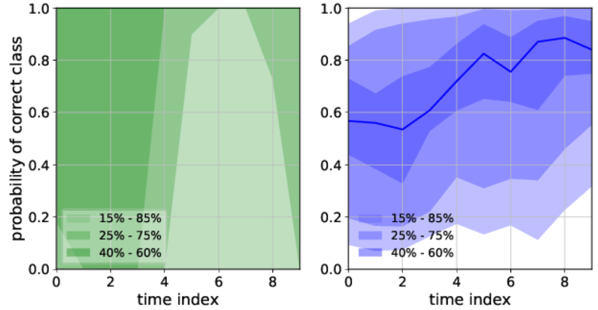
$$\mathbb{P}(c_{GT} \mid \mathcal{H})$$

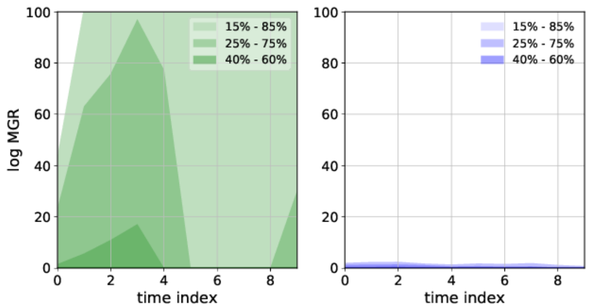
$$\log\left(\frac{\arg\max_c \; \mathbb{P}(c \mid \mathcal{H})}{\mathbb{P}(c_{GT} \mid \mathcal{H})}\right)$$



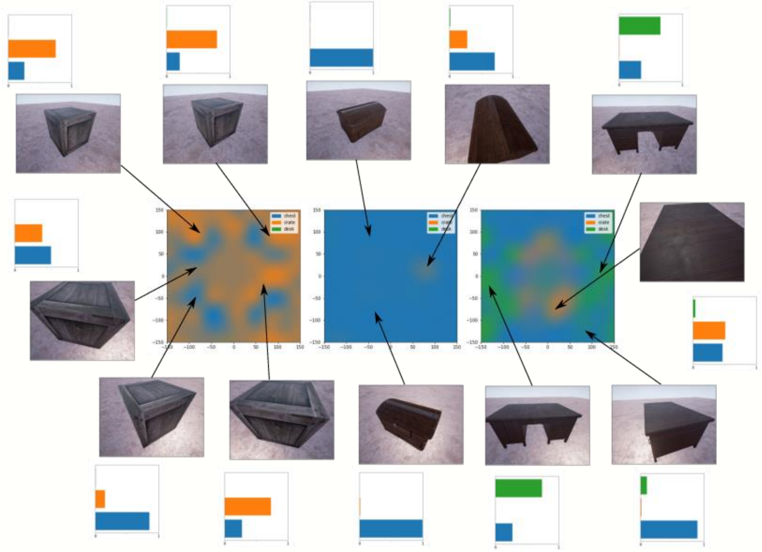


Model uncertainty:

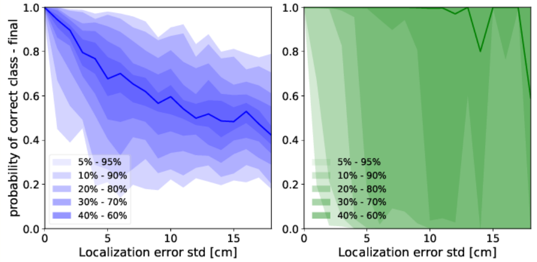


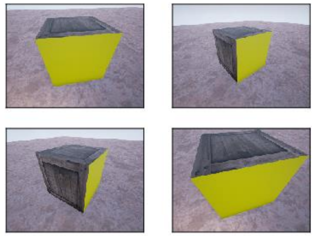


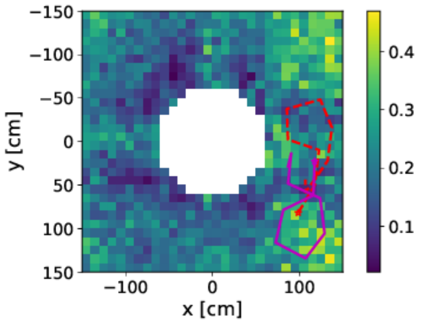
Learned GP models



sensitivity to localization error:







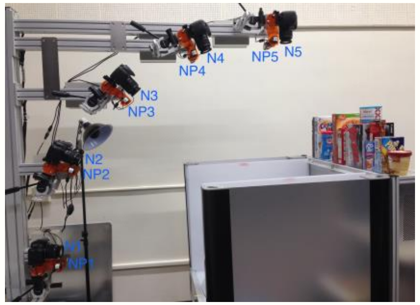
Results – Real Data

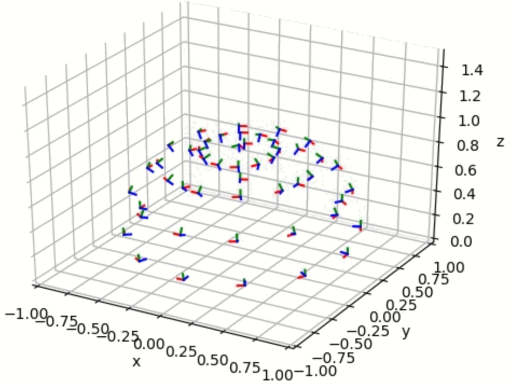
BigBIRD

(Big Berkeley Instance Recognition)

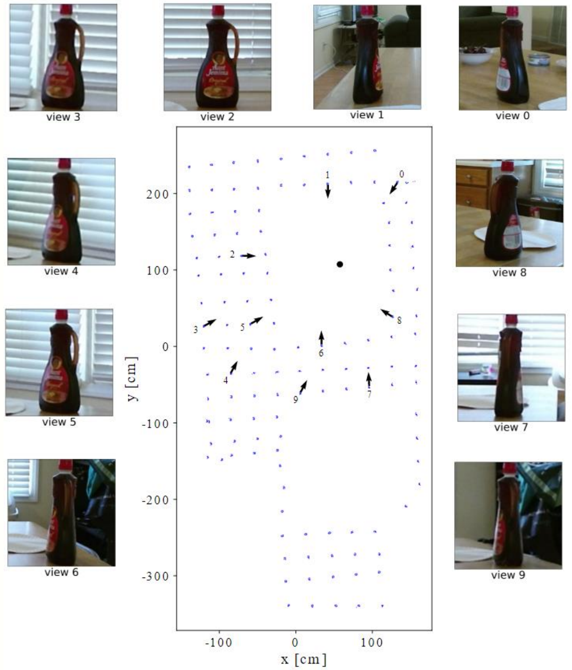
⇨ 125 objects

120 views / object





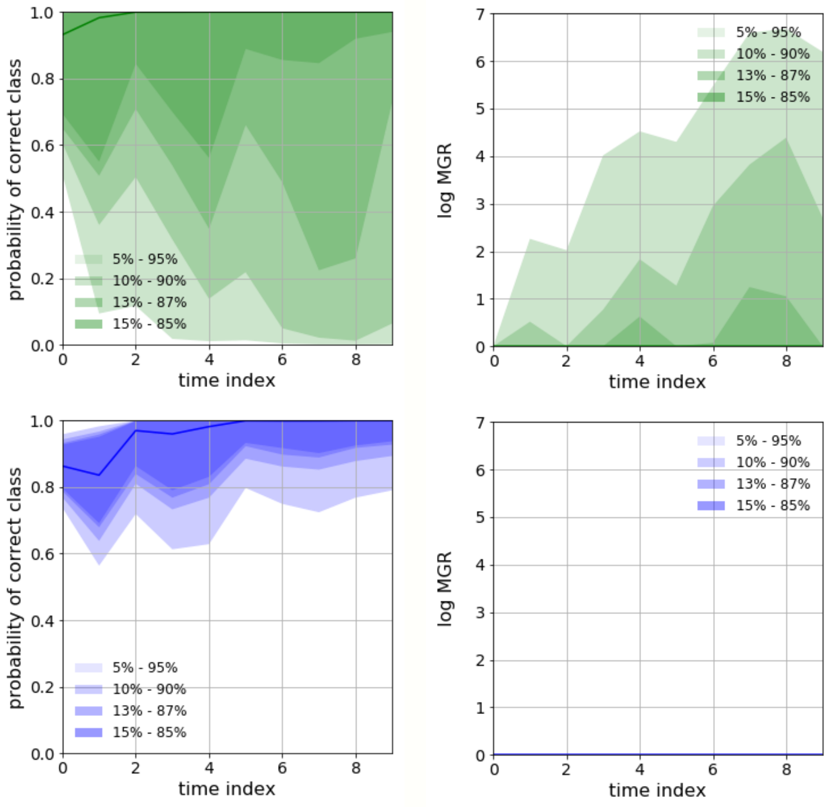
AVD (Active Vision Dataset)



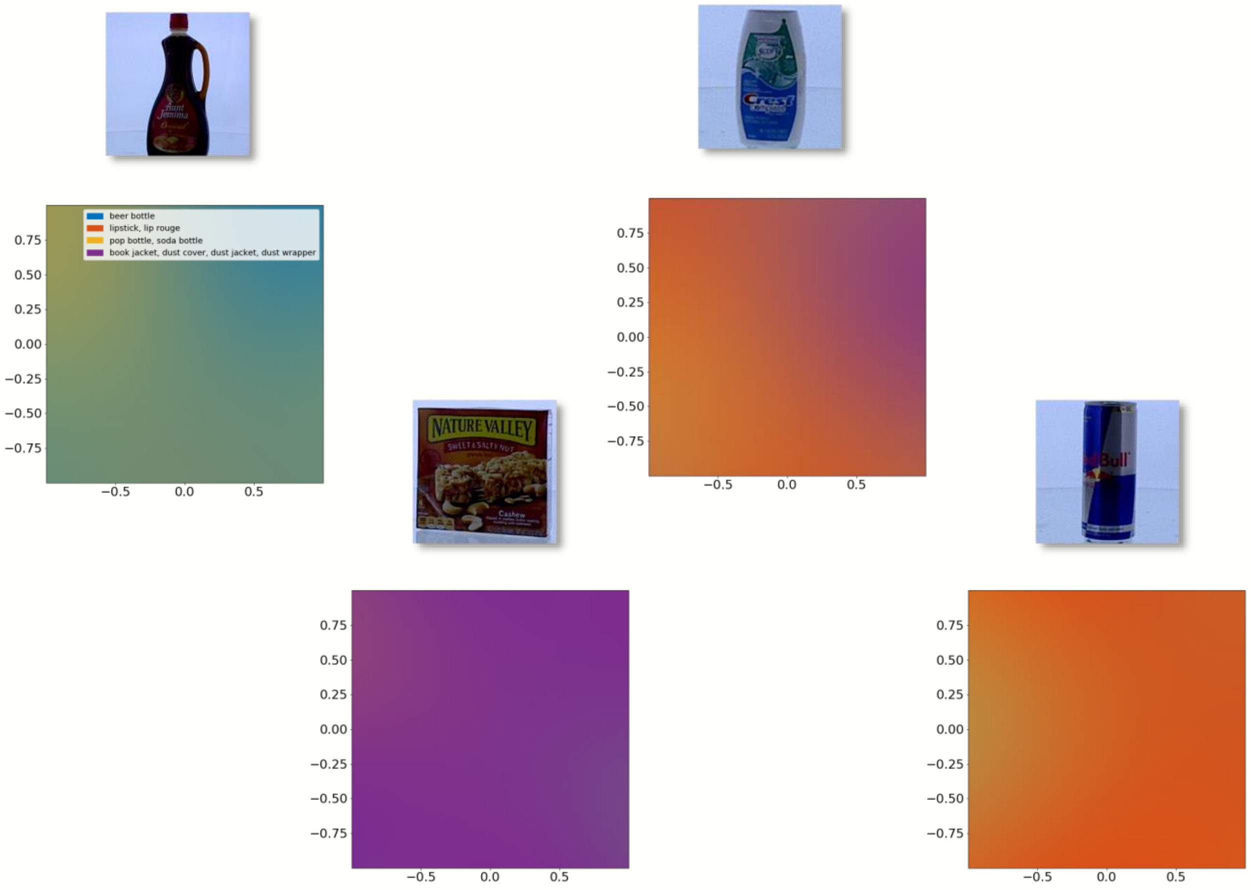
Results – Real Data

Localization uncertainty

$$\mathbb{P}(c_{GT} \mid \mathcal{H}) \qquad \log\left(\frac{\arg\max_c \ \mathbb{P}(c \mid \mathcal{H})}{\mathbb{P}(c_{GT} \mid \mathcal{H})}\right)$$



Example objects and learned GPs



Outline

Viewpoint-Dependent models for Semantic Perception under Uncertainty

✓ 1. ~~The semantic perception problem (Object-Level SLAM)~~

✓ 2. ~~Viewpoint-dependent semantic measurement models~~

3. Contributions:

    I. Classification under Model and Localization Uncertainty

    II. Data Association-Aware Semantic Mapping and Localization

    III. Semantic Perception with a Continuous Learned Representation

DA - Aware Semantic Mapping and Localization

"Data association aware semantic mapping and localization via a viewpoint-dependent classifier model", Tchuiev, Feldman and Indelman 19' IROS

Up until now: assumed data association solved.

Consider data association:

data association per measurement (discrete)

measurement and control history

$$\arg\max_{\mathcal{X},\mathcal{C},\mathcal{O},\beta} \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{C}, \mathcal{O}, \beta_{0:k} \mid \mathcal{H}_k)$$

robot track (continuous)

object categories (discrete)

object geometry (continuous)

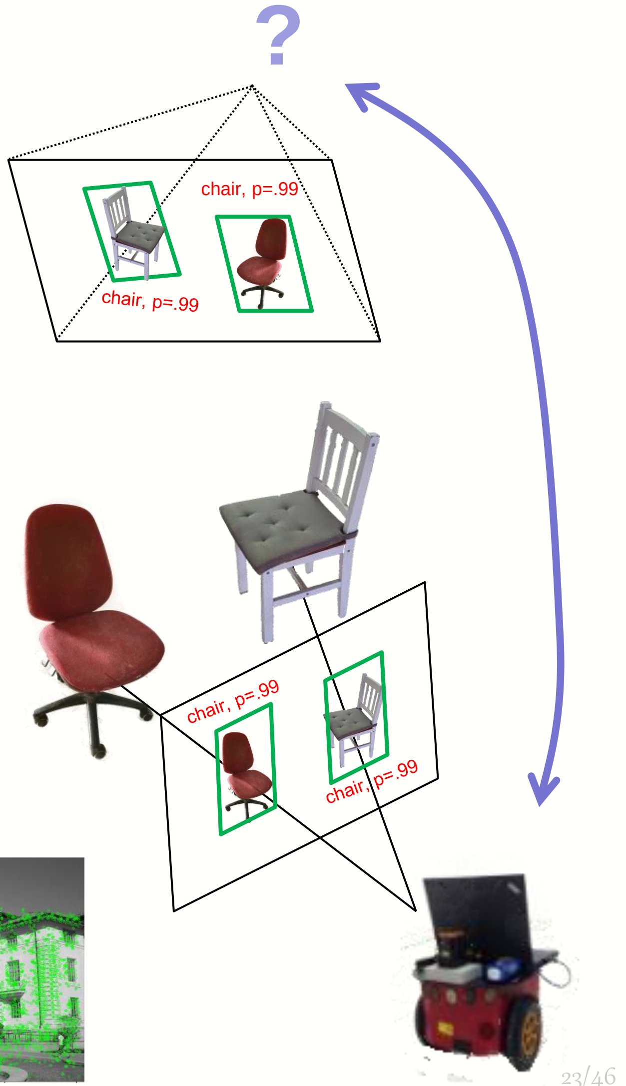
$$= \arg\max_{\mathcal{X},\mathcal{C},\mathcal{O},\beta} \underbrace{\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O} \mid \mathcal{C}, \beta_{0:k})}_{\text{(continuous) hypothesis}} \cdot \underbrace{\mathbb{P}(\mathcal{C}, \beta_{0:k} \mid \mathcal{H}_k)}_{\text{hypothesis weight}}$$

$$c \in \{1, \ldots, N\}$$

$$|\{(c_1, \ldots, c_m, \beta_1, \ldots, \beta_k)\}| = N^M \cdot m^k$$

…which is still better than …

© https://morioh.com/p/bd1b6fc9d9eb



DA - Aware Semantic Mapping and Localization

Contributions:

- DA-Aware Semantic Mapping through maintaining a full hybrid belief.
  - Maintain all plausible hypotheses until disambiguation is possible. Tractable in practice (subject to pruning).
- Viewpoint-Dependent model aids DA disambiguation by coupling between geometry and semantics.
- Approach operates with rich semantic feature vectors, not limited to most-likely-class measurements.

Approach - DA - Aware Semantic SLAM

Denote $b\left[\mathcal{X}_{0:k}, \mathcal{O}\right]^{\mathcal{C}}_{\beta_{1:k}} \doteq \underbrace{\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O} \mid \mathcal{C}, \beta_{1:k})}_{\text{(continuous) hypothesis}}$ $\quad w^{\mathcal{C}}_{\beta_{1:k}} \doteq \underbrace{\mathbb{P}(\mathcal{C}, \beta_{1:k} \mid \mathcal{H}_k)}_{\text{hypothesis weight}}$

hypothesis propagation $$b\left[\mathcal{X}_{0:k}, \mathcal{O}\right]^{\mathcal{C}}_{\beta_{1:k}} \propto b\left[\mathcal{X}_{0:k-1}, \mathcal{O}\right]^{\mathcal{C}}_{\beta_{1:k-1}} \cdot \underbrace{\mathbb{P}(\mathcal{X}_k \mid \mathcal{X}_{k-1}, \mathcal{A}_{k-1})}_{\text{motion model}} \cdot \underbrace{\mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k, \mathcal{O}_{\beta_k}, \mathcal{C})}_{\text{viewpoint-dependent model}}$$

weight update $$w^{\mathcal{C}}_{\beta_{1:k}} \propto w^{\mathcal{C}}_{\beta_{1:k-1}} \int_{\mathcal{X},\mathcal{O}} \underbrace{\mathbb{P}(\beta_k \mid \mathcal{X}_k, \mathcal{O}_{\beta_k})}_{\text{association probability}} \cdot b\left[\mathcal{X}_{0:k}, \mathcal{O}\right]^{\mathcal{C}}_{\beta_{1:k}} \, d\mathcal{X} d\mathcal{O}$$

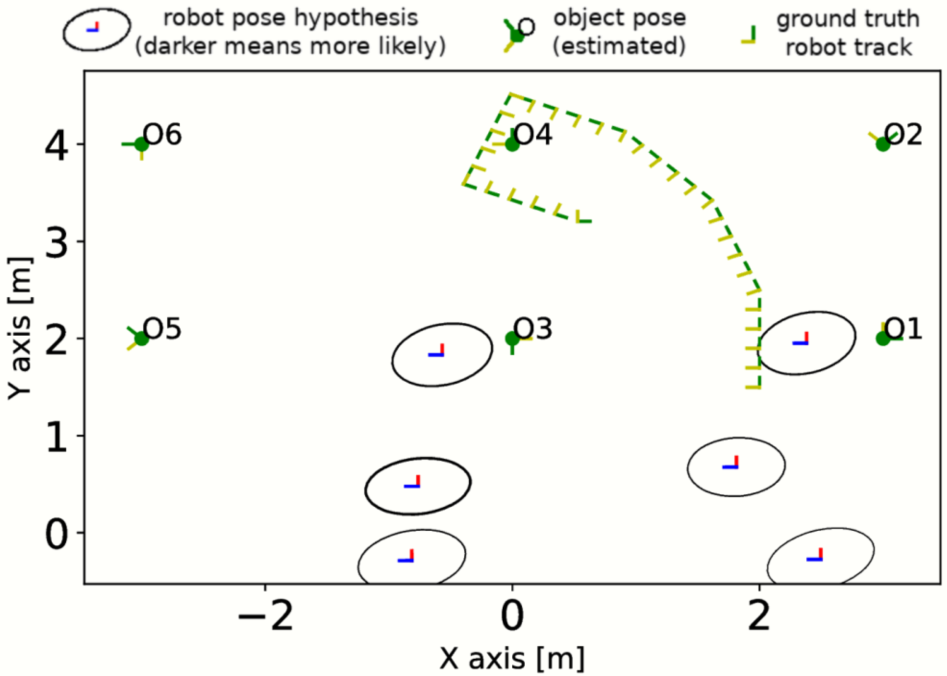
Weights that fall bellow a threshold are pruned

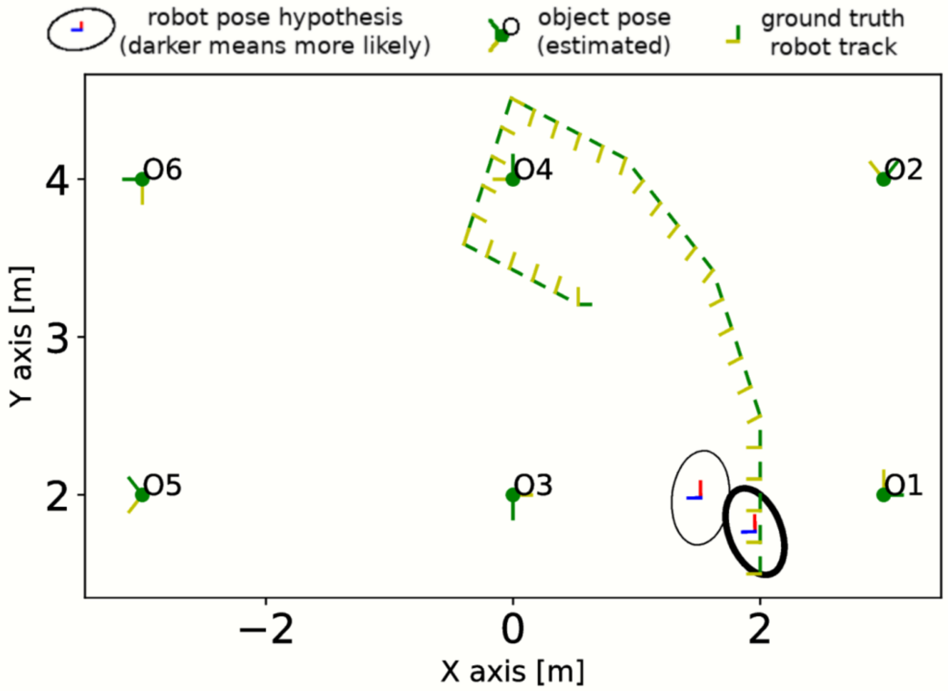
Results

Simulated environment:

- ❖ 6 identical objects
- ❖ 2 candidate classes with synthetic measurement models
- ❖ uninformative robot pose prior

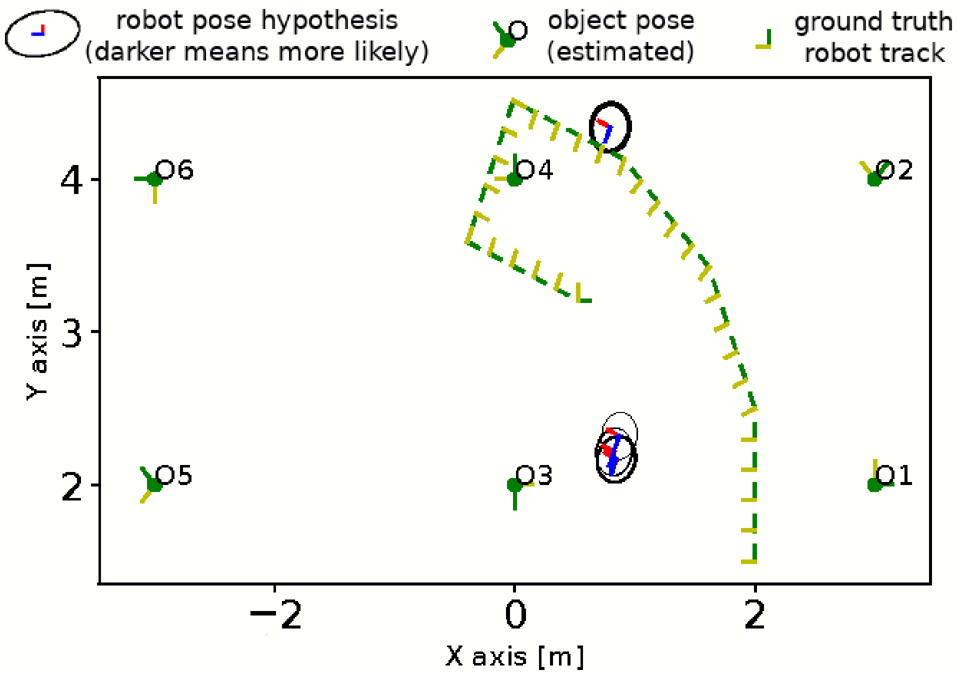
Time $k = 1$, without (left) and with (right) classifier model.

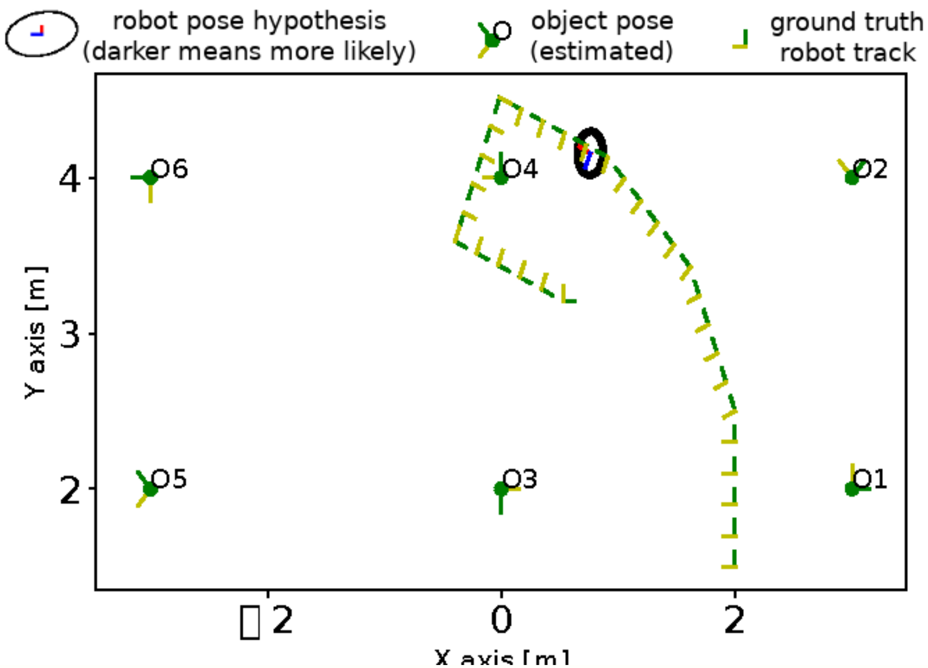




Results

Time $k = 15$ without (left) and with (right) classifier model



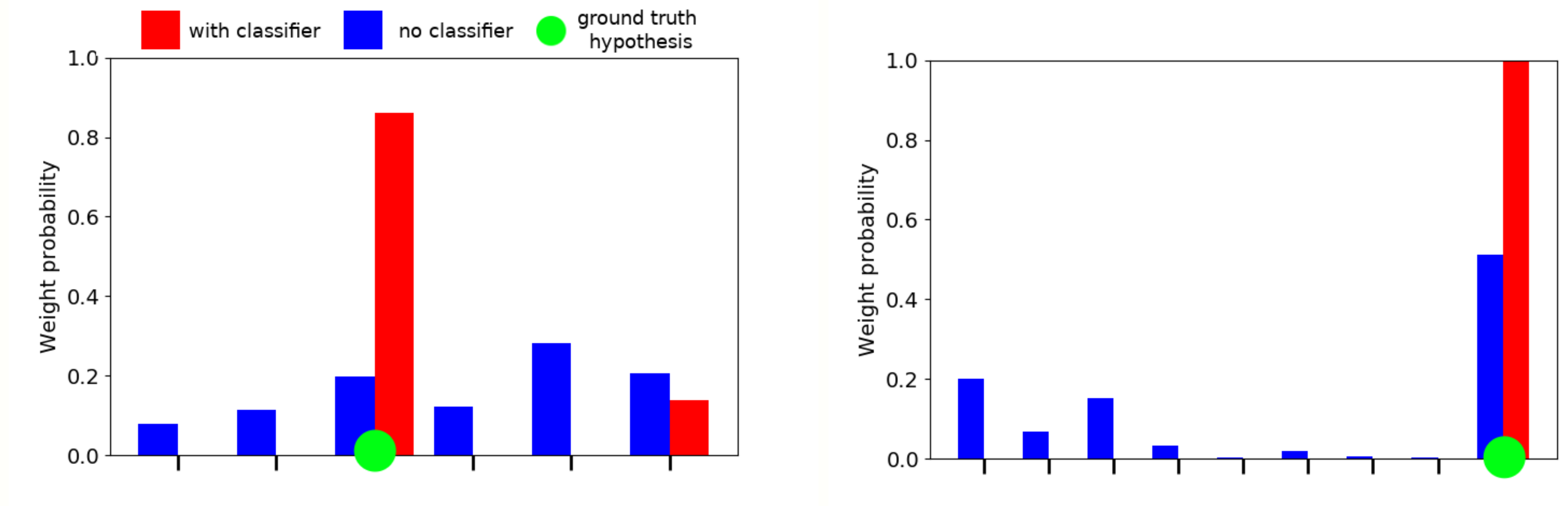


With classifier:

- ❖ Fewer hypotheses
- ❖ More accurate localization

Results

Hypothesis weight comparison, times $k = 1$ (left) and $k = 15$ (right)



With classifier:

- ❖ Fewer hypotheses
- ❖ Stronger disambiguation

Outline

Viewpoint-Dependent models for Semantic Perception under Uncertainty

✓ ~~1. The semantic perception problem (Object-Level SLAM)~~

✓ ~~2. Viewpoint-dependent semantic measurement models~~

3. Contributions:

    I. Classification under Model and Localization Uncertainty

    II. Data Association-Aware Semantic Mapping and Localization

    III. Semantic Perception with a Continuous Learned Representation

Semantic Perception with a Continuous Learned Representation

An initial version presented as “Towards Self-Supervised Semantic Representation with a Viewpoint-Dependent Observation Model” in proceedings of Workshop on Self-Supervised Robot Learning, in conjunction with RSS, July 2020

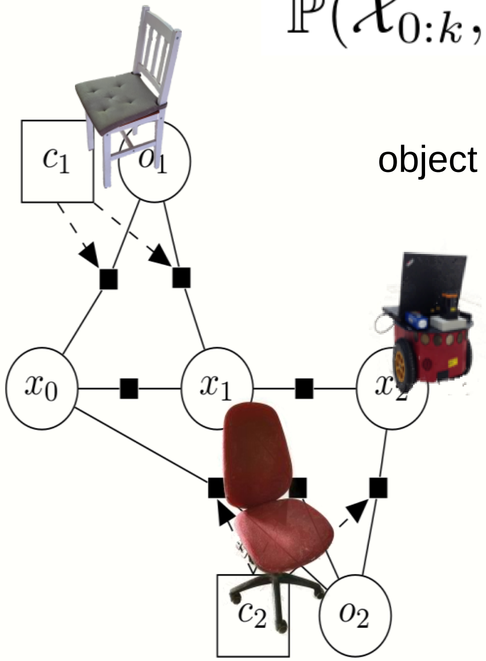
robot track

object categories (discrete)

measurement and control history

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{C}, \mathcal{O} \mid \mathcal{H}_k)$$

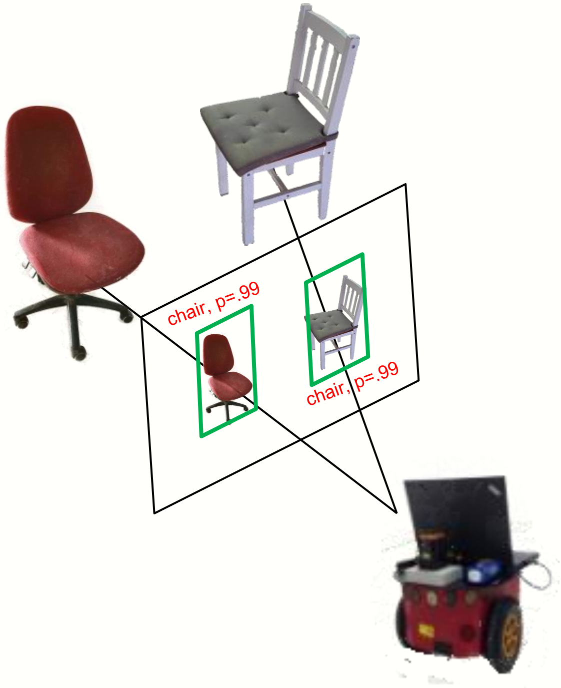
object geometry



⇨ Requires maintaining hypotheses (inefficient)

Requires per-class models

Limited granularity of semantic representation



Semantic Perception with a Continuous Learned Representation

robot track

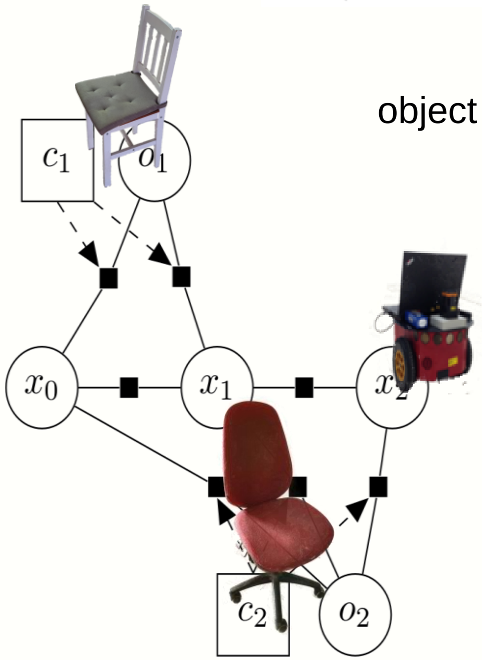
object categories (discrete)

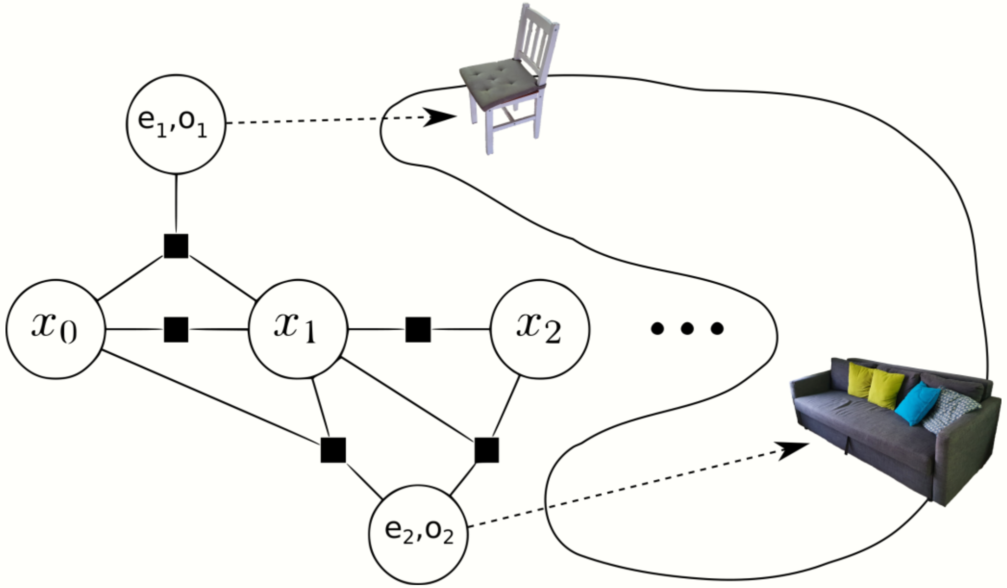
measurement and control history

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{C}, \mathcal{O} \mid \mathcal{H}_k) \implies \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{E}, \mathcal{O} \mid \mathcal{H}_k)$$

object geometry

continuous representation





Continuous Learned Representation – Inference

$$e_1, e_2 \in \mathbb{R}^n$$

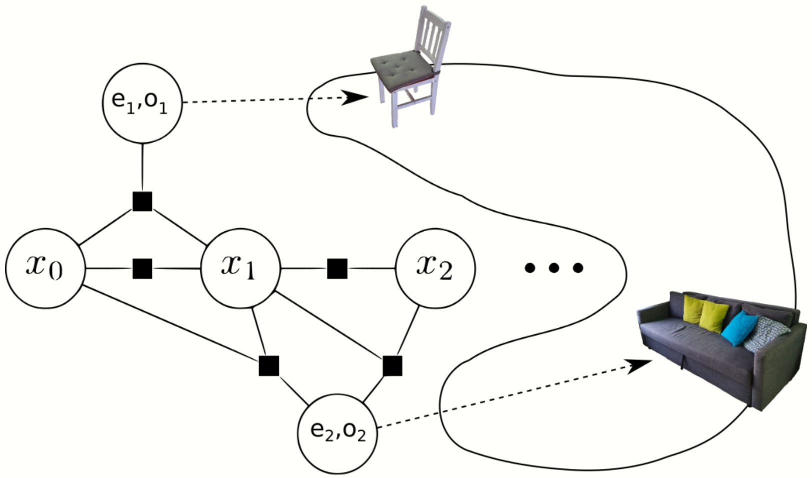
continuous representation

robot track

measurement and control history

maintain joint belief: $b_k \doteq \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{E}, \mathcal{O} \mid \mathcal{H}_k)$

object geometry



$$= \eta \cdot \underbrace{\mathbb{P}(z_k \mid \mathcal{X}_{0:k}, \mathcal{E}, \mathcal{O}, \mathcal{H}_k \setminus \{z_k\})}_{\text{semantic observation model}} \cdot \underbrace{\mathbb{P}(\mathcal{X}_k \mid \mathcal{X}_{0:k-1}, \mathcal{A}_{k-1})}_{\text{motion model}} \cdot b_{k-1}$$

⇨ Need a single semantic observation model (as opposed to per-class previously)

Continuous inference

No discretization on semantic representation

Fitting the Viewpoint-Dependent Model – Take 1

Assume a Gaussian viewpoint-dependent model conditioned on continuous representation:

frame

semantic description

$$\mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_{0:k}, \mathcal{E}, \mathcal{O}, \mathcal{H}_k \setminus \{z_k\}) \doteq \mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k^{(rel)}, \mathcal{E})$$

relative pose
to object

$$\doteq \mathcal{N}\left(\mathcal{Z}_k;\ \mu_\theta(\mathcal{X}_k^{(rel)}, \mathcal{E}), \Sigma\right)$$

Use Maximum-a-Posteriori to fit $\mathcal{E}, \theta$

observations

representation variables
for n objects

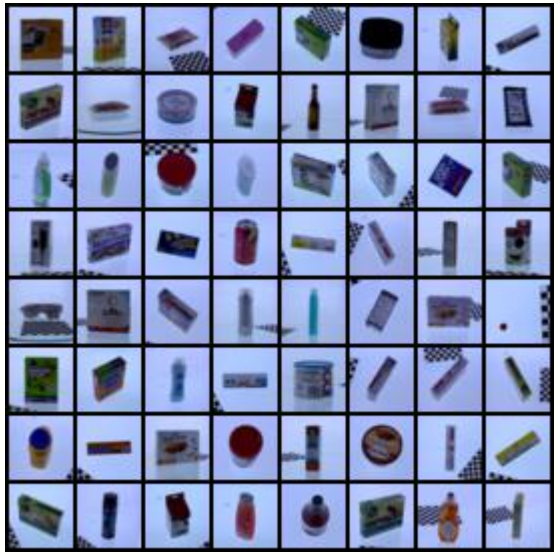
corresponding
relative poses

data association

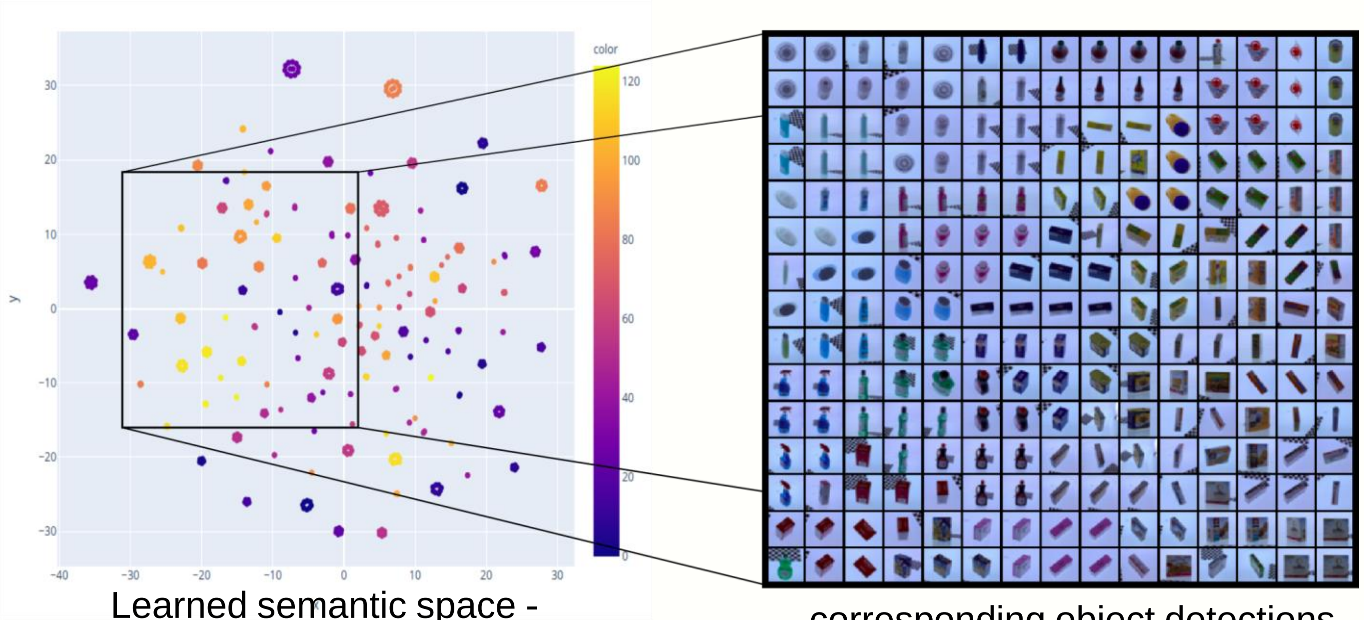
$$\arg\max_{\theta, \mathcal{E}_{1:n}} \mathbb{P}\left(\mathcal{Z}_{0:k}, \mathcal{E}_{1:n} \mid \mathcal{X}_{0:k}^{(rel)}, \beta_{0:k}\right)$$

$$= \arg\max_{\theta, \mathcal{E}_{1:n}} \sum \underbrace{\log \mathbb{P}_\theta\left(\mathcal{Z}_i \mid \mathcal{E}_{\beta_i}, \mathcal{X}_i^{(rel)}\right)}_{\text{viewpoint - dependent observation model}} + \underbrace{\log \mathbb{P}(\mathcal{E}_{1:n})}_{\text{prior on representation vectors}}$$

Take 1 - Results for Fitting $\mathbb{P}_\theta\left(\mathcal{Z} \mid \mathcal{E}, \mathcal{X}^{(rel)}\right)$



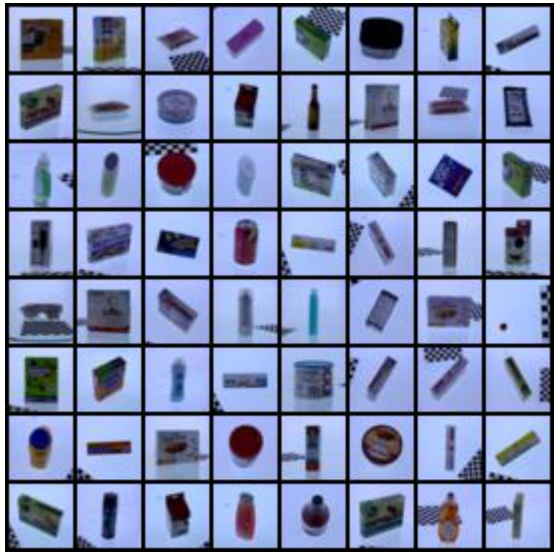
example images



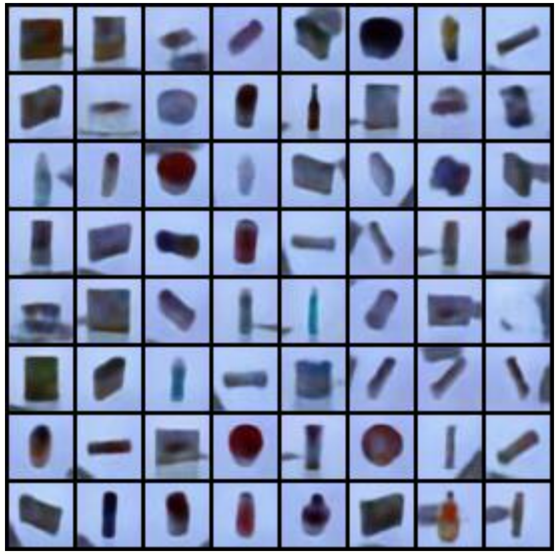
Learned semantic space - tSNE

corresponding object detections

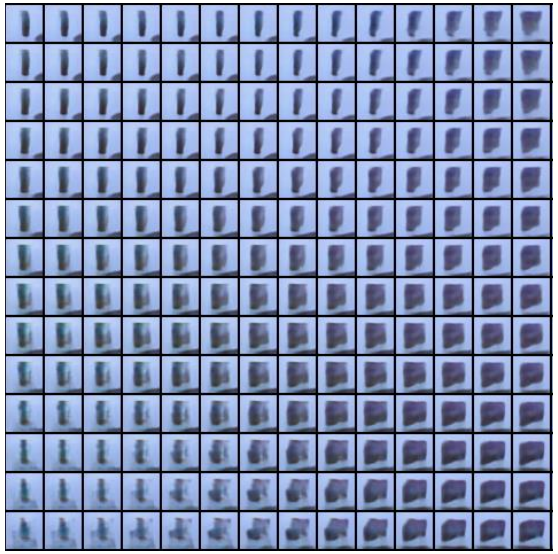
Take 1 - Results for Fitting $\mathbb{P}_\theta\left(\mathcal{Z} \mid \mathcal{E}, \mathcal{X}^{(rel)}\right)$



example images



mean predictions
(at ground truth)

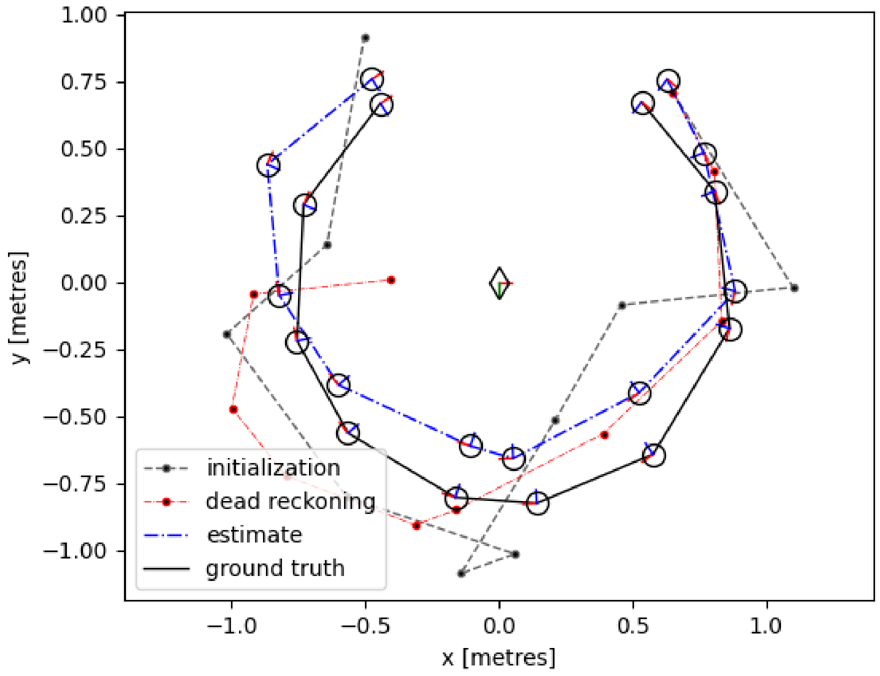


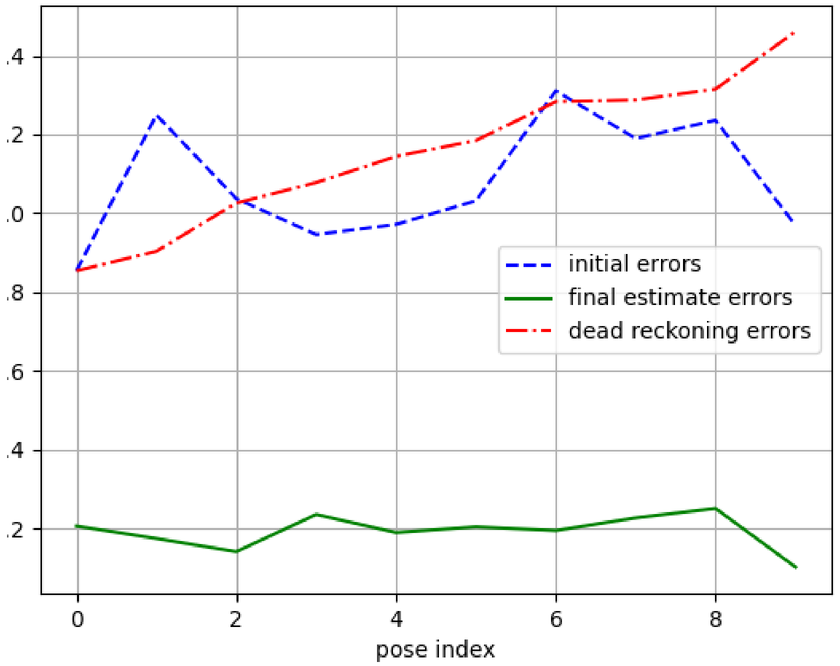
predictions for varying X
(around ground truth)

Take1 - Inference Using the Model

Simulation: use frames from viewpoints along a simulated track

- ❖ Also using odometry





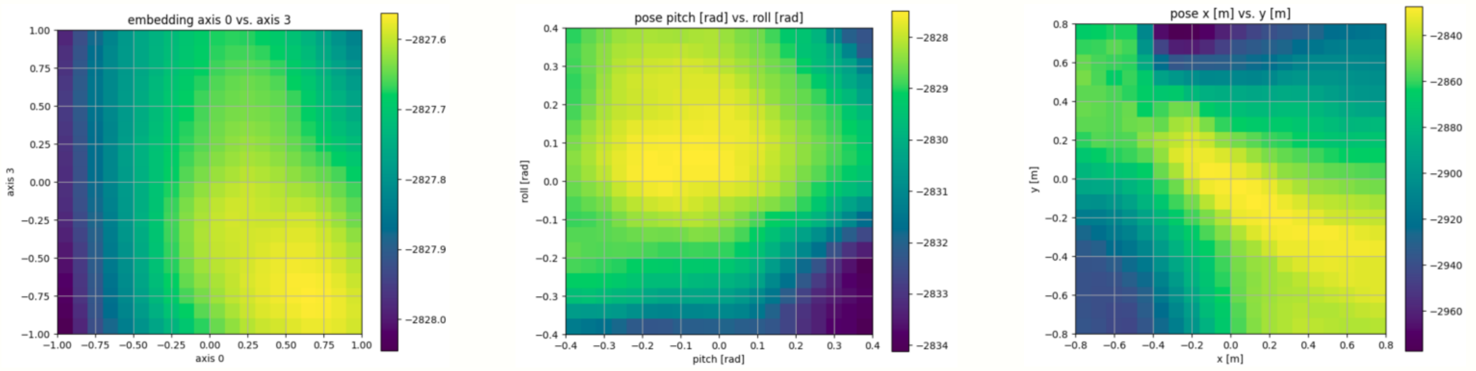
example track frames

Take 1 - Limitations

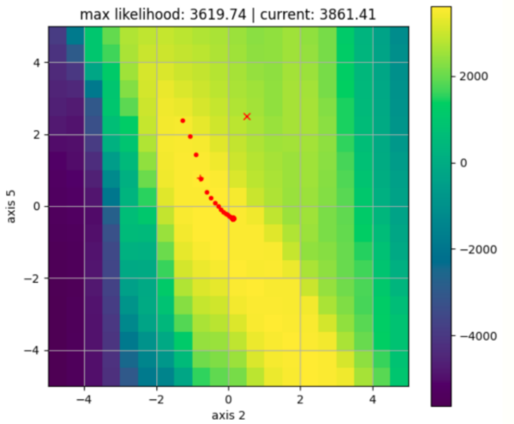
- Factor is huge (32×32 frame $\Longrightarrow$ 1024 Jacobian rows / keyframe!)
- Model expressiveness?
- <u>Maximum likelihood does not provide sufficient gradients for optimization</u>

$$\underset{\theta, \mathcal{E}_{1:n}}{\arg\max} \underbrace{\sum \log \mathbb{P}_\theta \left(\mathcal{Z}_i \mid \mathcal{E}_{\beta_i}, \mathcal{X}_i^{(rel)} \right)}_{\text{only constraints at training examples!}} + \log \mathbb{P}(\mathcal{E}_{1:n})$$

Values of likelihood $\mathbb{P}_\theta \left(\mathcal{Z} \mid \mathcal{E}, \mathcal{X}^{(rel)} \right)$ around ground truth point:



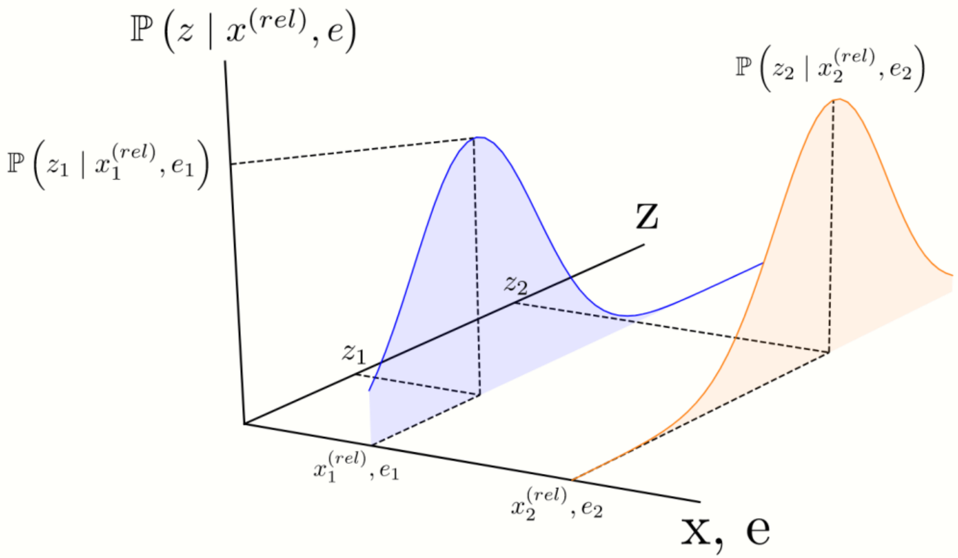
Example optimization path:

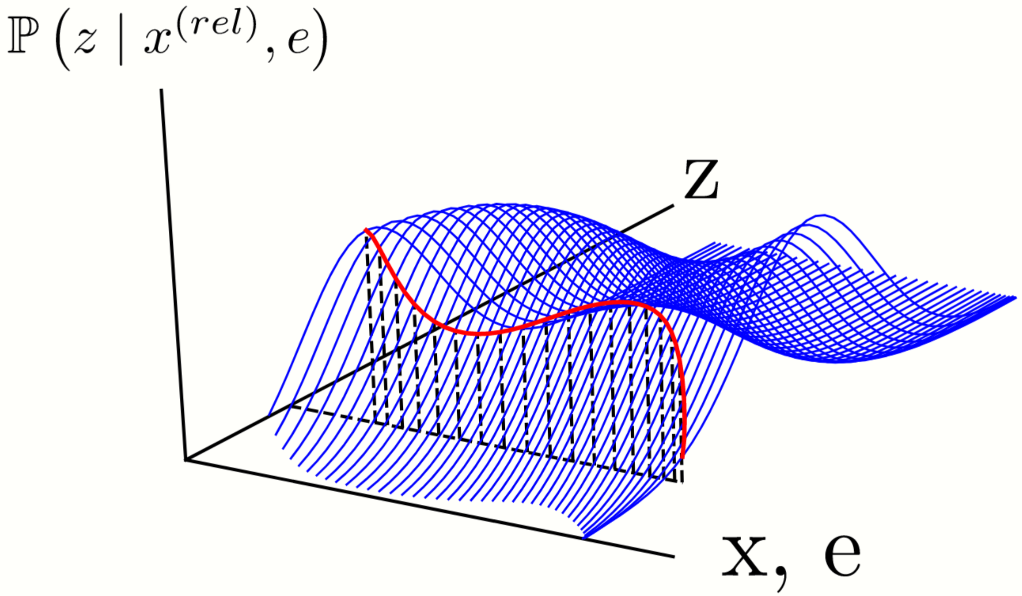


A Closer Look…

The learned model is constrained on a sparse set of triplets $\mathcal{Z}, \mathcal{X}^{(rel)}, \mathcal{E}$

⇨ Can try to improve inference by shaping the model elsewhere





Fitting the Viewpoint-Dependent Model – Take 2 (a)

Use a feature extractor to reduce factor dimensionality

$$\mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k^{(rel)}, \mathcal{E}) \doteq \mathbb{P}_\theta\left(f_\psi\left(\mathcal{Z}_k\right) \mid \mathcal{X}_k^{(rel)}, \mathcal{E}\right)$$

$$\doteq \mathcal{N}\left(f_\psi\left(\mathcal{Z}_k\right);\ \mu_\theta(\mathcal{X}_k^{(rel)}, \mathcal{E}), \Sigma\right)$$

In practice we use $\dim(\mathcal{E}) = 16 \quad \dim\left(f_\psi\left(\mathcal{Z}_k\right)\right) = 12$.

Jointly fit all parameters

data term

representation regularization / prior

feature extractor regularization

$$\underset{\theta,\psi,\mathcal{E}_{1:n}}{\arg\min}\ J\left(\theta, \psi, \mathcal{E}_{1:n}\right) = \underset{\theta,\psi,\mathcal{E}_{1:n}}{\arg\min}\ J_d\left(\theta, \psi, \mathcal{E}_{1:n}\right) + J_r\left(\mathcal{E}_{1:n}\right) + J_f(\psi)$$

Fitting the Viewpoint-Dependent Model – Take 2 (b)

Data term

$$\underbrace{J_d^{(i)}\left(\theta,\psi,\mathcal{E}_{1:n}\right)}_{\text{data term for training example (i)}} = \mathop{\mathbb{E}}_{\Delta\mathcal{X}^{(rel)},\Delta\mathcal{E}}\left\{\frac{\left\|\left[\Delta\mathcal{X}^{(rel)},\Delta\mathcal{E}\right]+s\left(\mathcal{X}^{(rel)}+\Delta\mathcal{X}^{(rel)},\mathcal{E}+\Delta\mathcal{E}\right)\right\|^2}{\left\|\left[\Delta\mathcal{X}^{(rel)},\Delta\mathcal{E}\right]\right\|^2}\right\}$$

where s is the optimization step at $\mathcal{X}^{(rel)}+\Delta\mathcal{X}^{(rel)},\mathcal{E}+\Delta\mathcal{E}$ (that will be used at inference)

For gradient descent

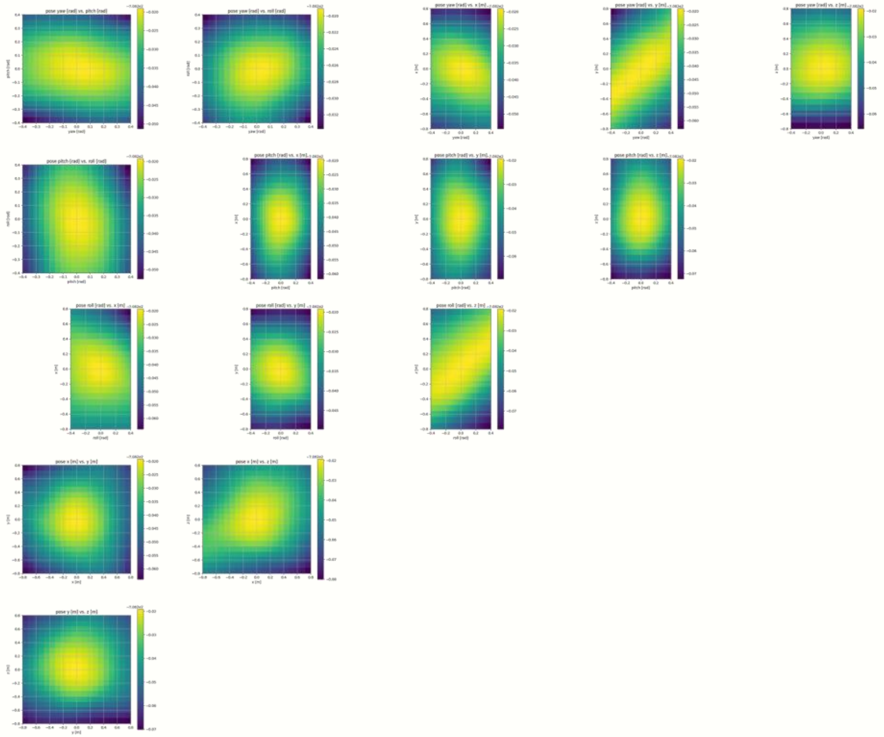
$$s^{GD} \doteq -\frac{\partial}{\partial\mathcal{X}^{(rel)},\mathcal{E}}\log\mathbb{P}_\theta\left(\mathcal{Z}\mid\mathcal{X}^{(rel)}+\Delta\mathcal{X}^{(rel)},\mathcal{E}+\Delta\mathcal{E}\right)$$

Learned Model

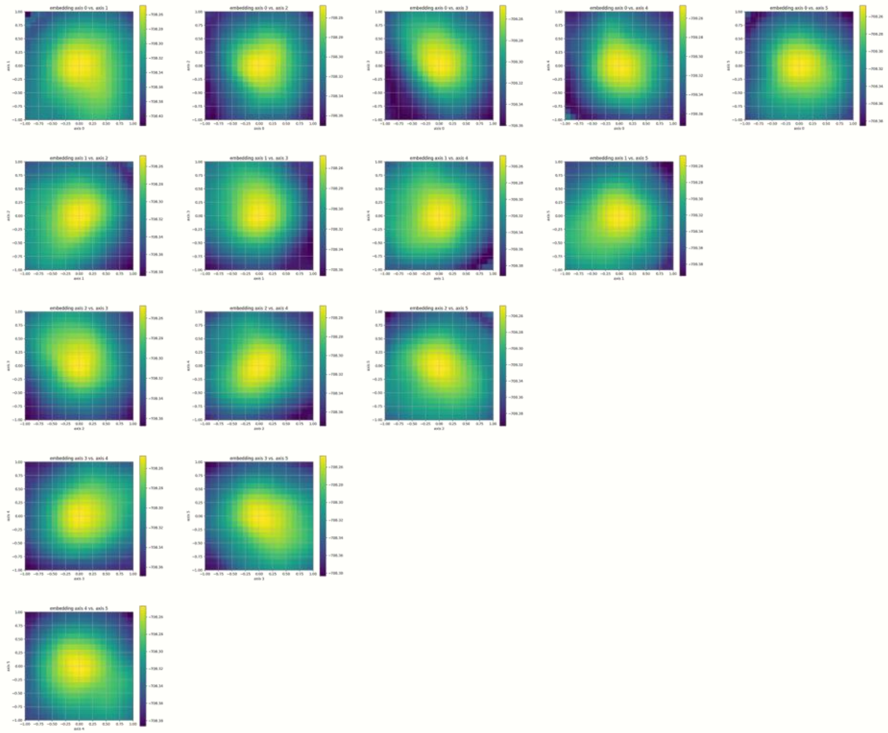
This finally gives useful gradients with respect to both $\mathcal{X}^{(rel)}$ and $\mathcal{E}$

Inference experiments still in progress.

likelihood - offsets in relative pose



likelihood - offsets in semantic representation



Outline

Viewpoint-Dependent models for Semantic Perception under Uncertainty

✓ ~~1. The semantic perception problem (Object-Level SLAM)~~

✓ ~~2. Viewpoint-dependent semantic measurement models~~

✓ ~~3. Contributions:~~

   I. Classification under Model and Localization Uncertainty

   II. Data Association-Aware Semantic Mapping and Localization

   III. Semantic Perception with a Continuous Learned Representation

4. Summary

Summary

❖ We showed how to address semantic perception under uncertainty by exploiting the coupling between semantics and geometry provided by viewpoint-dependent models.

Contributions:

I. Classification aware of Model uncertainty and correlations among viewpoints

II. Data Association-Aware Semantic Mapping and Localization

III. A novel approach to semantic SLAM through inference in a learned latent space

Thanks for listening!

Questions?