SLAM-AWARE, SELF-SUPERVISED PERCEPTION IN MOBILE ROBOTS

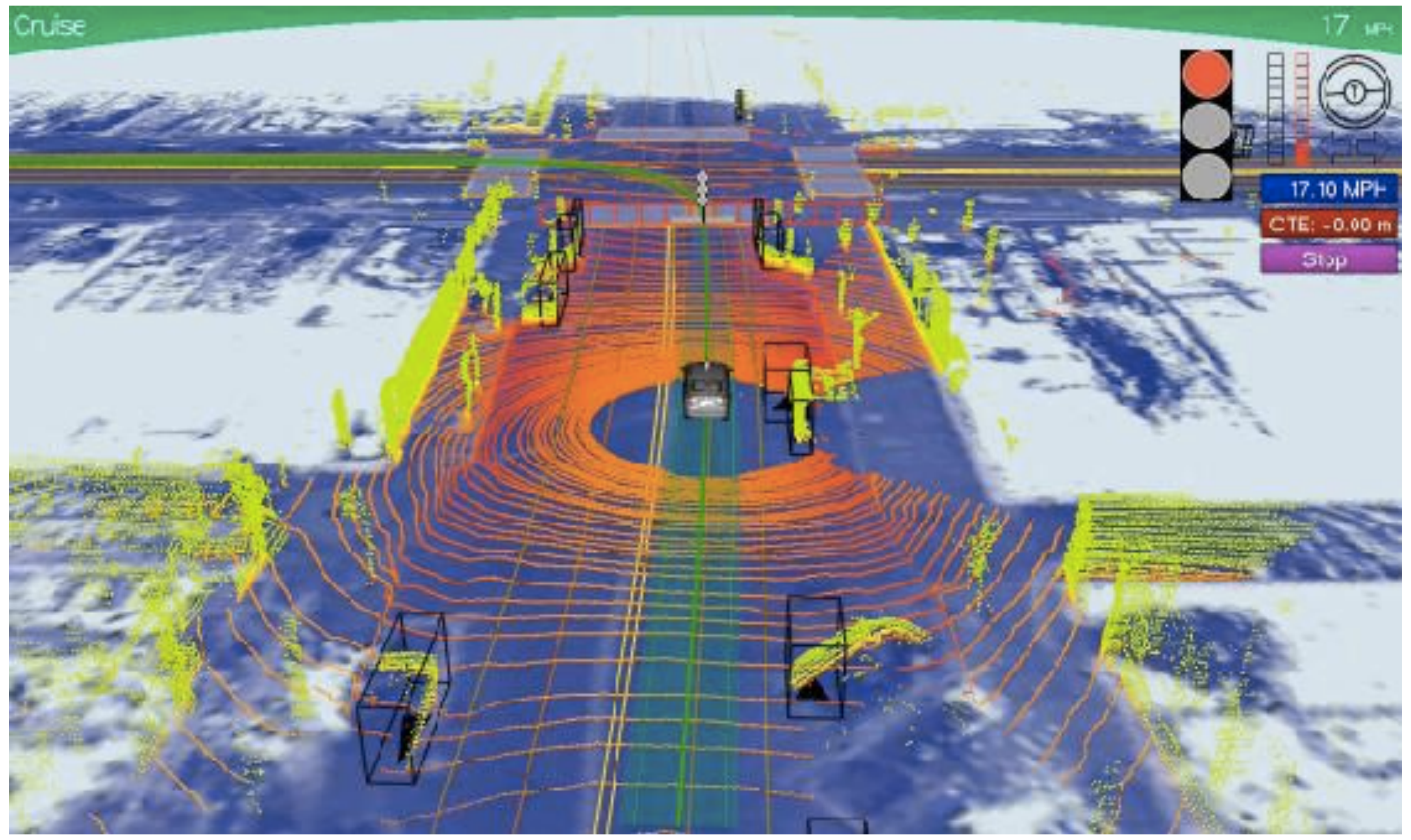
Sudeep Pillai

PhD Thesis Defense
Aug 29, 2017

MOTIVATION

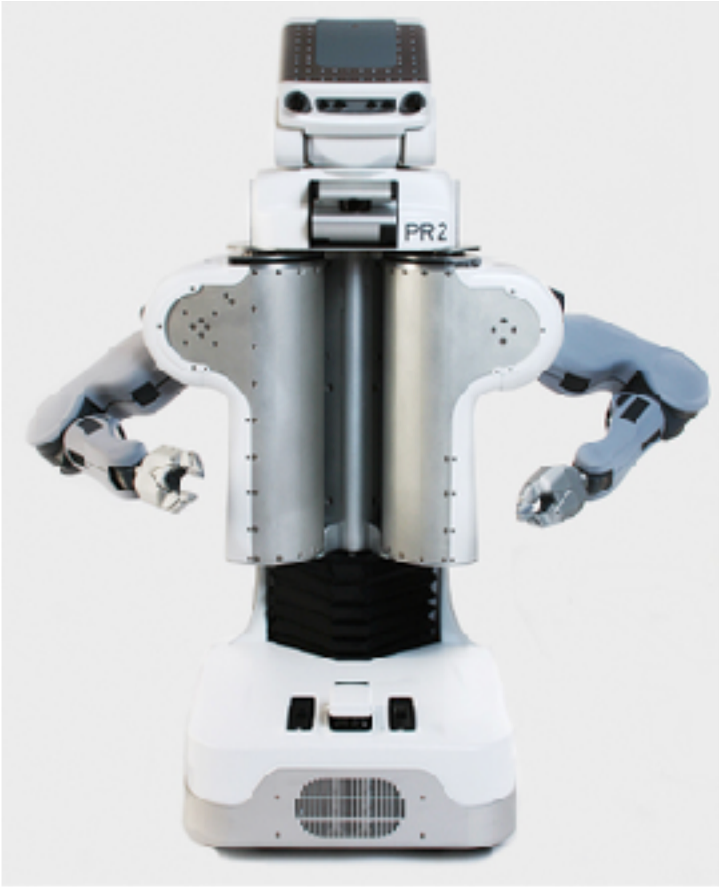
Mobile robots today are endowed with rich spatial models to effectively understand and navigate in the world

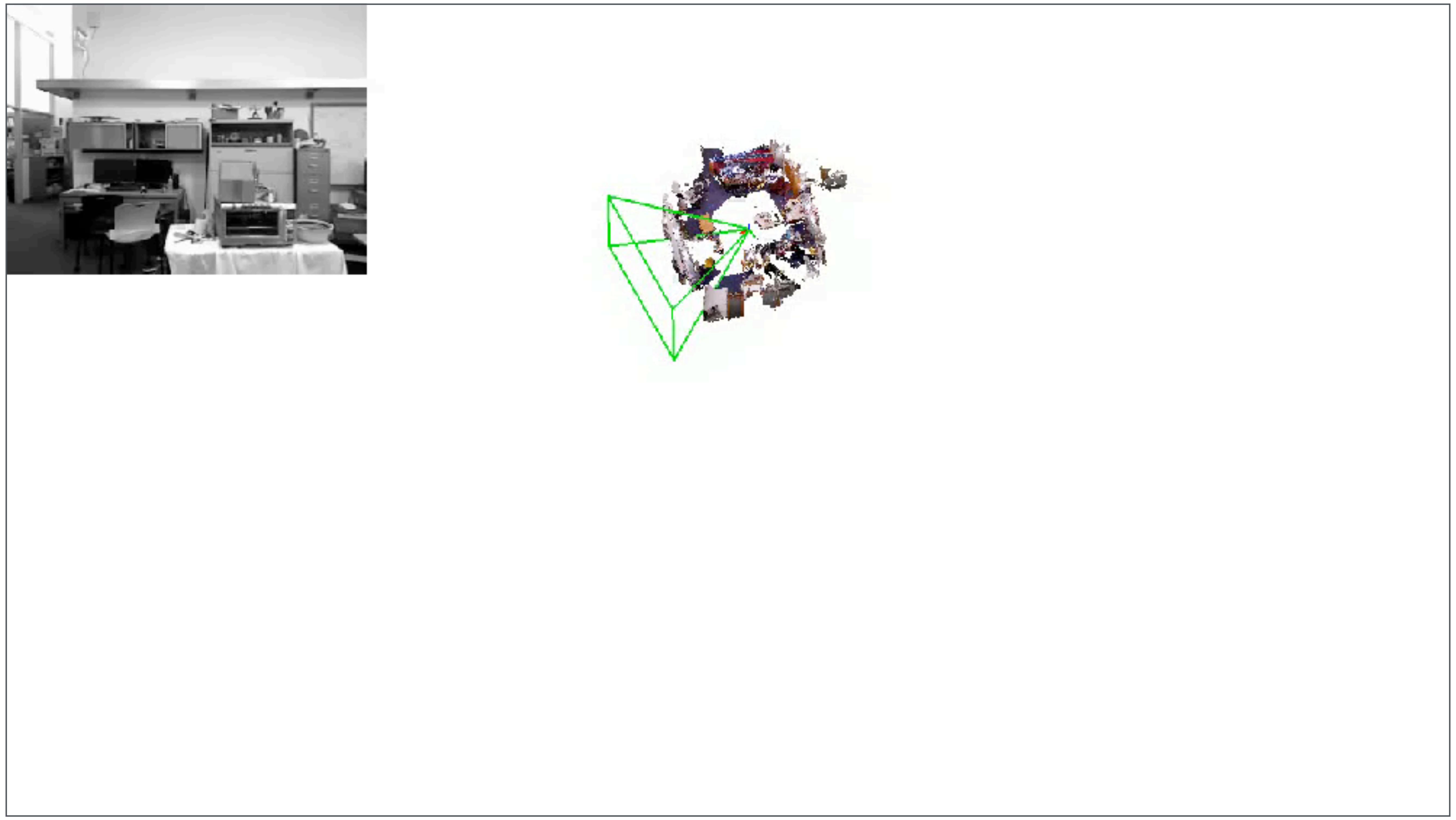
GEOMETRIC SCENE UNDERSTANDING
FOR NAVIGATION



MOTIVATION

SPATIALLY-COGNIZANT ROBOTS with SLAM



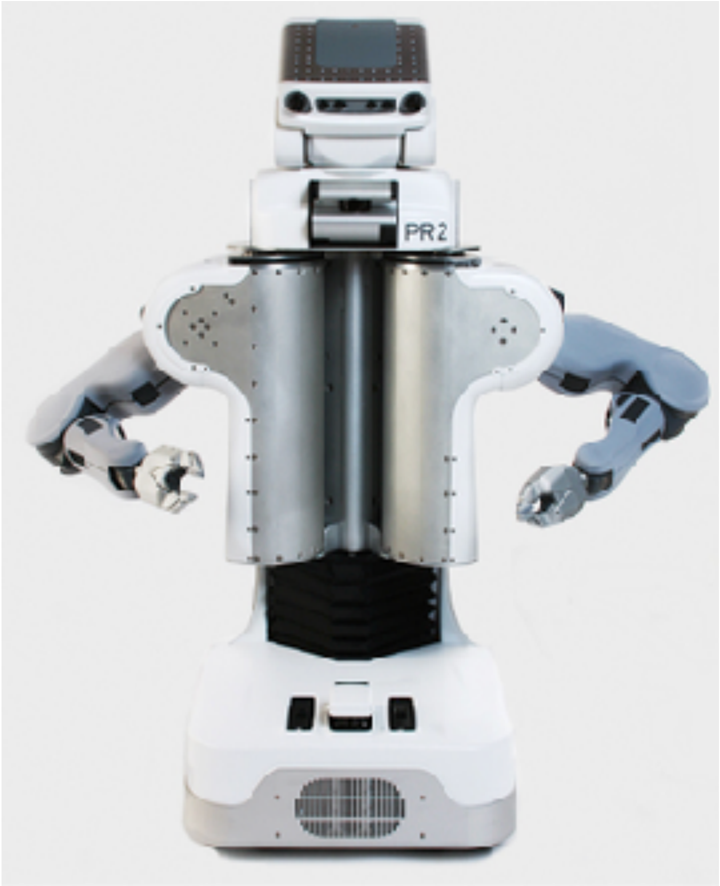


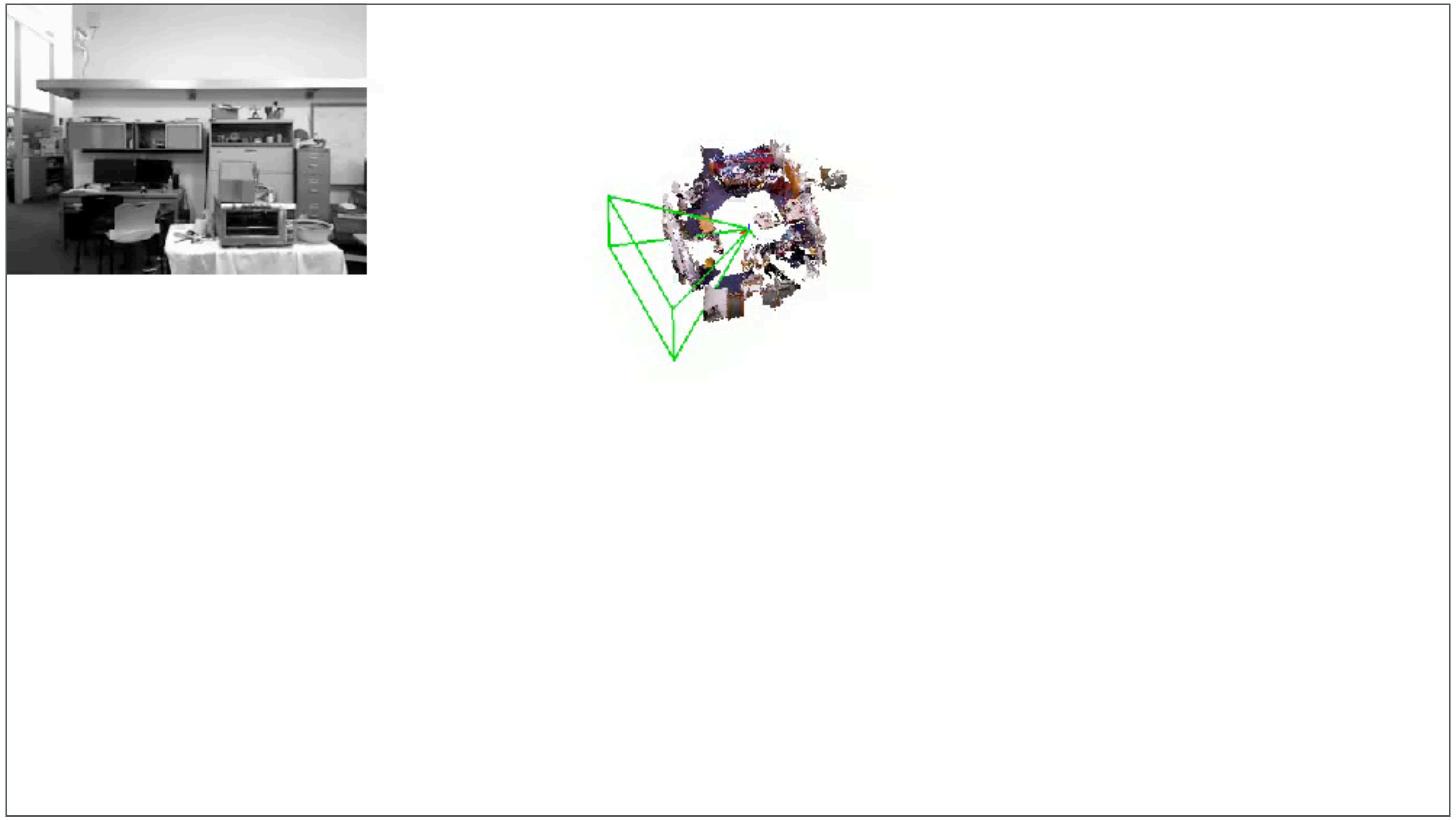
Temporally Scalable Visual SLAM using a Reduced Pose Graph

[Johannsson et. al 2013]

MOTIVATION

SPATIALLY-COGNIZANT ROBOTS with SLAM





Temporally Scalable Visual SLAM using a Reduced Pose Graph

[Johannsson et. al 2013]

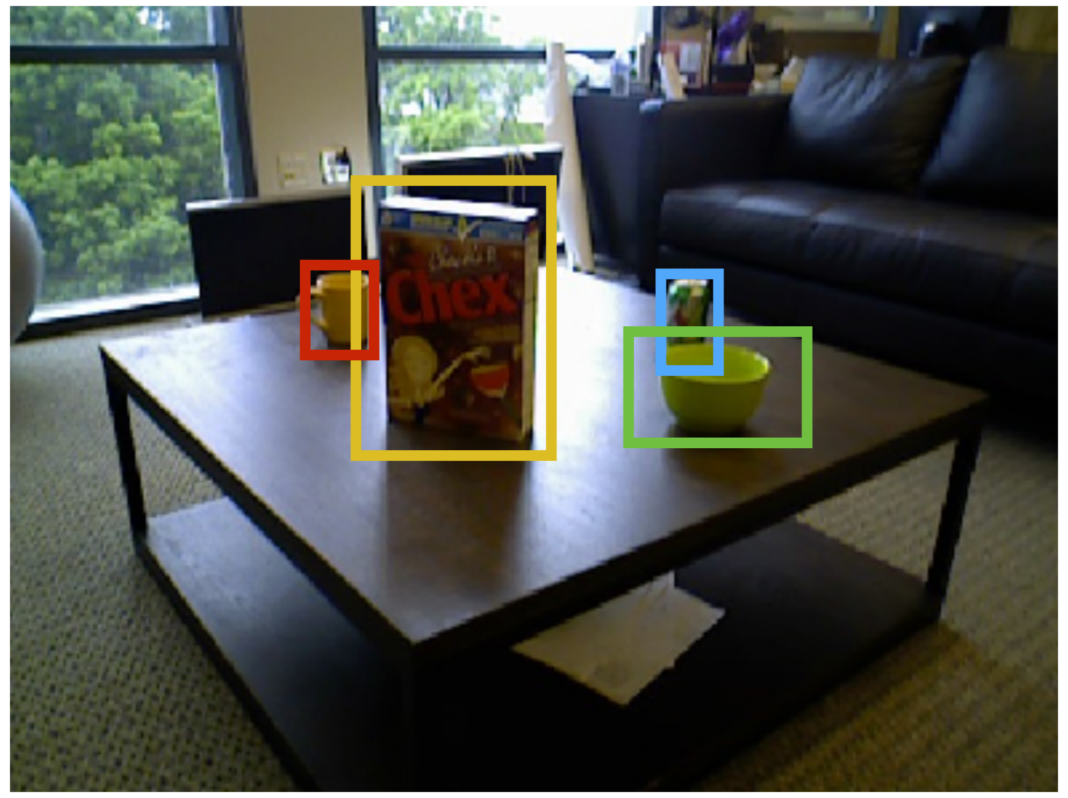
SLAM AS A SUPERVISORY SIGNAL

Mobile robots need to be endowed with SLAM-aware perceptual models for navigation and scene understanding, effectively using SLAM as a supervisory signal

SLAM AS A SUPERVISORY SIGNAL

Mobile robots need to be endowed with SLAM-aware perceptual models for navigation and scene understanding, effectively using SLAM as a supervisory signal

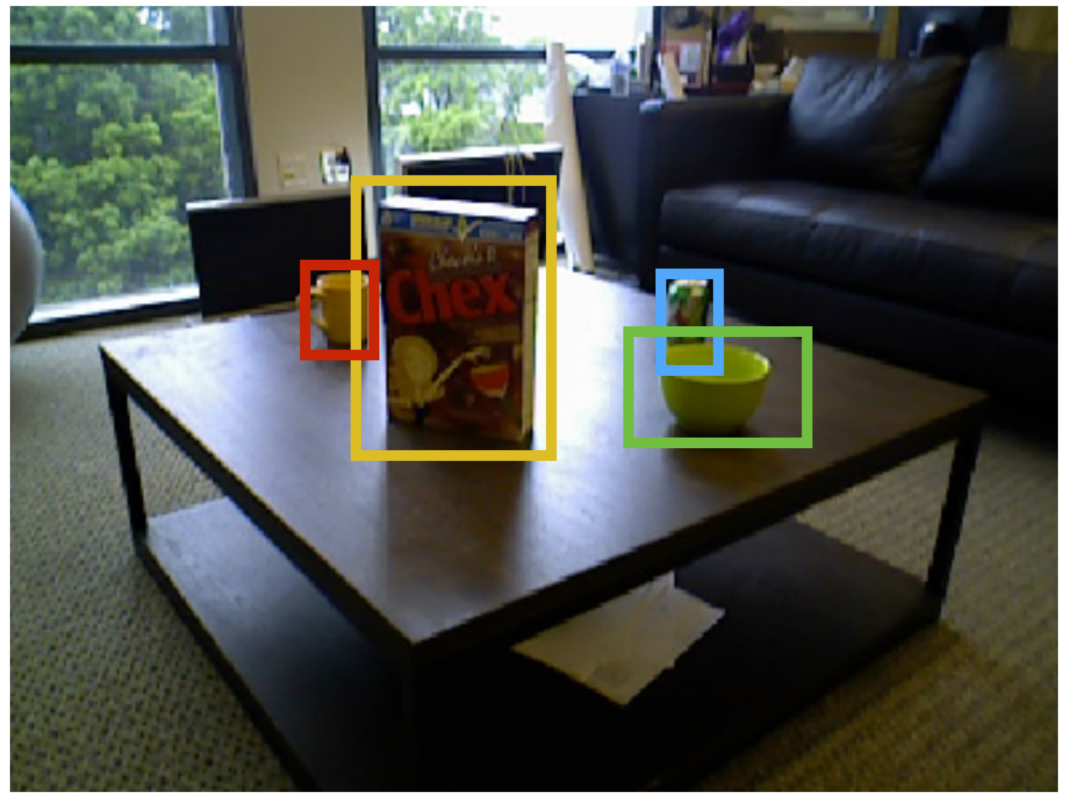
OBJECT RECOGNITION



SLAM AS A SUPERVISORY SIGNAL

Mobile robots need to be endowed with SLAM-aware perceptual models for navigation and scene understanding, effectively using
SLAM as a supervisory signal

OBJECT RECOGNITION



LEARNING VIA SELF-SUPERVISION

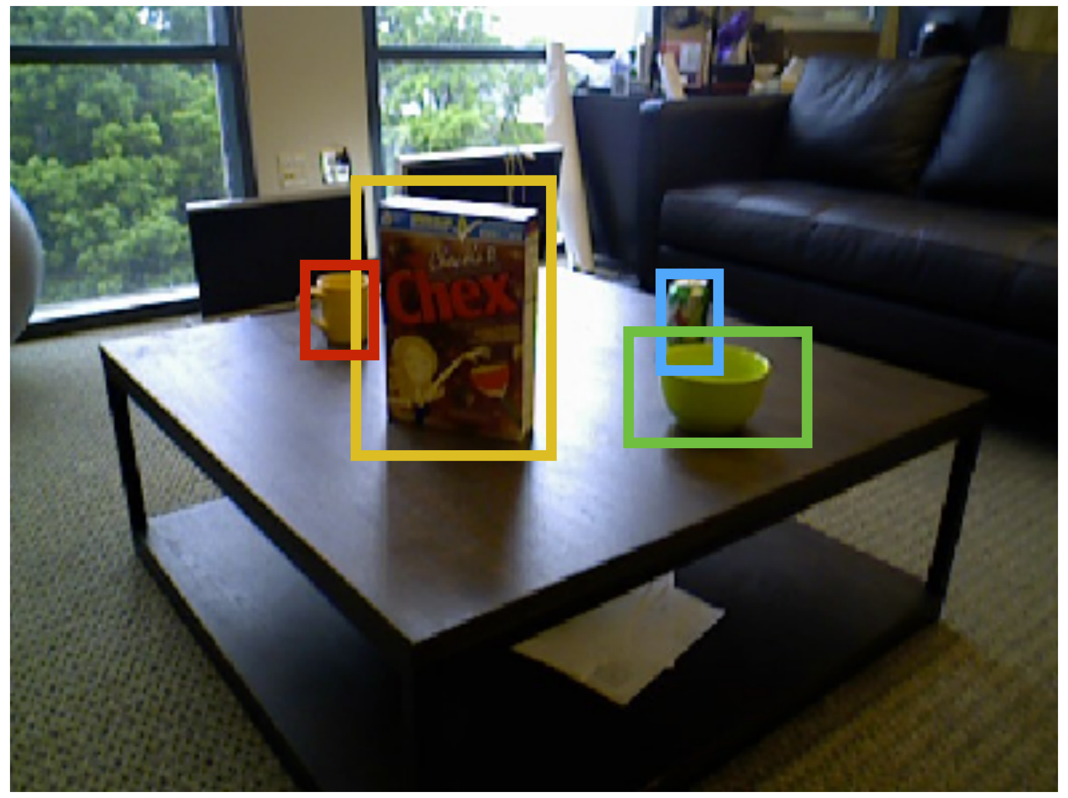
MIT DGC Vehicle (2007)

Uber ATG Vehicle (2017)

SLAM AS A SUPERVISORY SIGNAL

Mobile robots need to be endowed with SLAM-aware perceptual models for navigation and scene understanding, effectively using
SLAM as a supervisory signal

OBJECT RECOGNITION



LEARNING VIA SELF-SUPERVISION

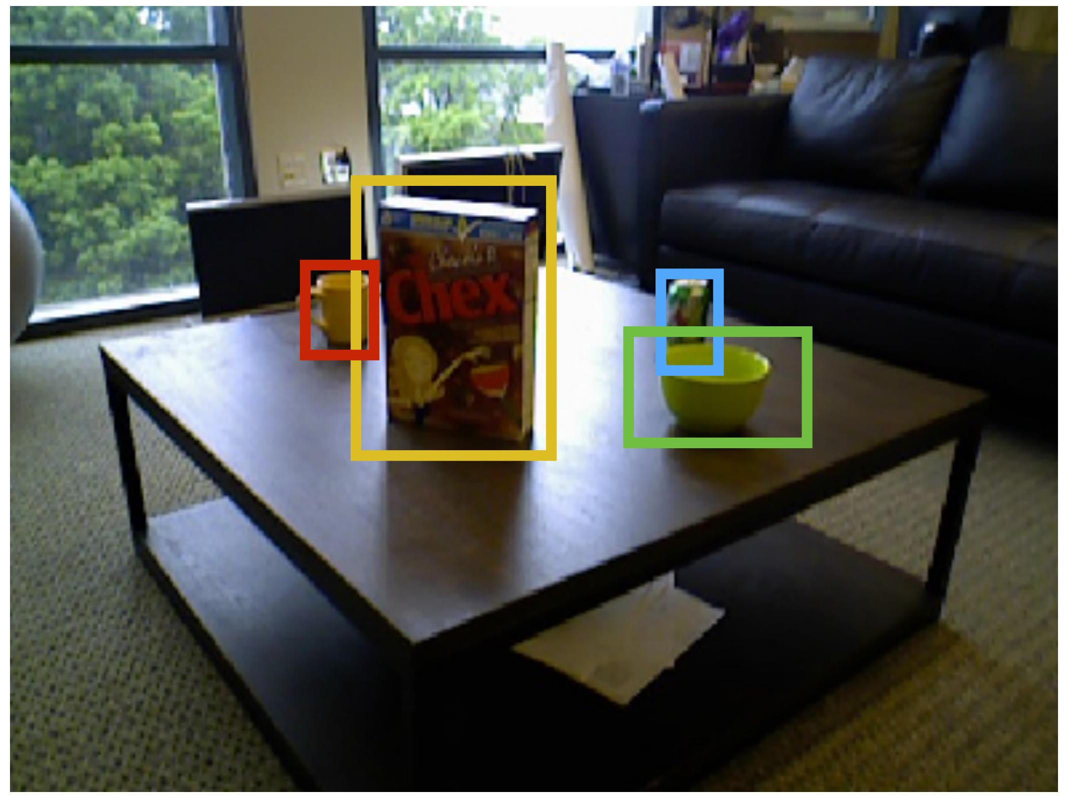
MIT DGC Vehicle (2007)

Uber ATG Vehicle (2017)

SLAM AS A SUPERVISORY SIGNAL

Mobile robots need to be endowed with SLAM-aware perceptual models for navigation and scene understanding, effectively using
SLAM as a supervisory signal

OBJECT RECOGNITION

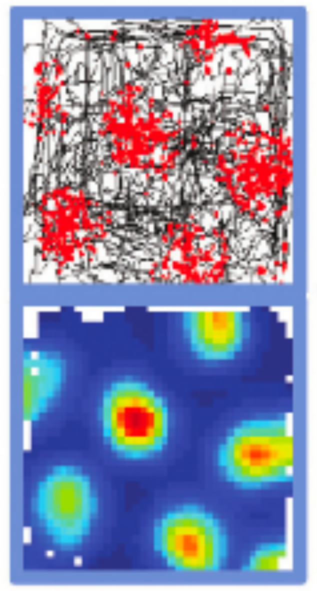


LEARNING VIA SELF-SUPERVISION

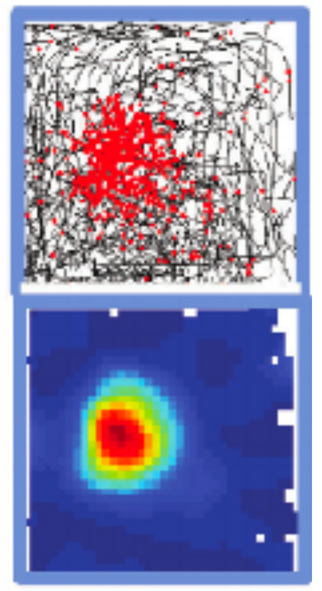
MIT DGC Vehicle (2007)

Uber ATG Vehicle (2017)

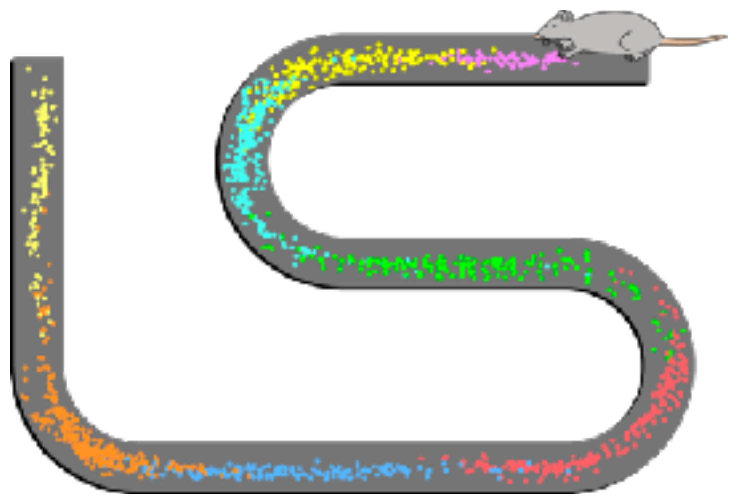
LEARNING TO LOCALIZE



Grid cells



Place-cells

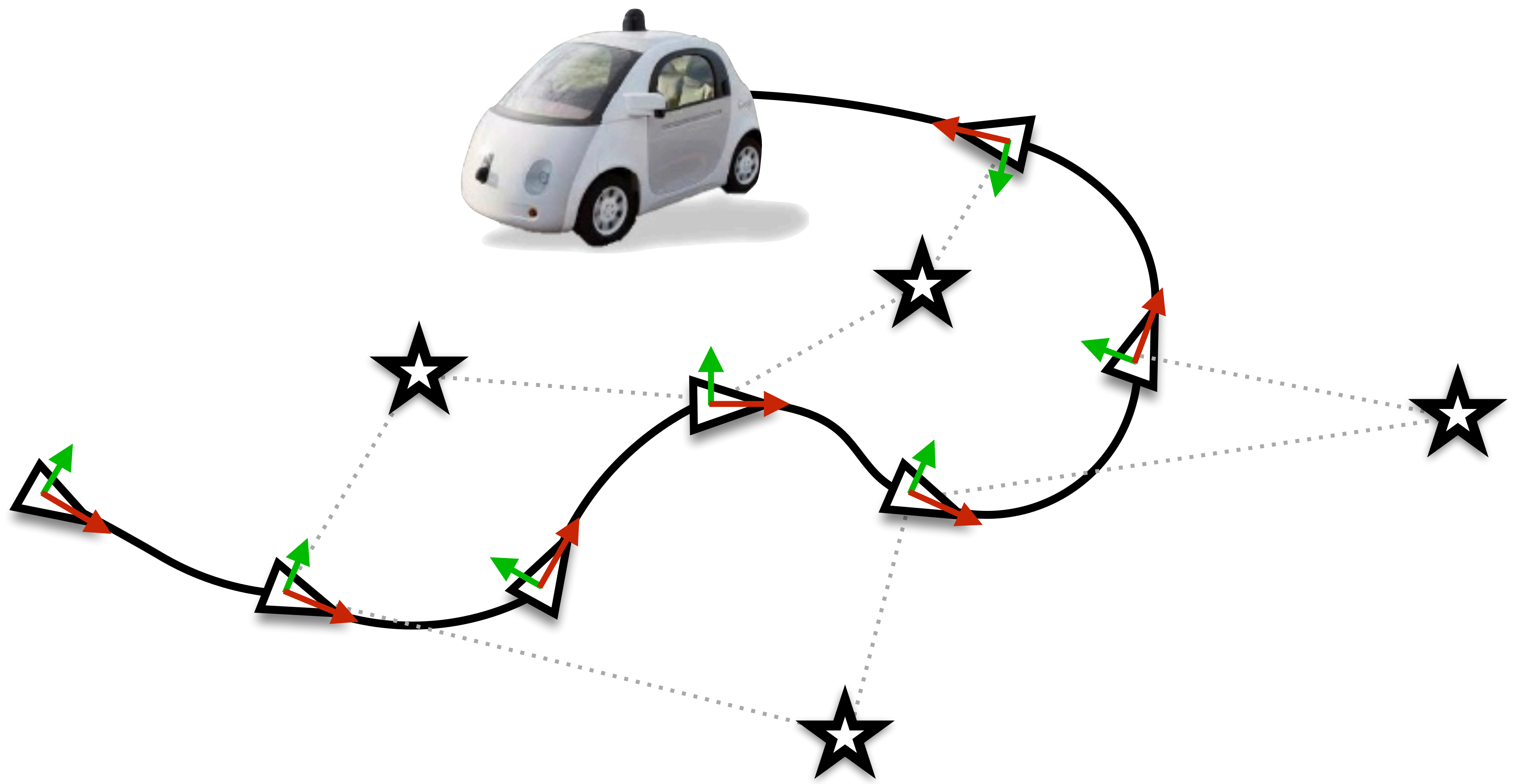


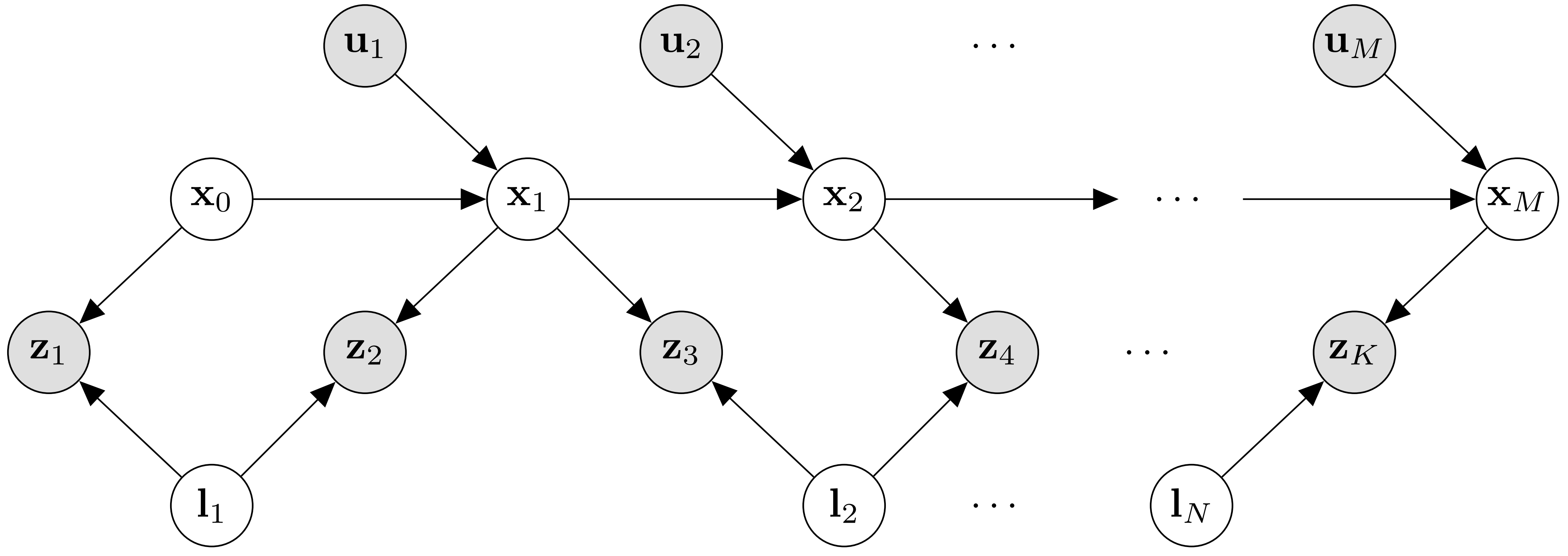
Place-cells

2014 Nobel Prize in Physiology or Medicine
Spatial Cells in the Hippocampal Formation
John O'Keefe, May-Britt Moser, Edvard I. Moser

SLAM

- Simultaneous Localization and Mapping
  - Joint probability distribution
  - Factored and represented as a DGM





SLAM as a Bayes Net

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{U}, \mathbf{Z}) \propto p(\mathbf{x}_0) \prod_{i=1}^{M} p(\mathbf{x}_i \mid \mathbf{x}_{i-1}, \mathbf{u}_i) \prod_{k=1}^{K} p(\mathbf{z}_k \mid \mathbf{x}_{ik}, \mathbf{l}_{jk})$$

$$\propto \underbrace{\prod_{i=1}^{M} \exp\left(-\frac{1}{2} \|f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i\|^2_{\Sigma_u}\right)}_{\text{Influence of odometry measurements}} \underbrace{\prod_{k=1}^{K} \exp\left(-\frac{1}{2} \|h_k(\mathbf{x}_{ik}, \mathbf{l}_{jk}) - \mathbf{z}_k\|^2_{\Sigma_k}\right)}_{\text{Influence of landmark measurements}}$$

Latent variables

$\mathbf{x}$ Robot state

$\mathbf{l}$ Landmarks

Measurements

$\mathbf{u}$ Odometry

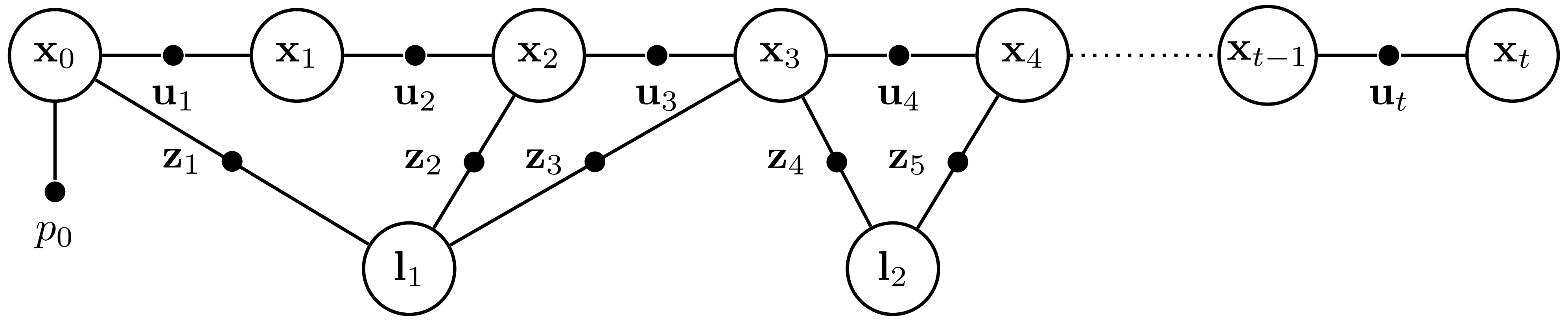
$\mathbf{z}$ Landmark sightings

Factored Joint probability distribution

FACTOR GRAPHS FOR SLAM

Visual-SLAM

Bundle Adjustment with odometry



$$
\begin{aligned}
\mathbf{X}^*, \mathbf{L}^* &= \underset{\mathbf{X},\mathbf{L}}{\arg\max}\, p(\mathbf{X}, \mathbf{L} \mid \mathbf{U}, \mathbf{Z}_{\mathbf{l}}) \\
&= \underset{\mathbf{X},\mathbf{L}}{\arg\min} \left\{ \underbrace{\sum_{i=1}^{M} \| f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i \|^2_{\sum_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{k=1}^{K} \| h_k(\mathbf{x}_{ik}, \mathbf{l}_{jk}) - \mathbf{z}_k \|^2_{\sum_k}}_{\text{Bundle Adjustment Problem}} \right\}
\end{aligned}
$$

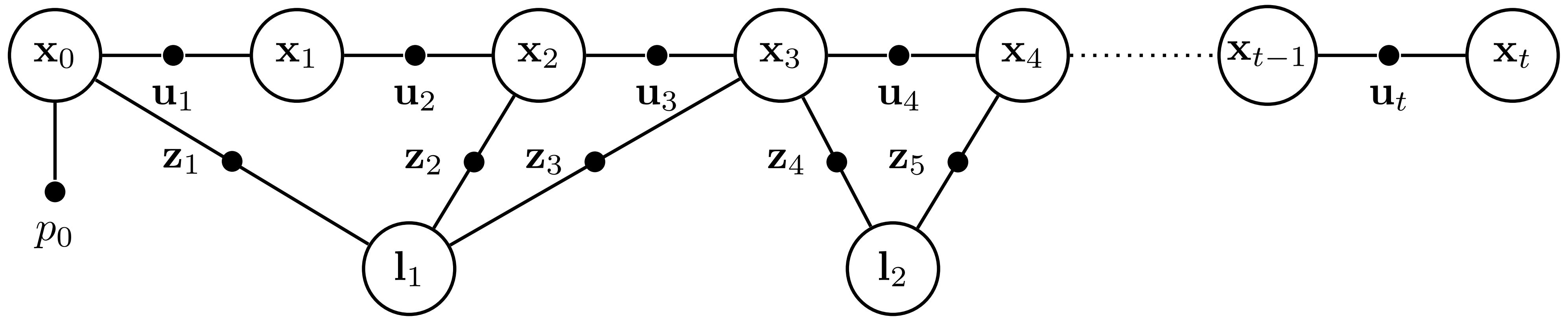
Latent variables $\mathbf{x}$ Robot state $\mathbf{l}$ Landmarks

Measurements $\mathbf{u}$ Odometry $\mathbf{z}$ Landmark sightings p Prior

FACTOR GRAPHS FOR SLAM

Visual-SLAM

Bundle Adjustment with odometry

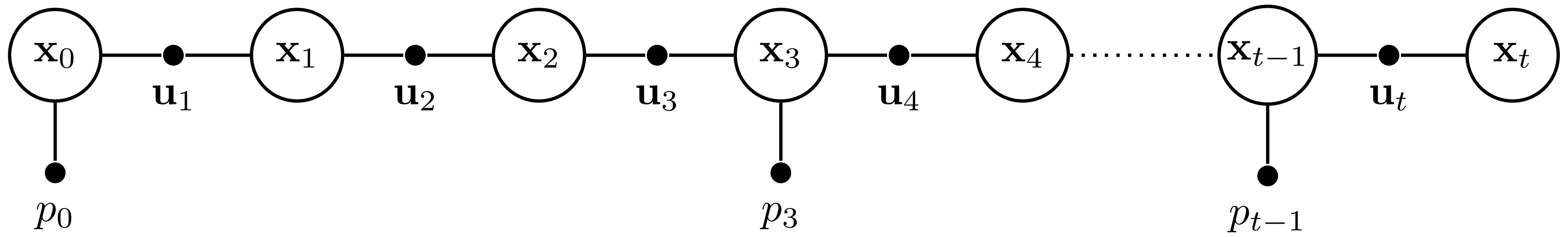


$$\mathbf{X}^*, \mathbf{L}^* = \arg\max_{\mathbf{X},\mathbf{L}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{U}, \mathbf{Z_l})$$

$$= \arg\min_{\mathbf{X},\mathbf{L}} \left\{ \underbrace{\sum_{i=1}^{M} \| f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i \|^2_{\Sigma_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{k=1}^{K} \| h_k(\mathbf{x}_{ik}, \mathbf{l}_{jk}) - \mathbf{z}_k \|^2_{\Sigma_k}}_{\text{Bundle Adjustment Problem}} \right\}$$

GPS-aided Localization

Fusing odometry with intermittent GPS updates



$$\mathbf{X}^* = \arg\max_{\mathbf{X}} p(\mathbf{X} \mid \mathbf{U}, \mathbf{Z_g})$$

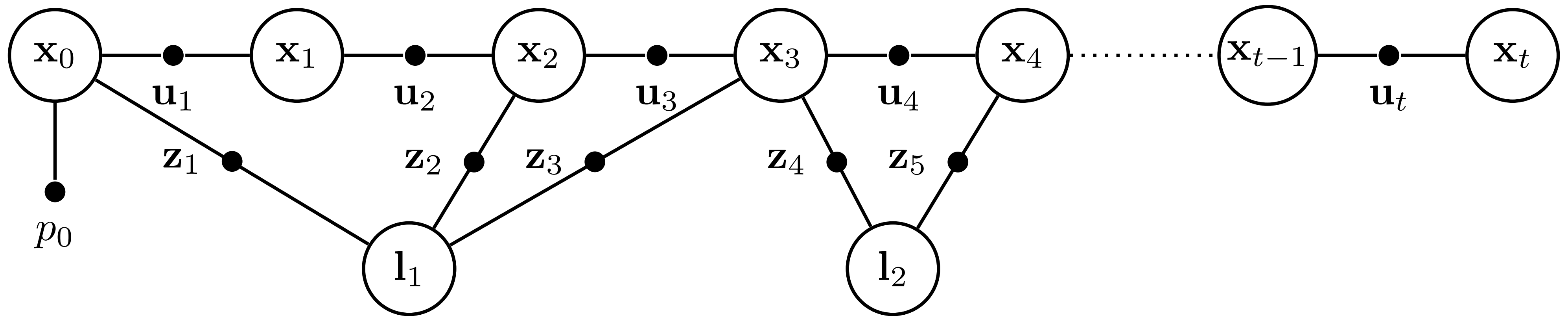
$$= \arg\min_{\mathbf{X}} \left\{ \underbrace{\sum_{i=1}^{M} \| f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i \|^2_{\Sigma_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{j=1}^{G} \| h_g(\mathbf{x}_j) - \mathbf{z}_j \|^2_{\Sigma_g}}_{\text{GPS Measurement Priors}} \right\}$$

Latent variables $\mathbf{x}$ Robot state $\mathbf{l}$ Landmarks

Measurements $\mathbf{u}$ Odometry $\mathbf{z}$ Landmark sightings p Prior

FACTOR GRAPHS FOR SLAM

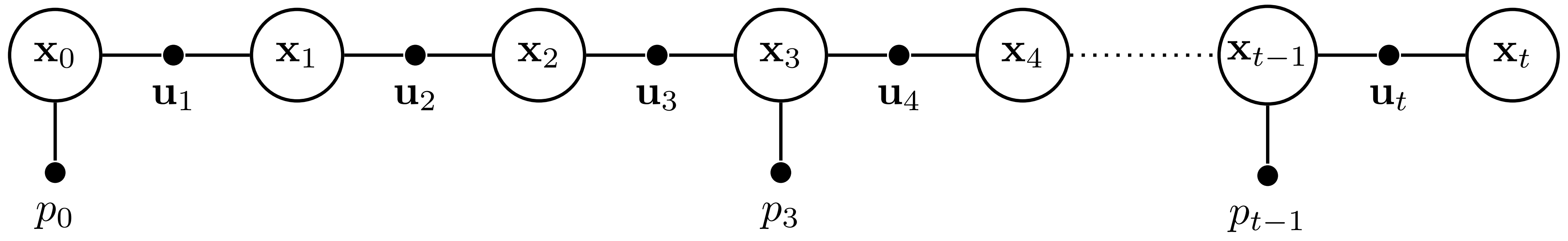
Visual-SLAM
Bundle Adjustment with odometry



$$\mathbf{X}^*, \mathbf{L}^* = \arg\max_{\mathbf{X},\mathbf{L}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{U}, \mathbf{Z_l})$$

$$= \arg\min_{\mathbf{X},\mathbf{L}} \left\{ \underbrace{\sum_{i=1}^{M} \|f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i\|^2_{\sum_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{k=1}^{K} \|h_k(\mathbf{x}_{ik}, \mathbf{l}_{jk}) - \mathbf{z}_k\|^2_{\sum_k}}_{\text{Bundle Adjustment Problem}} \right\}$$

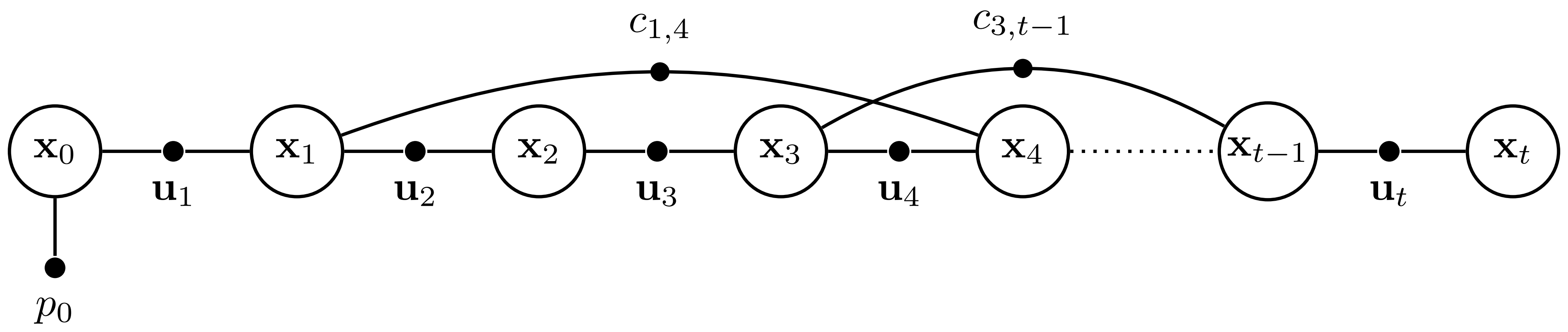
GPS-aided Localization
Fusing odometry with intermittent GPS updates



$$\mathbf{X}^* = \arg\max_{\mathbf{X}} p(\mathbf{X} \mid \mathbf{U}, \mathbf{Z_g})$$

$$= \arg\min_{\mathbf{X}} \left\{ \underbrace{\sum_{i=1}^{M} \|f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i\|^2_{\sum_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{j=1}^{G} \|h_g(\mathbf{x}_j) - \mathbf{z}_j\|^2_{\sum_g}}_{\text{GPS Measurement Priors}} \right\}$$

Pose-Graph SLAM
Fusing odometry with loop-closure constraints



$$\mathbf{X}^* = \arg\max_{\mathbf{X}} p(\mathbf{X} \mid \mathbf{U}, \mathbf{Z_c})$$

$$= \arg\min_{\mathbf{X}} \left\{ \underbrace{\sum_{i=1}^{M} \|f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i\|^2_{\sum_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{(j,k)\in\mathcal{C}} \|h_c(\mathbf{x}_j, \mathbf{x}_k) - \mathbf{z}_{jk}\|^2_{\sum_c}}_{\text{Loop-Closure Constraint Factors}} \right\}$$

Latent variables $\mathbf{x}$ Robot state $\mathbf{l}$ Landmarks

Measurements $\mathbf{u}$ Odometry $\mathbf{z}$ Landmark sightings p Prior

SLAM AS A SUPERVISORY SIGNAL

Monocular SLAM-Supported Object Recognition

Self-Supervised Visual Ego-motion Learning

Self-Supervised Visual Place Recognition Learning

SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

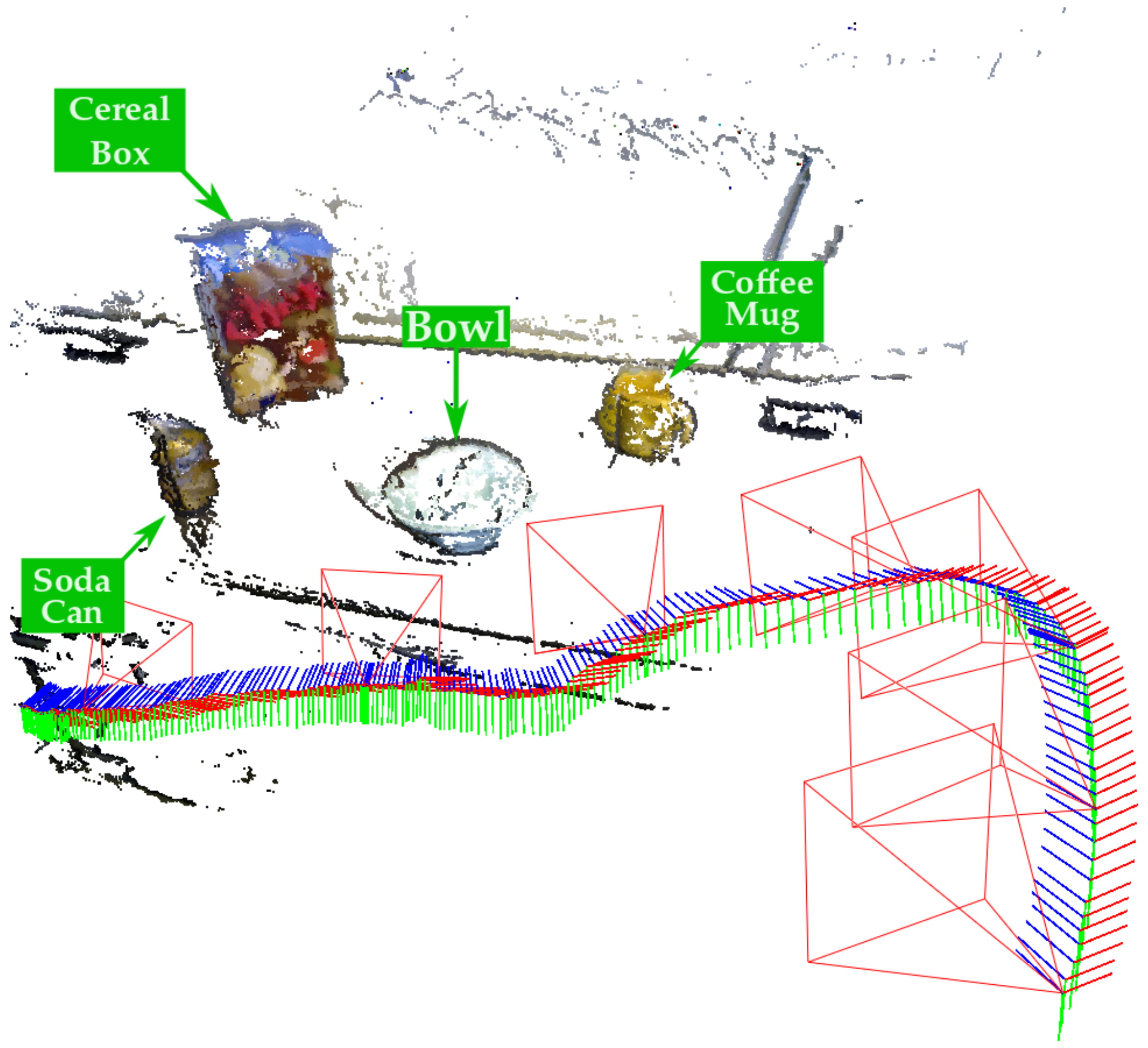
Correspondence Engine
(Geometric data association)

Self-Supervision
(SLAM-aided supervision)

Knowledge Transfer
(Bootstrapping)

SLAM AS A SUPERVISORY SIGNAL

Monocular SLAM-Supported Object Recognition



Self-Supervised Visual Ego-motion Learning

Self-Supervised Visual Place Recognition Learning

SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

Correspondence Engine
(Geometric data association)

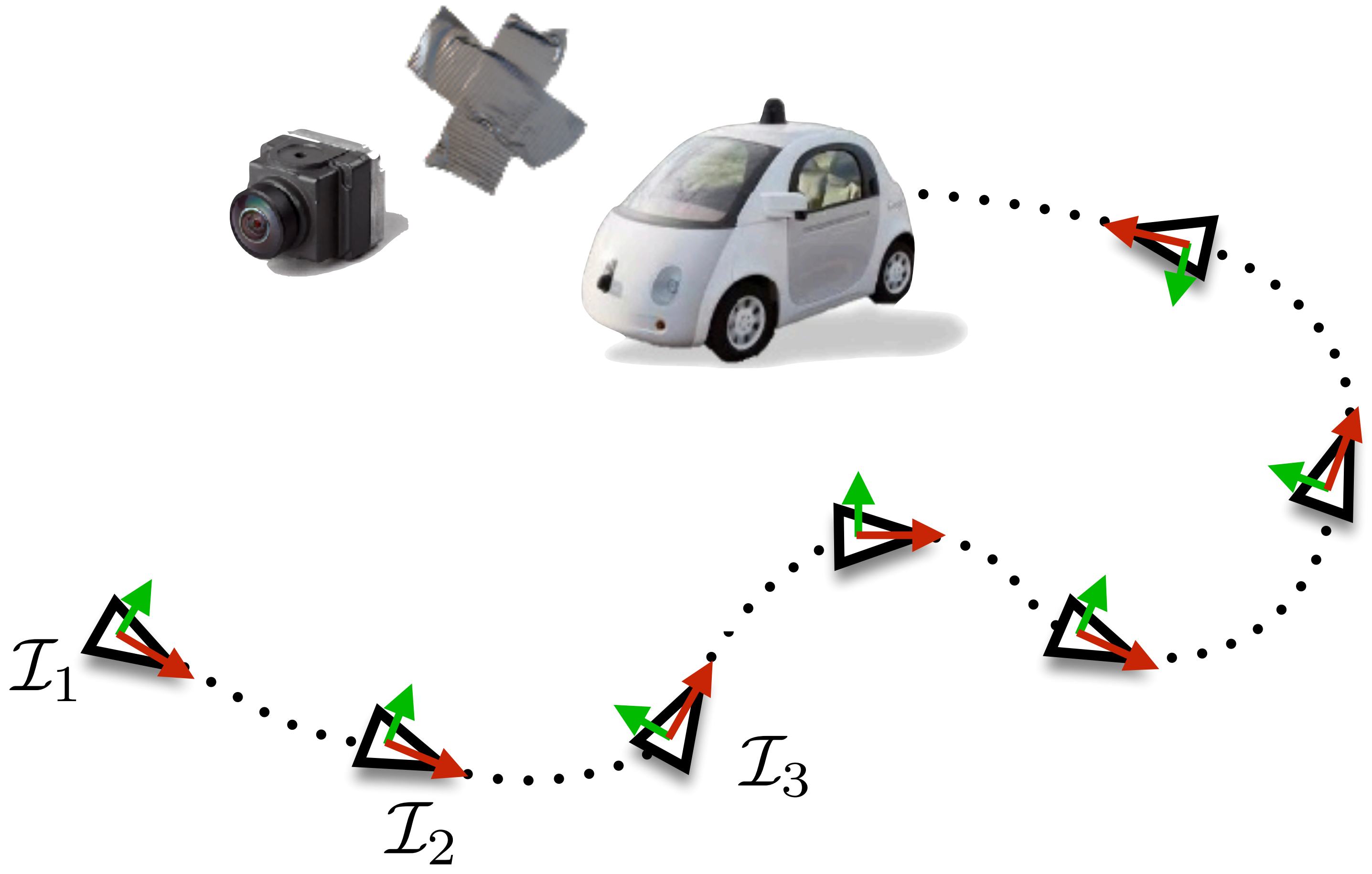
Self-Supervision
(SLAM-aided supervision)

Knowledge Transfer
(Bootstrapping)

SLAM AS A SUPERVISORY SIGNAL

Monocular SLAM-Supported Object Recognition

Self-Supervised Visual Ego-motion Learning



Self-Supervised Visual Place Recognition Learning

SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

Correspondence Engine
(Geometric data association)

Self-Supervision
(SLAM-aided supervision)

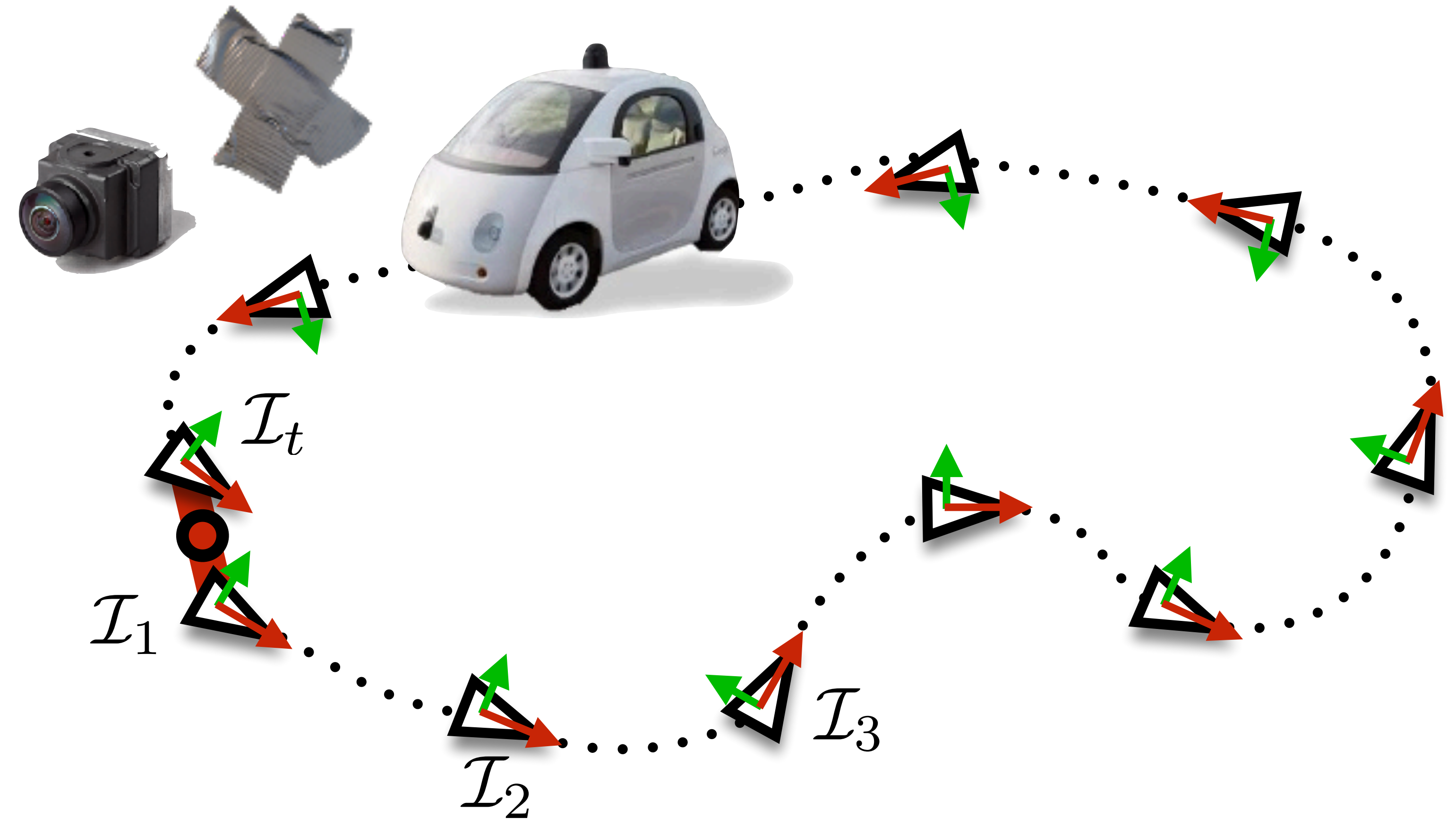
Knowledge Transfer
(Bootstrapping)

SLAM AS A SUPERVISORY SIGNAL

Monocular SLAM-Supported Object Recognition

Self-Supervised Visual Ego-motion Learning

Self-Supervised Visual Place Recognition Learning



SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS
with SLAM

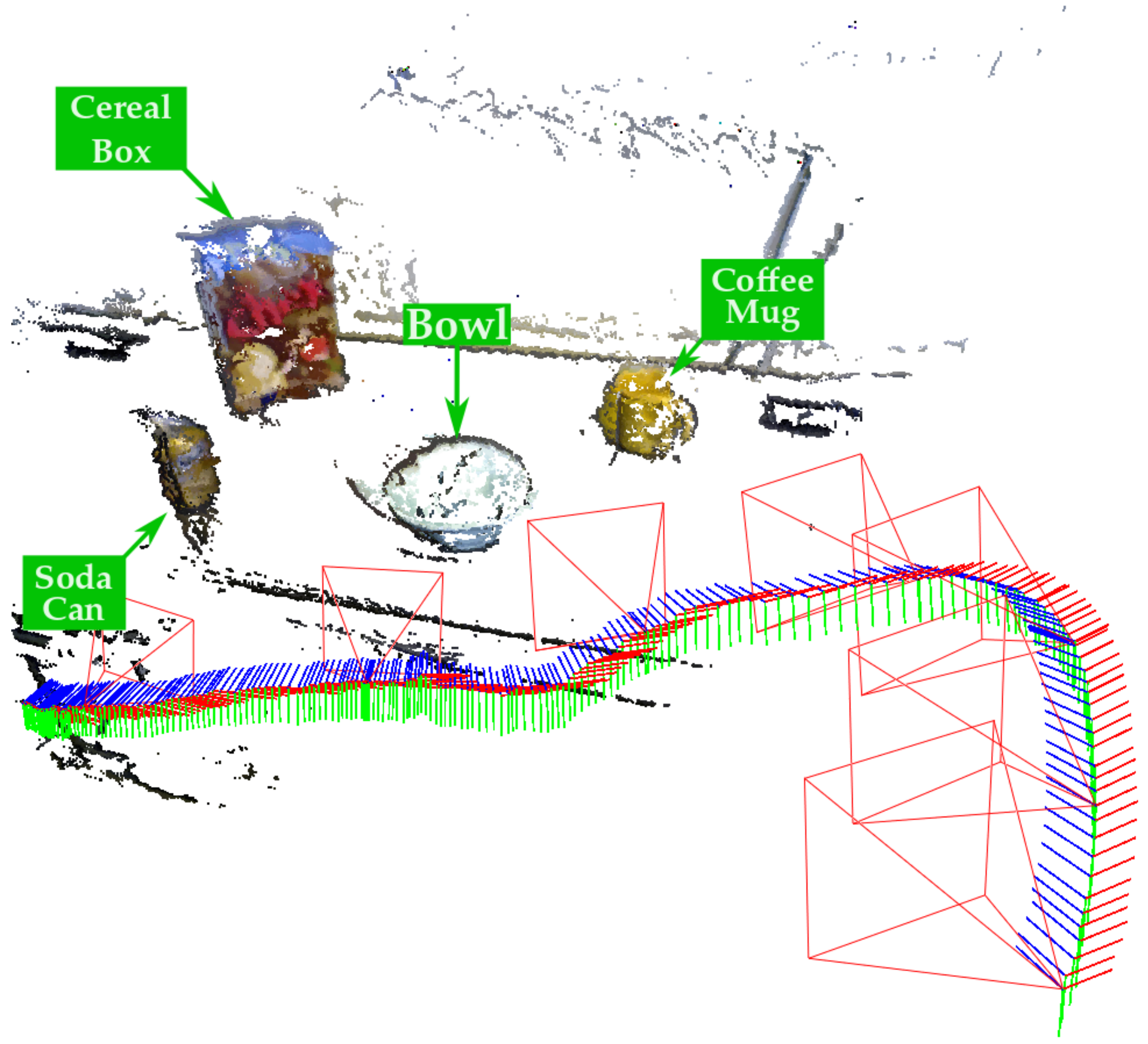
Correspondence Engine
(Geometric data association)

Self-Supervision
(SLAM-aided supervision)

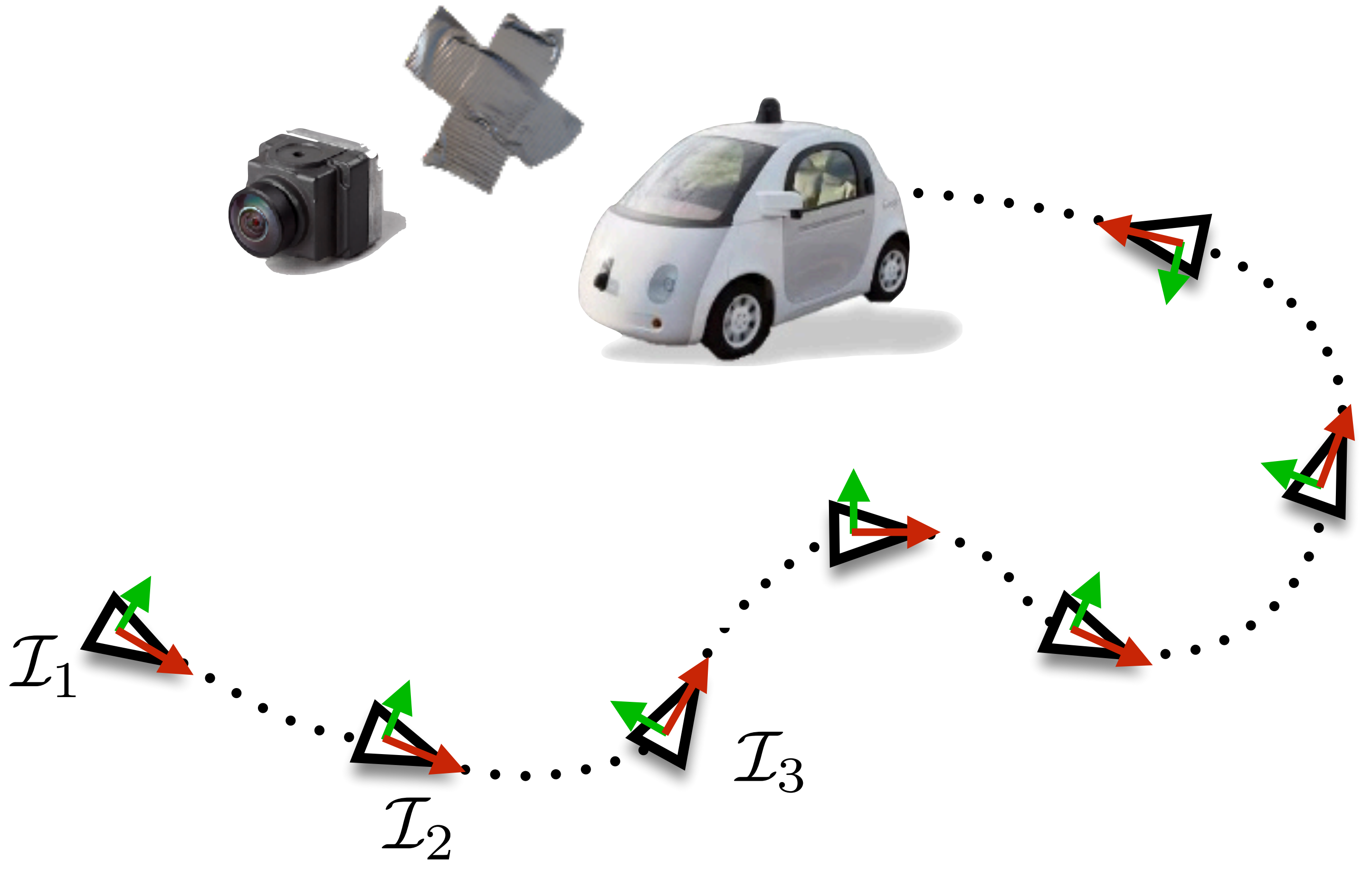
Knowledge Transfer
(Bootstrapping)

SLAM AS A SUPERVISORY SIGNAL

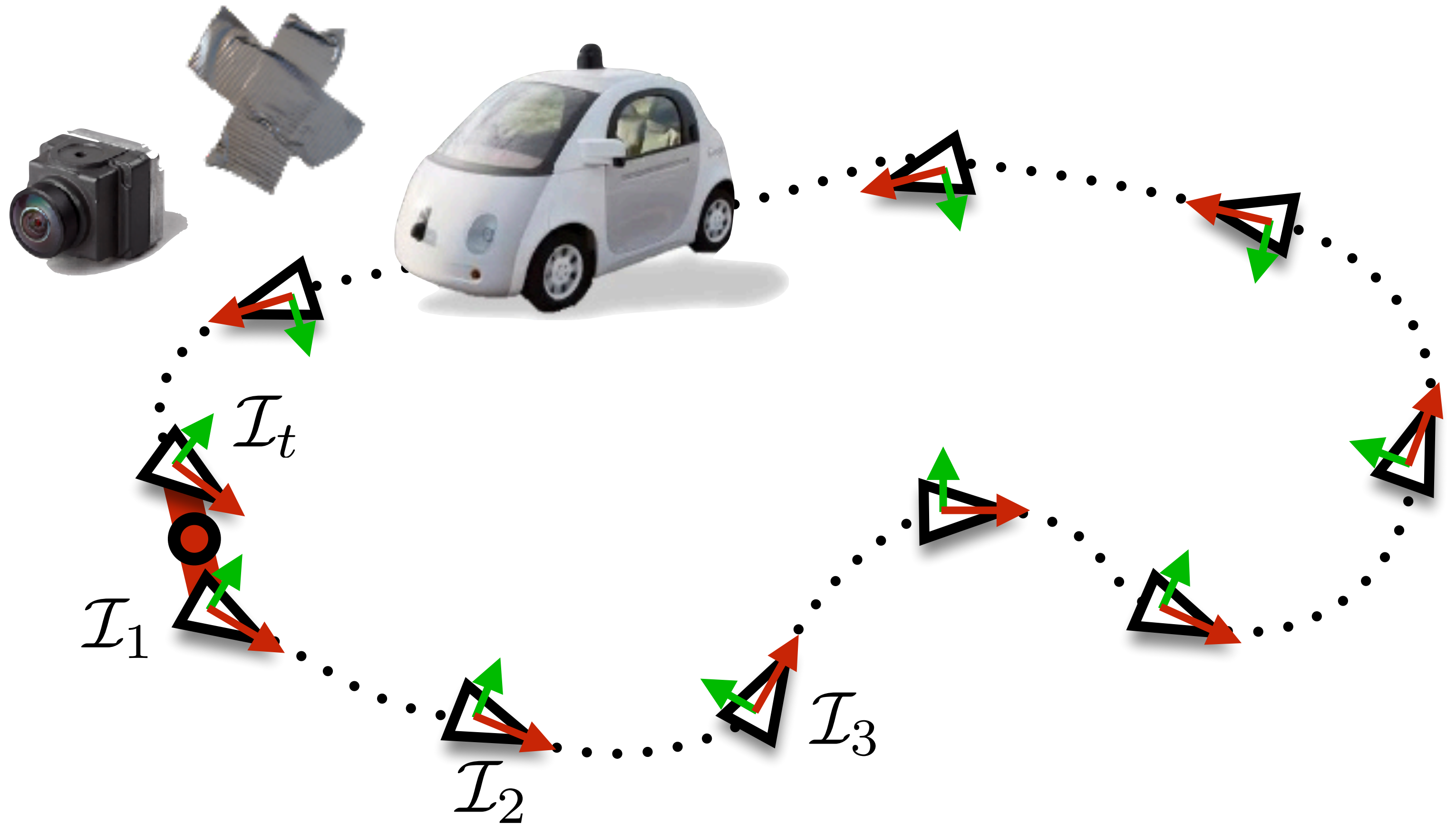
Monocular SLAM-Supported Object Recognition



Self-Supervised Visual Ego-motion Learning



Self-Supervised Visual Place Recognition Learning



SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

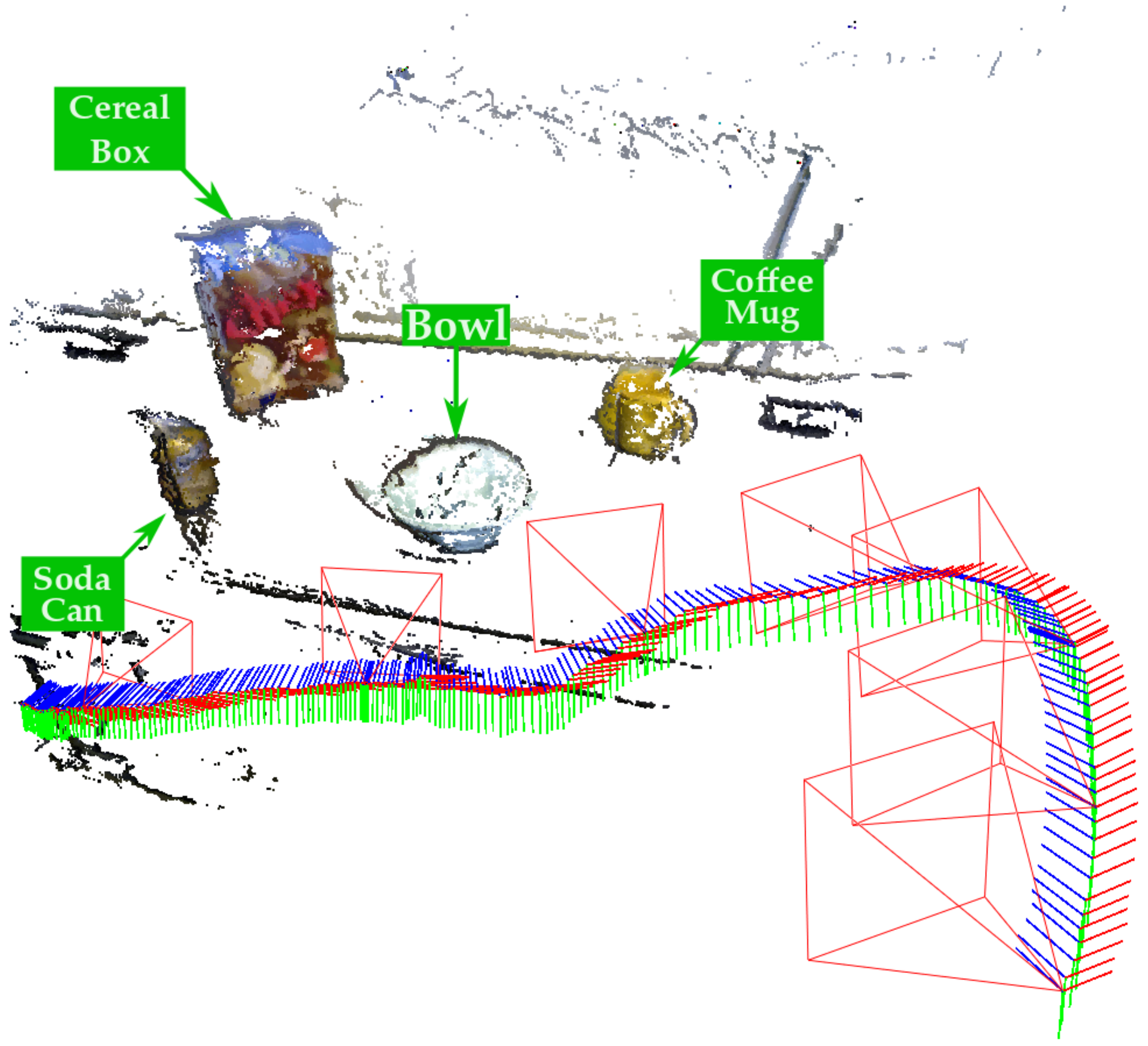
Correspondence Engine
(Geometric data association)

Self-Supervision
(SLAM-aided supervision)

Knowledge Transfer
(Bootstrapping)

SLAM AS A SUPERVISORY SIGNAL

Monocular SLAM-Supported Object Recognition



Self-Supervised Visual Ego-motion Learning

Self-Supervised Visual Place Recognition Learning

SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

Correspondence Engine
(Geometric data association)

Self-Supervision
(SLAM-aided supervision)

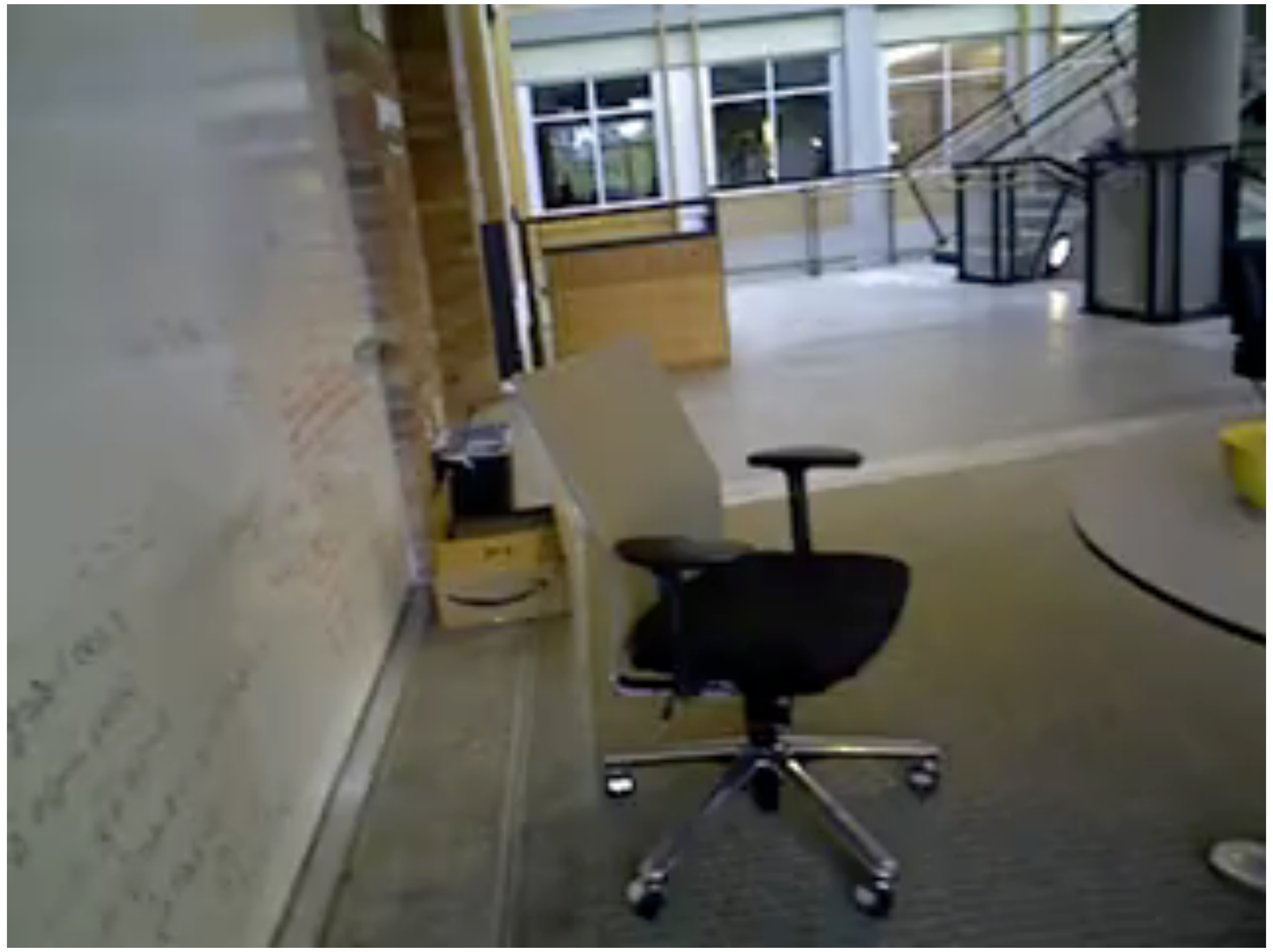
Knowledge Transfer
(Bootstrapping)

OBJECT RECOGNITION IN ROBOTS

Robots equipped with a **single RGB camera** need to **continuously recognize** and **localize all potential objects** in its immediate environment

Single RGB Camera
Versatile

Multi-view Object Detection
Camera & Object localization by leveraging SLAM



Input RGB Video

Robust
Avoid spurious detection/mis-classification

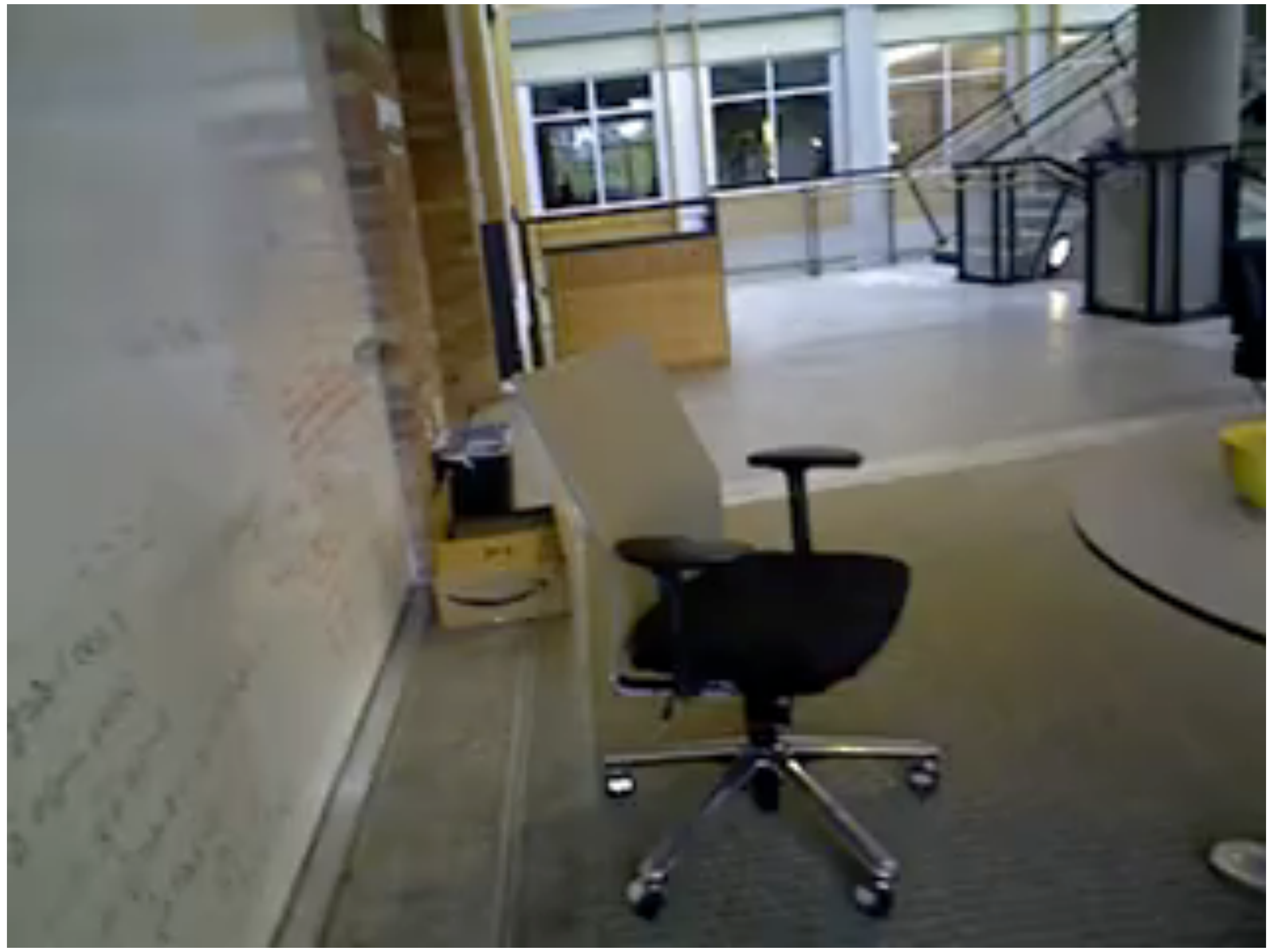
Real-time
Scalable recognition

OBJECT RECOGNITION IN ROBOTS

Robots equipped with a **single RGB camera** need to **continuously recognize** and **localize all potential objects** in its immediate environment

Single RGB Camera
Versatile

Multi-view Object Detection
Camera & Object localization by leveraging SLAM



Input RGB Video

Robust
Avoid spurious detection/mis-classification

Real-time
Scalable recognition

SEMANTIC AND GEOMETRIC SCENE UNDERSTANDING LANDSCAPE

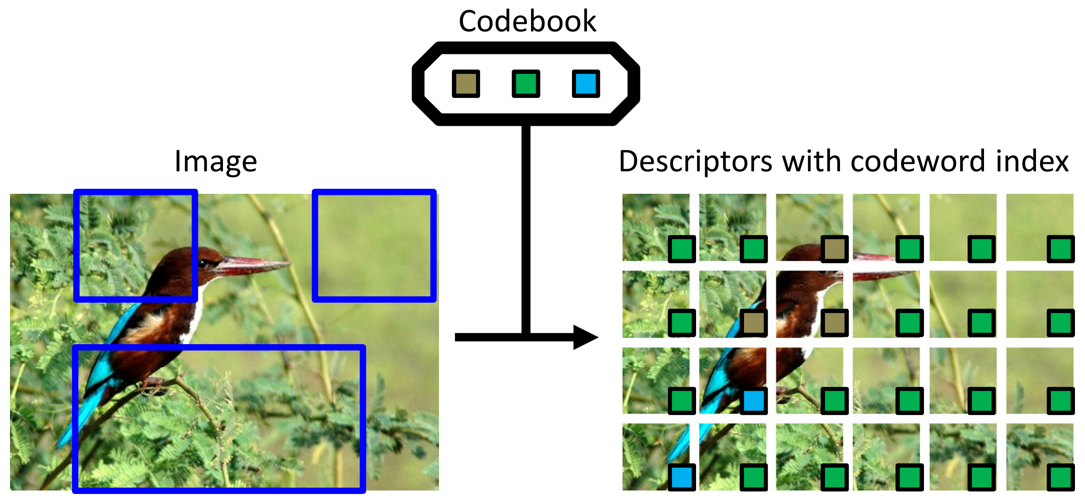
SEMANTIC AND GEOMETRIC SCENE UNDERSTANDING LANDSCAPE

- Shift in Visual-SLAM and Object Detection capabilities

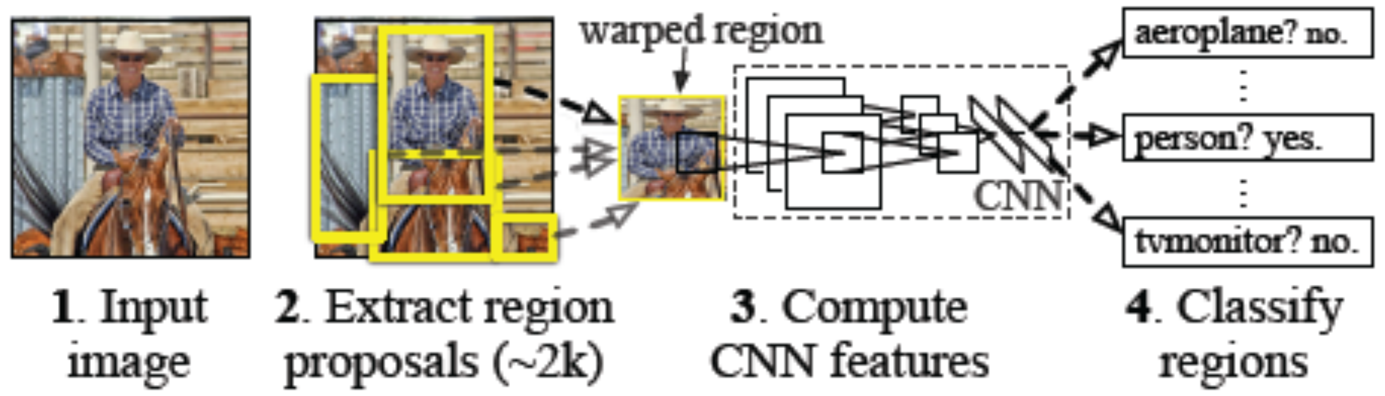
SEMANTIC AND GEOMETRIC SCENE UNDERSTANDING LANDSCAPE

- Shift in Visual-SLAM and Object Detection capabilities
  - **Richer Semantics**: Object Proposals, R-CNN, RoI Pooling/Align

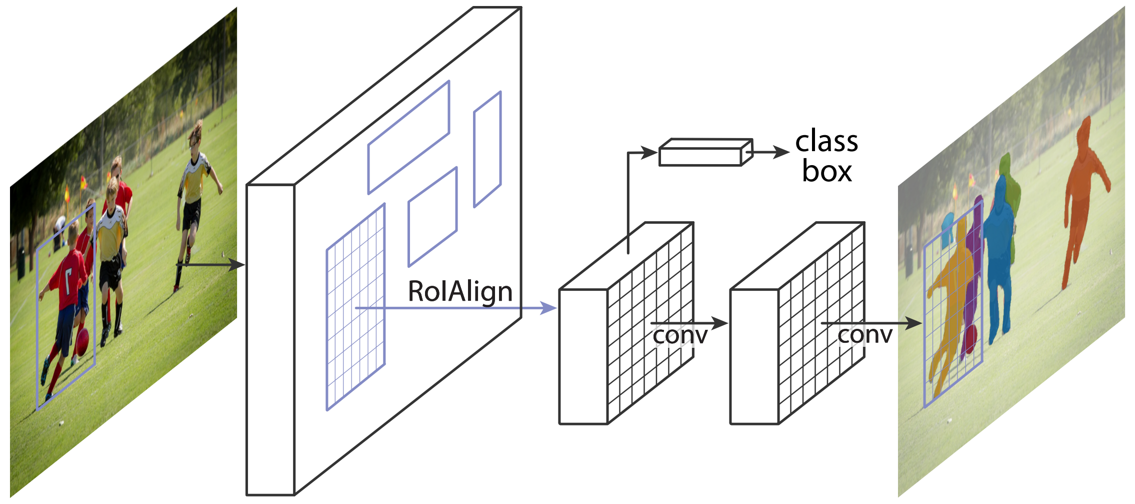
RICHER SEMANTICS



FLAIR (van de Sande 2014)



R-CNN, Fast(er) R-CNN (Girshick et al. 2014-5)

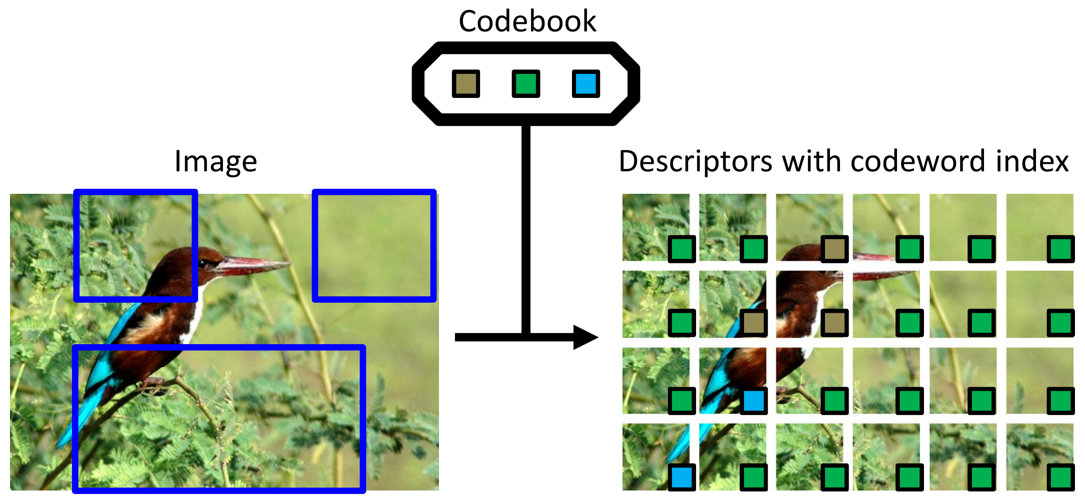


Mask-RCNN (He et al 2017)

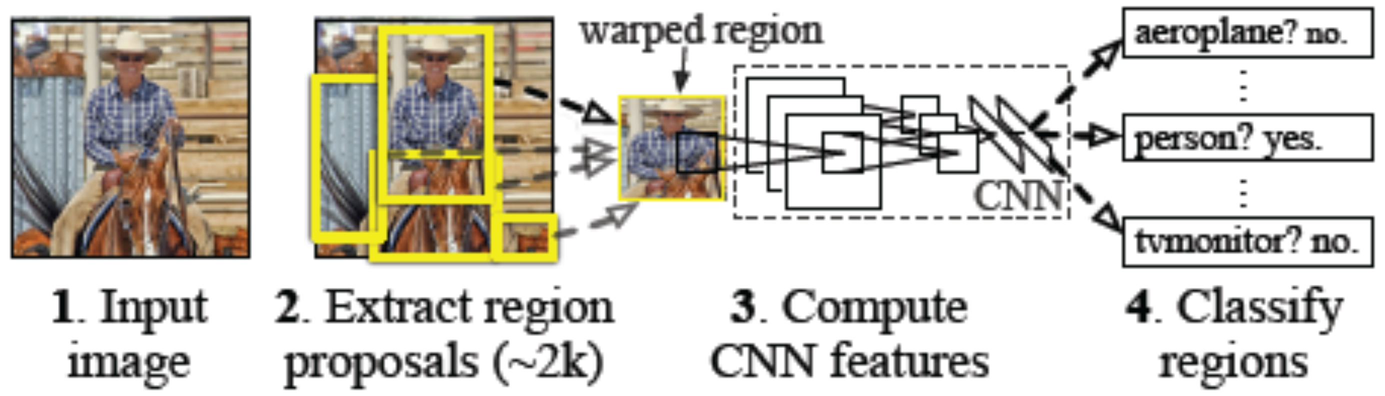
SEMANTIC AND GEOMETRIC SCENE UNDERSTANDING LANDSCAPE

- Shift in Visual-SLAM and Object Detection capabilities
  - **Richer Semantics**: Object Proposals, R-CNN, RoI Pooling/Align
  - **Robust vSLAM:** Sparse and Semi-dense Monocular Reconstruction

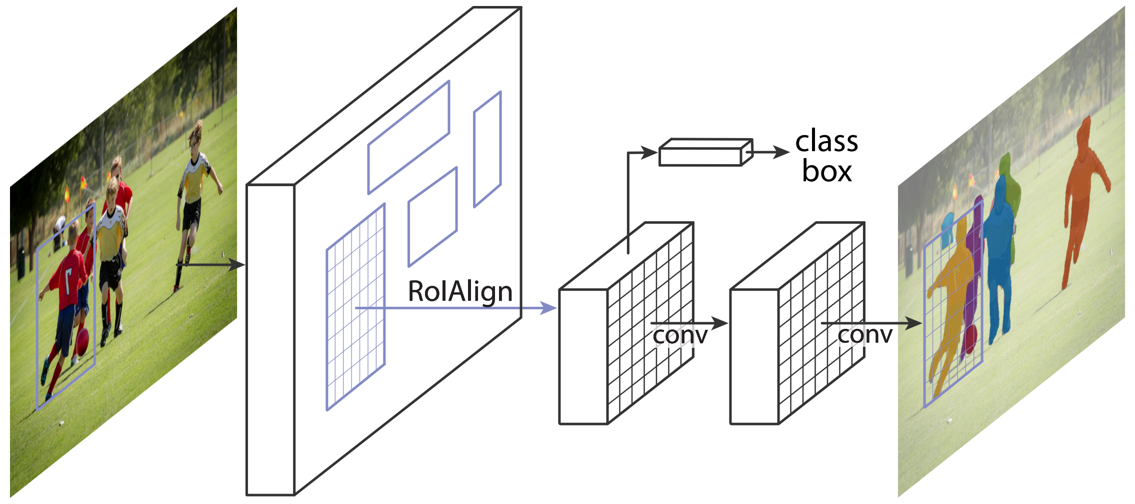
RICHER SEMANTICS



FLAIR (van de Sande 2014)



R-CNN, Fast(er) R-CNN (Girshick et al. 2014-5)



Mask-RCNN (He et al 2017)

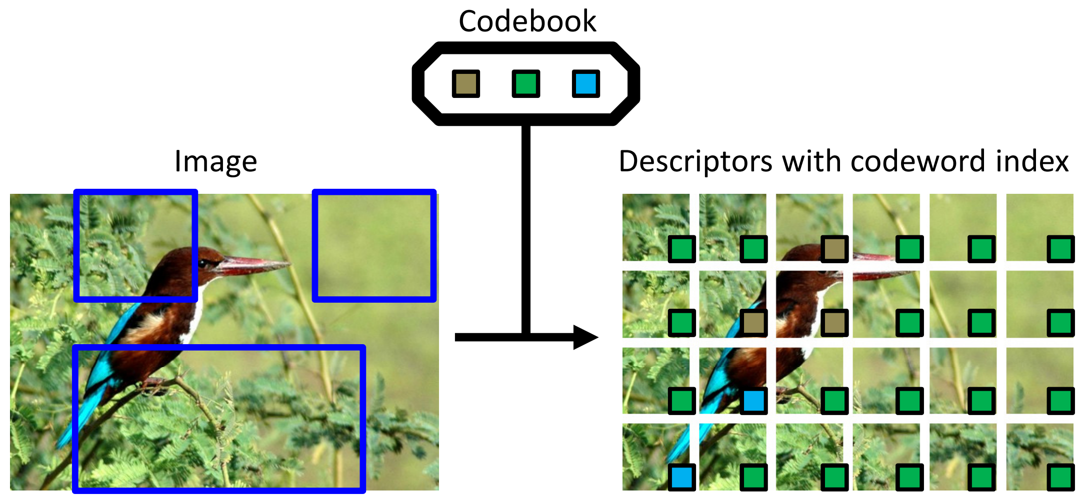
ROBUST vSLAM

Semi-Dense Mapping with ORB-SLAM
Mur-Artal et al. (RSS 2015)

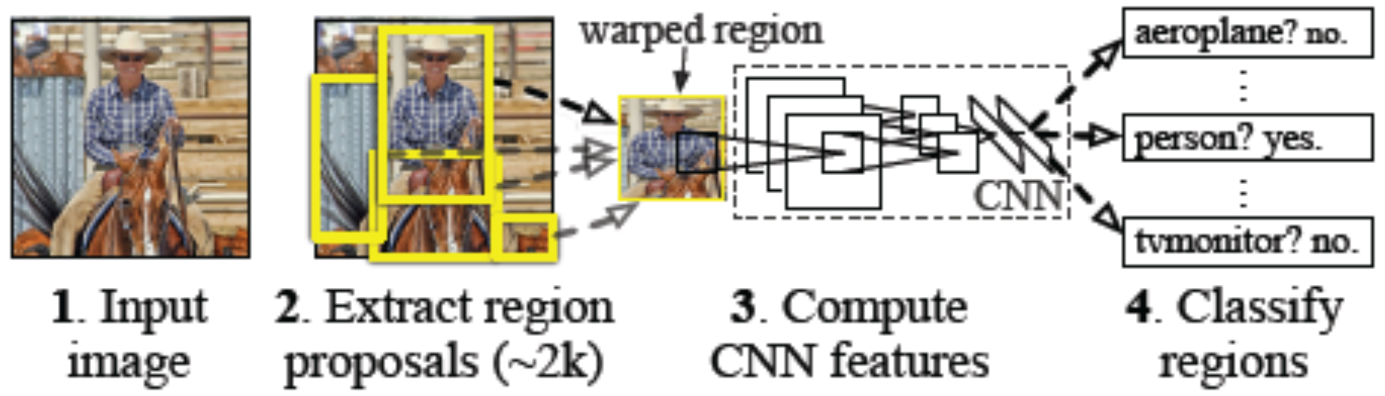
SEMANTIC AND GEOMETRIC SCENE UNDERSTANDING LANDSCAPE

- Shift in Visual-SLAM and Object Detection capabilities
  - **Richer Semantics**: Object Proposals, R-CNN, RoI Pooling/Align
  - **Robust vSLAM:** Sparse and Semi-dense Monocular Reconstruction
  - **Semantic SFM/SLAM:** Semantics measurements for SLAM

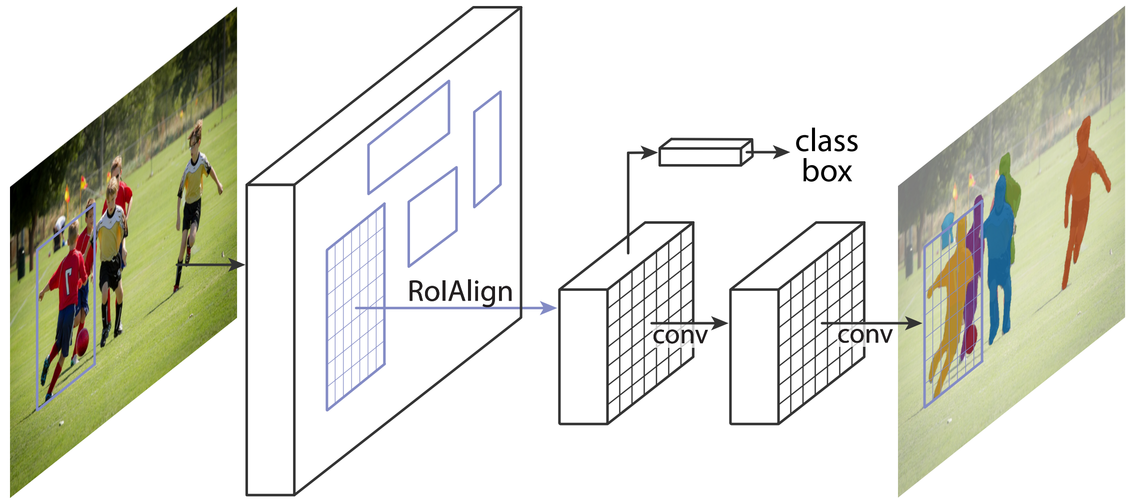
RICHER SEMANTICS



FLAIR (van de Sande 2014)

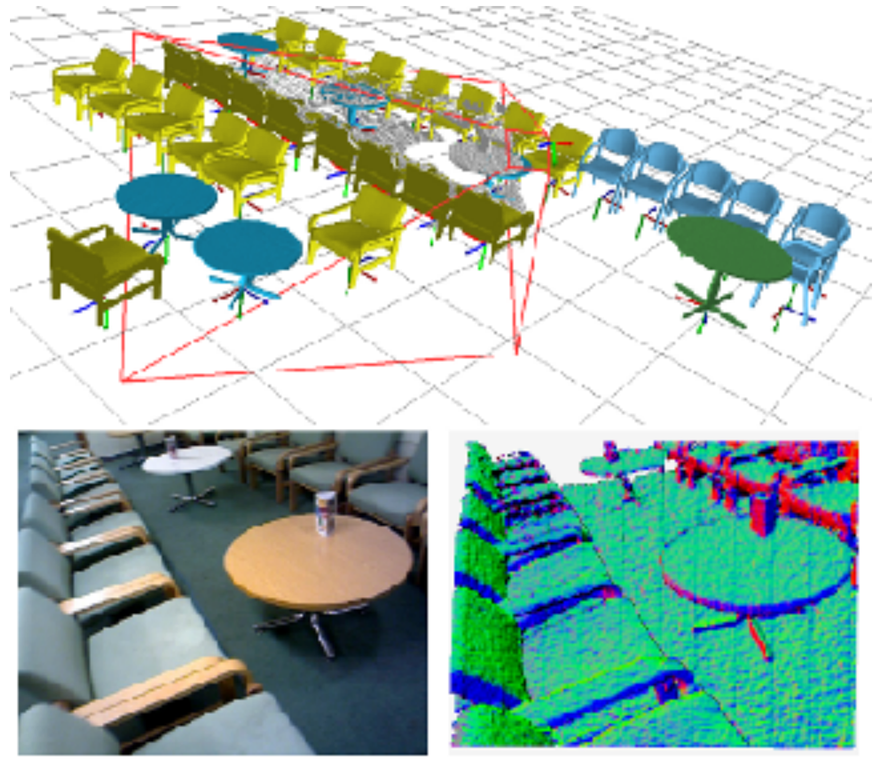


R-CNN, Fast(er) R-CNN (Girshick et al. 2014-5)



Mask-RCNN (He et al 2017)

SEMANTIC SLAM



SLAM++

Salas-Moreno et al. (CVPR 2013)

ROBUST vSLAM

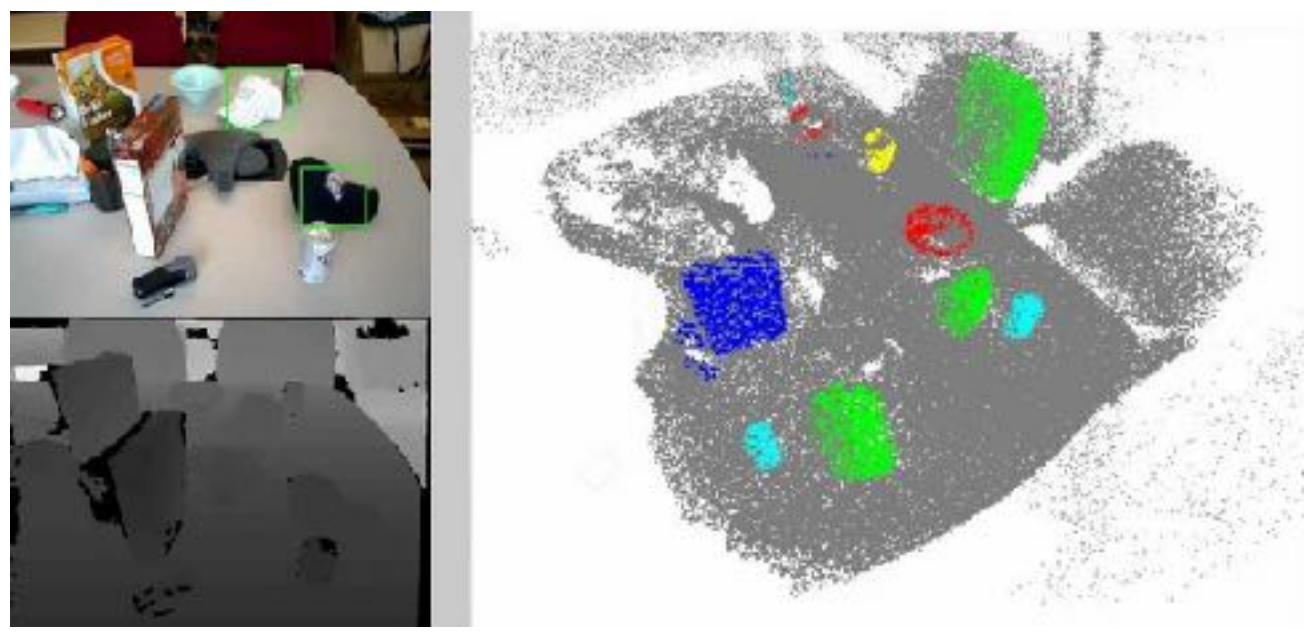
Semi-Dense Mapping with ORB-SLAM

Mur-Artal et al. (RSS 2015)

SEMANTIC AND GEOMETRIC SCENE UNDERSTANDING LANDSCAPE

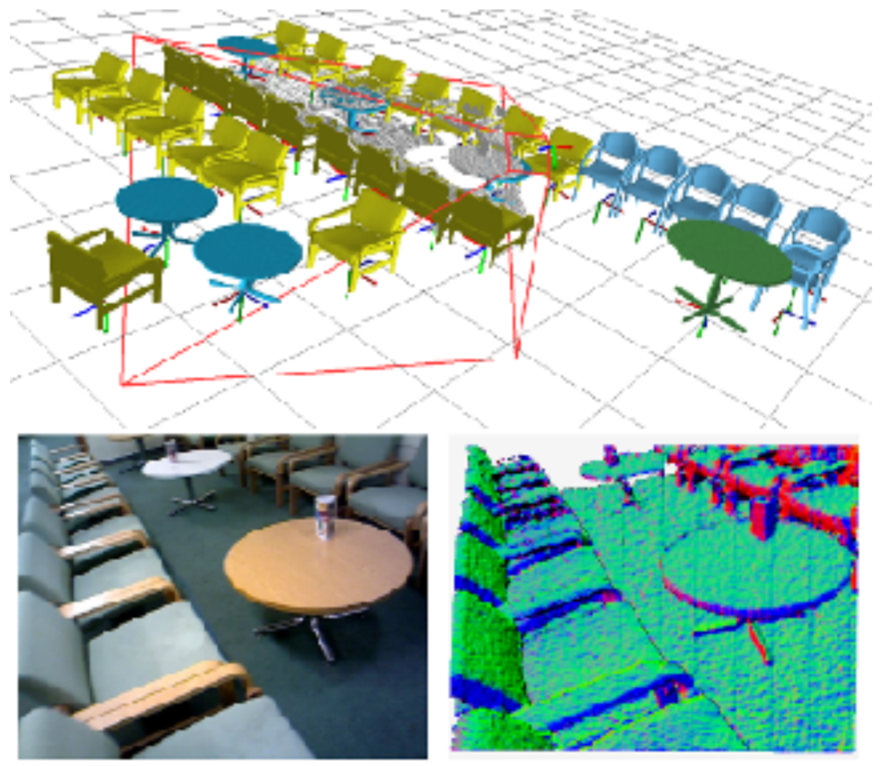
- Shift in Visual-SLAM and Object Detection capabilities
  - **Richer Semantics**: Object Proposals, R-CNN, RoI Pooling/Align
  - **Robust vSLAM:** Sparse and Semi-dense Monocular Reconstruction
  - **Semantic SFM/SLAM:** Semantics measurements for SLAM
  - **RGB-D Detection:** Map-driven detection with RGB-D SLAM

RGB-D DETECTION



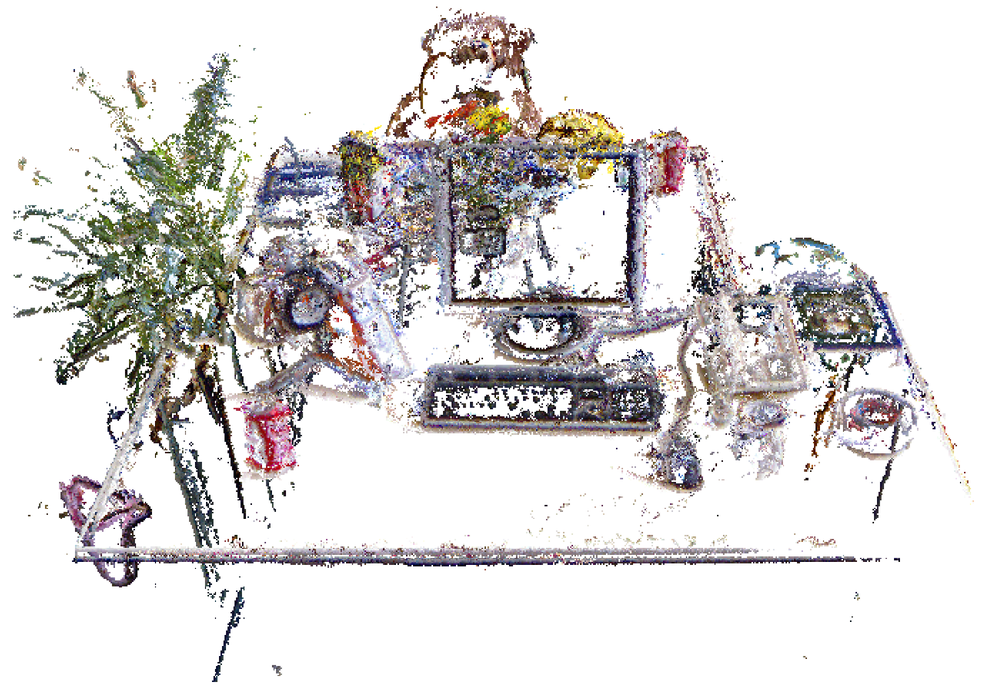
Detection-based Object Labeling in 3D scenes
Lai et al. (ICRA 2012)

SEMANTIC SLAM



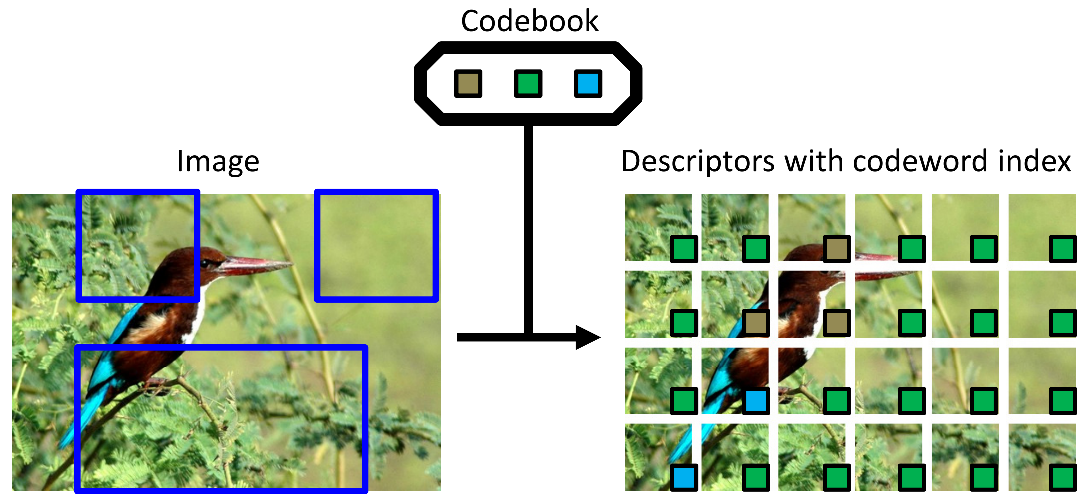
SLAM++
Salas-Moreno et al. (CVPR 2013)

ROBUST vSLAM

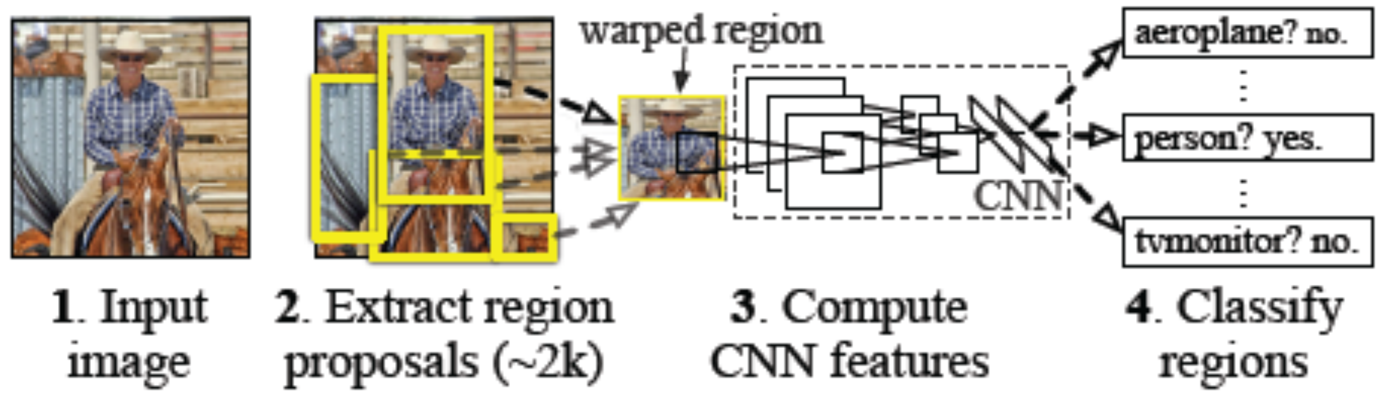


Semi-Dense Mapping with ORB-SLAM
Mur-Artal et al. (RSS 2015)

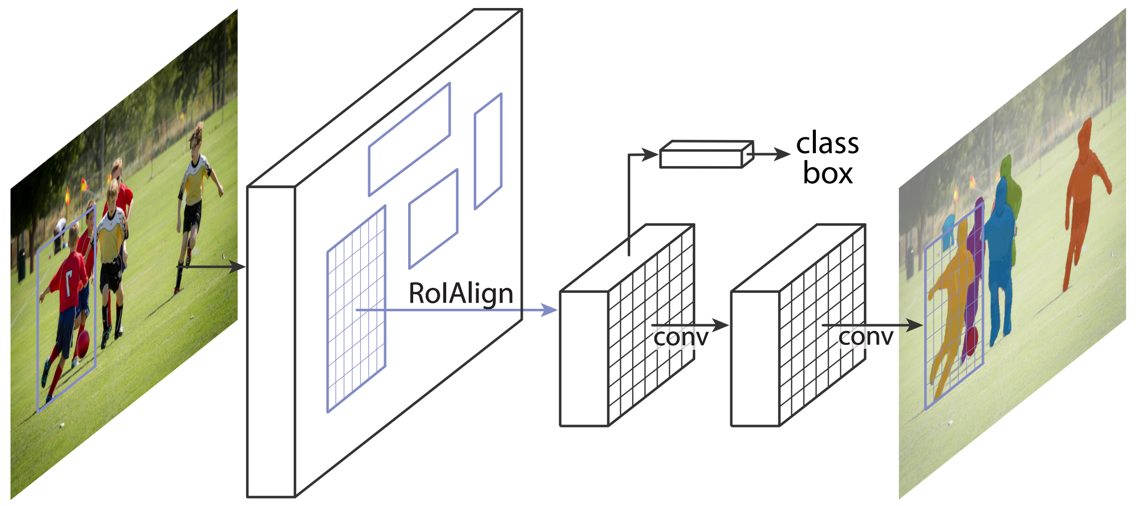
RICHER SEMANTICS



FLAIR (van de Sande 2014)



R-CNN, Fast(er) R-CNN (Girshick et al. 2014-5)

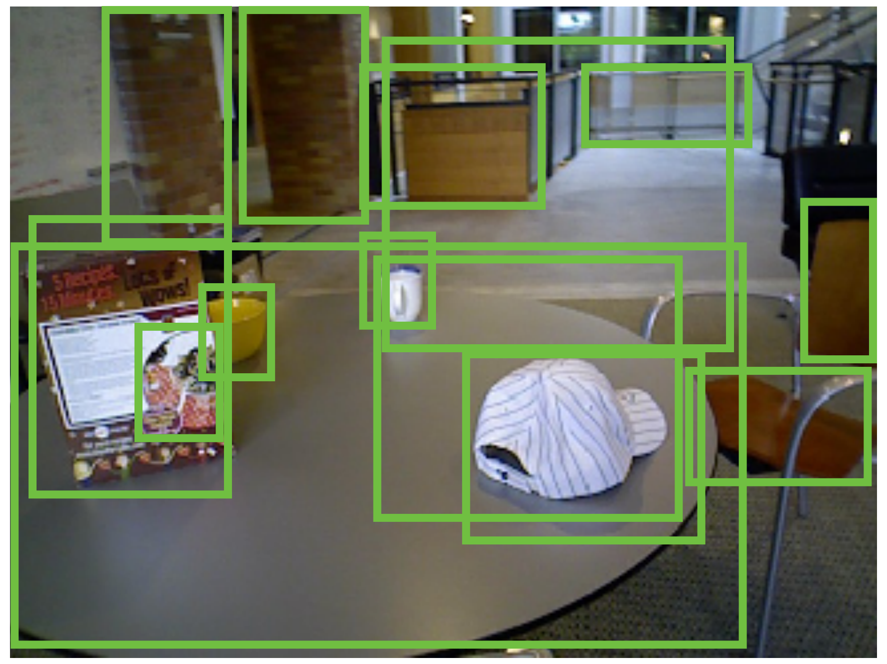


Mask-RCNN (He et al 2017)

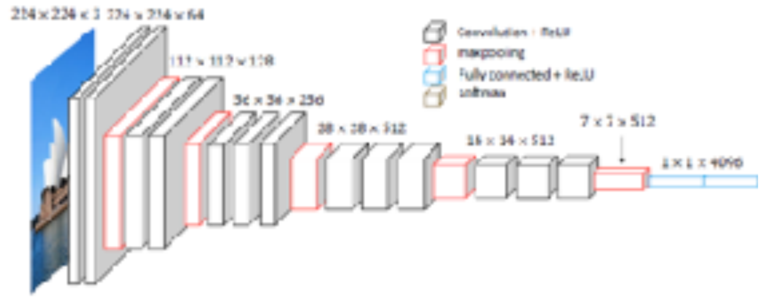
STATE-OF-THE-ART OBJECT RECOGNITION

- **Frame-based object-recognition**
  - Good overall recognition performance
  - Some viewpoint, lighting invariance
  - No memory, context or scene knowledge
  - Spurious false positives

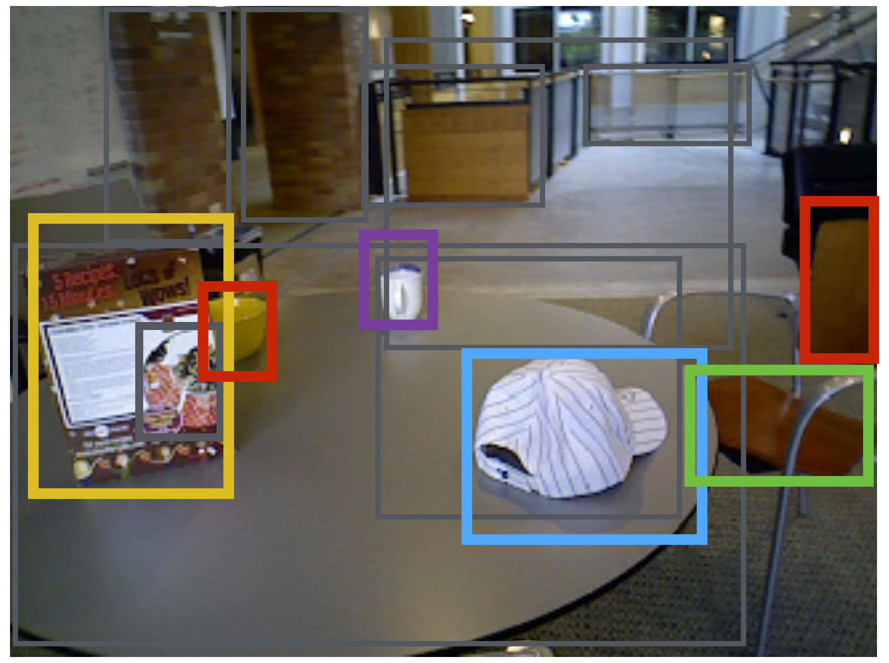
Category-agnostic Object Proposals



Feed-forward CNN



Region-of-Interest Pooling and Classification

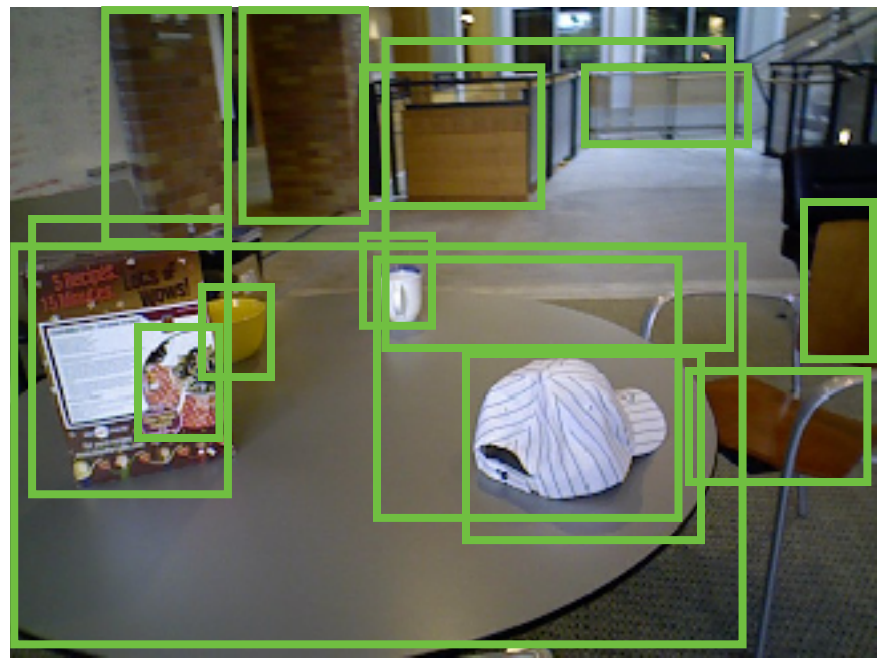


Two Stage Object Recognition

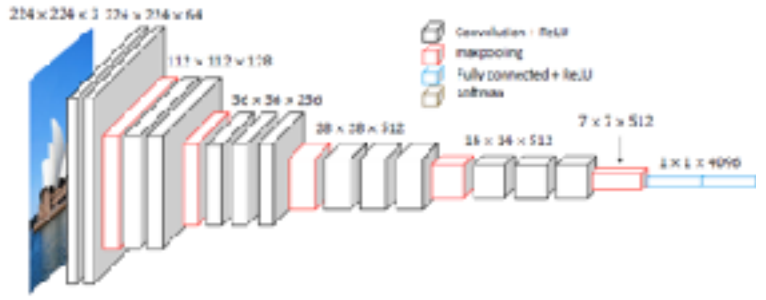
STATE-OF-THE-ART OBJECT RECOGNITION

- **Frame-based object-recognition**
  - Good overall recognition performance
  - Some viewpoint, lighting invariance
  - No memory, context or scene knowledge
  - Spurious false positives

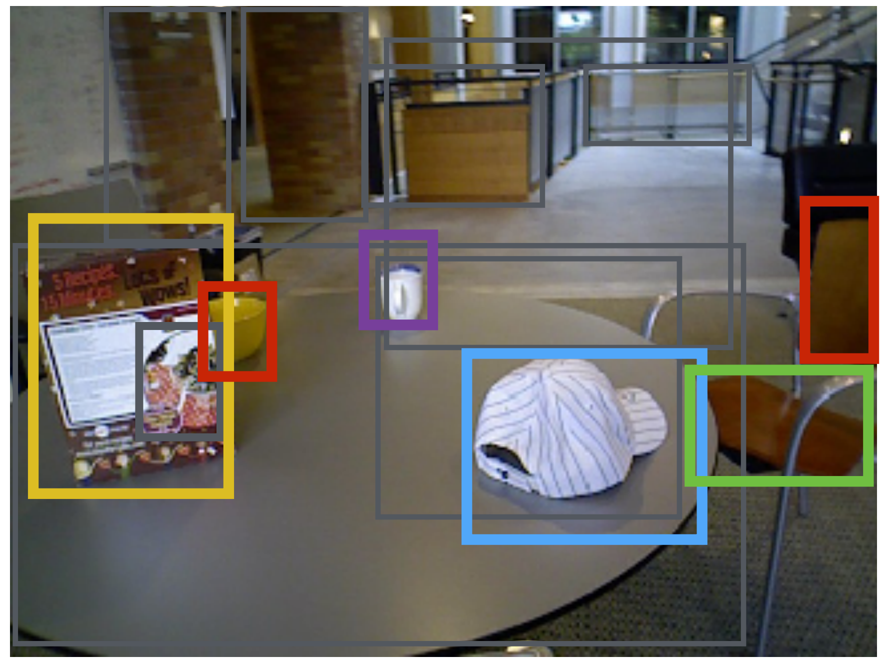
Category-agnostic Object Proposals



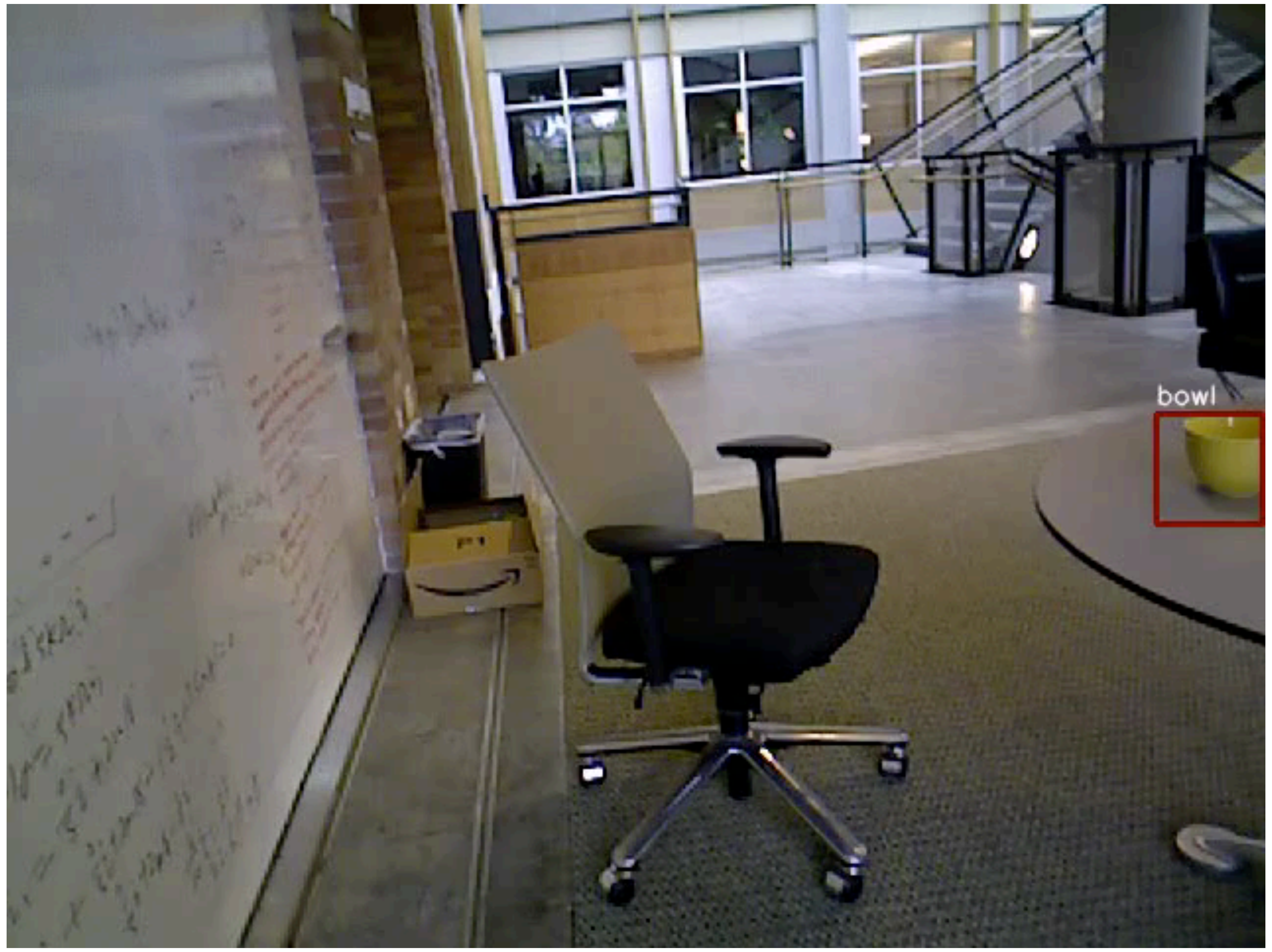
Feed-forward CNN



Region-of-Interest Pooling and Classification



Two Stage Object Recognition



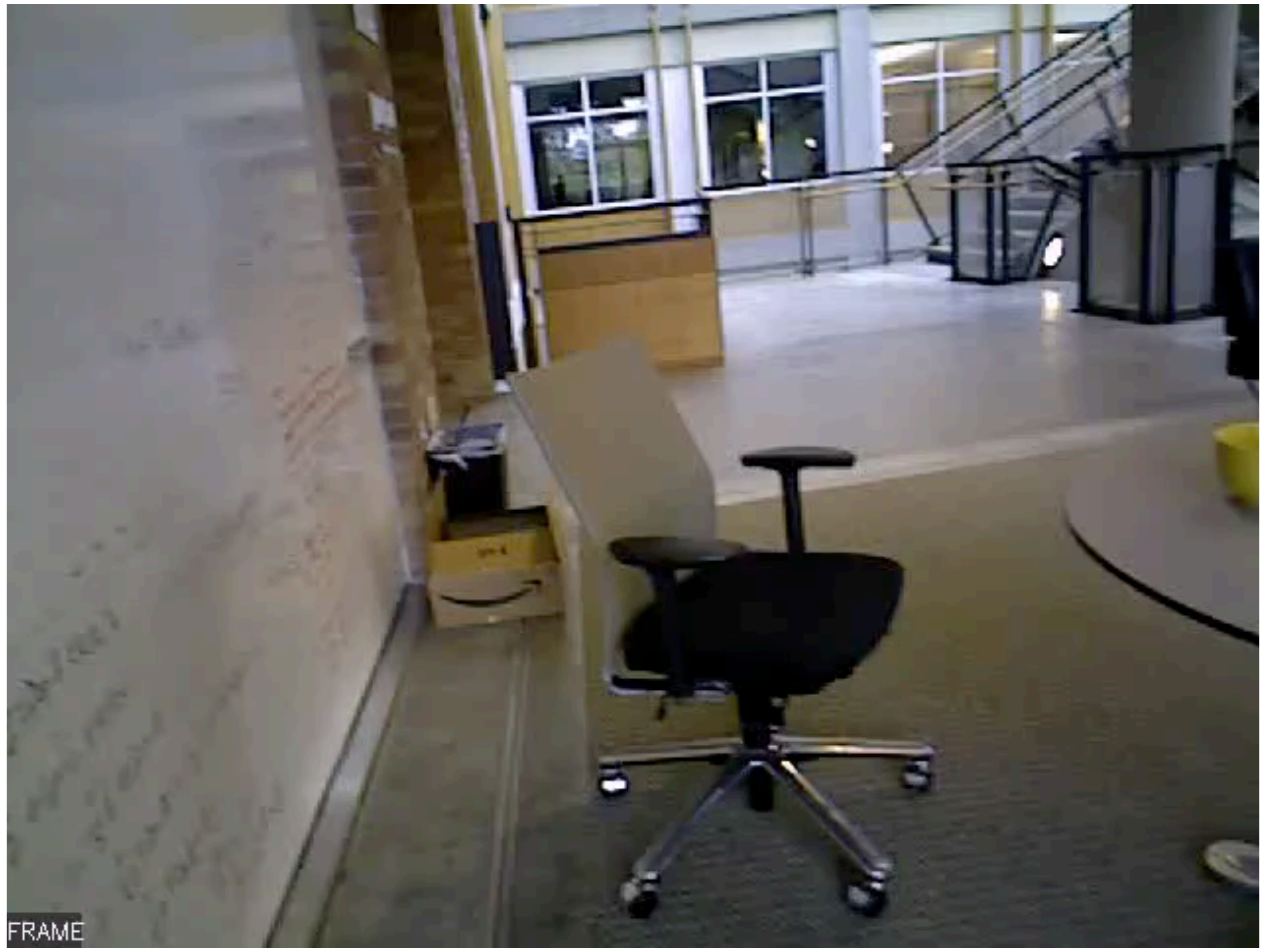
Geodesic Object Proposals with Fast-RCNN

Fast R-CNN, *Girshick 2015*

Geodesic Object Proposals *Krahenbuhl et al 2014*

OBJECT PROPOSALS **with** SLAM

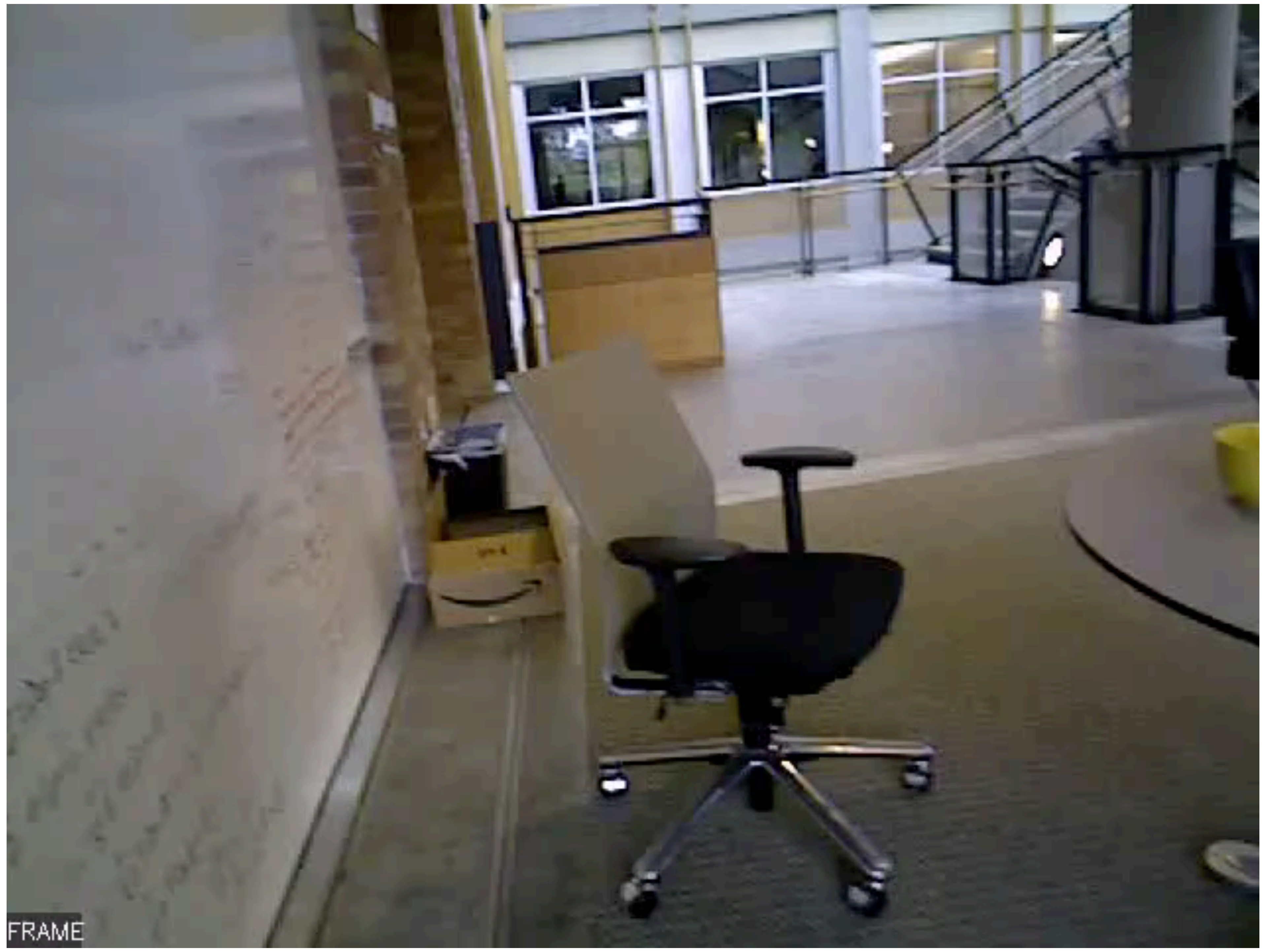
- **SLAM-aware object proposals**
  - Strong overall recognition performance
  - Better viewpoint, lighting invariance
  - Provides some spatial context and knowledge
  - Spurious false positives



OBJECT PROPOSALS **with** SLAM

- **SLAM-aware object proposals**
  - Strong overall recognition performance
  - Better viewpoint, lighting invariance
  - Provides some spatial context and knowledge
  - Spurious false positives

SLAM as a correspondence-engine for spatially-consistent object proposals

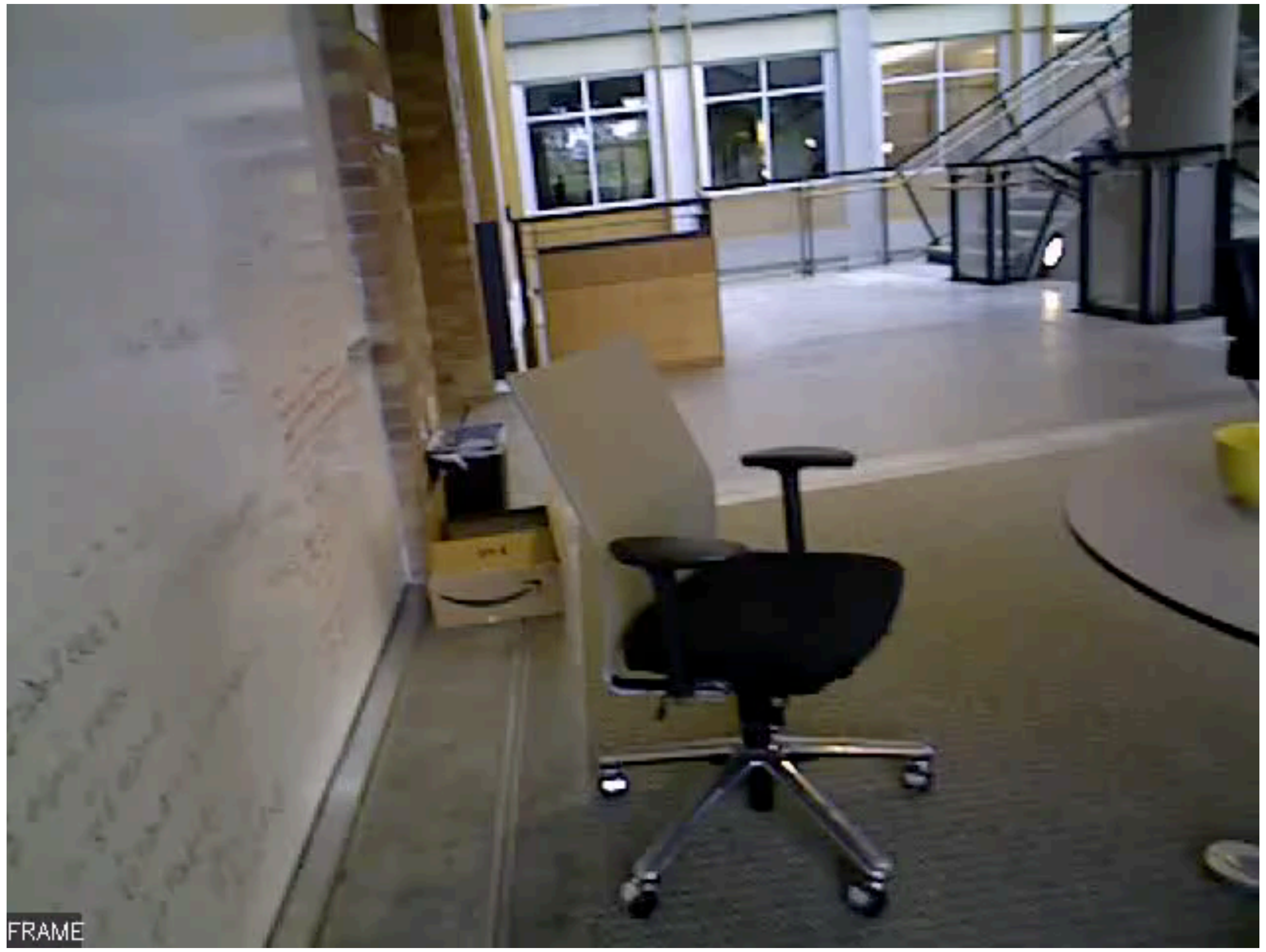


OBJECT PROPOSALS **with** SLAM

- **SLAM-aware object proposals**
  - Strong overall recognition performance
  - Better viewpoint, lighting invariance
  - Provides some spatial context and knowledge
  - Spurious false positives

SLAM as a correspondence-engine for spatially-consistent object proposals

Each frame is individually classified

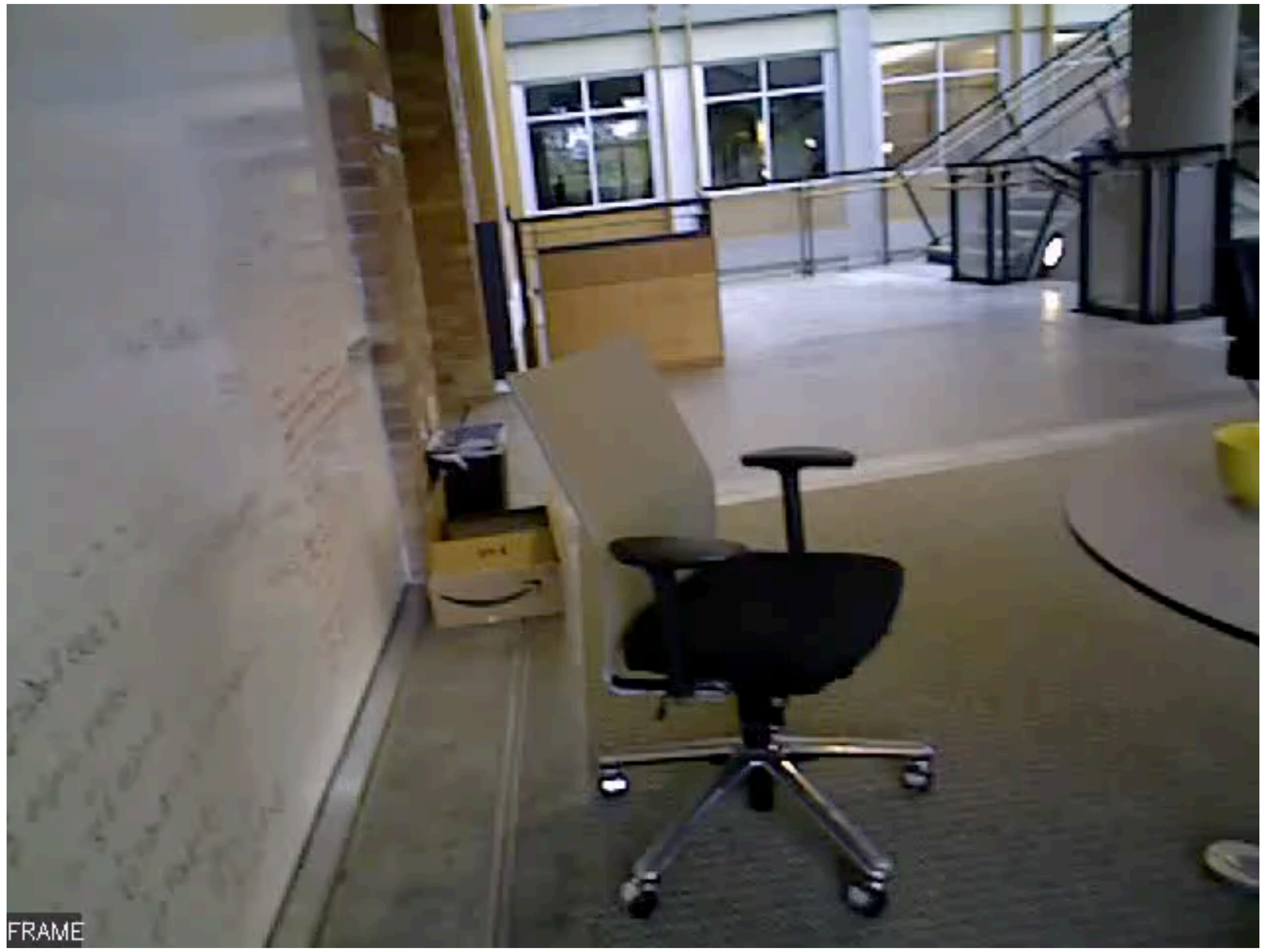


OBJECT PROPOSALS **with** SLAM

- **SLAM-aware object proposals**
  - Strong overall recognition performance
  - Better viewpoint, lighting invariance
  - Provides some spatial context and knowledge
  - Spurious false positives

SLAM as a correspondence-engine for spatially-consistent object proposals

Each frame is individually classified

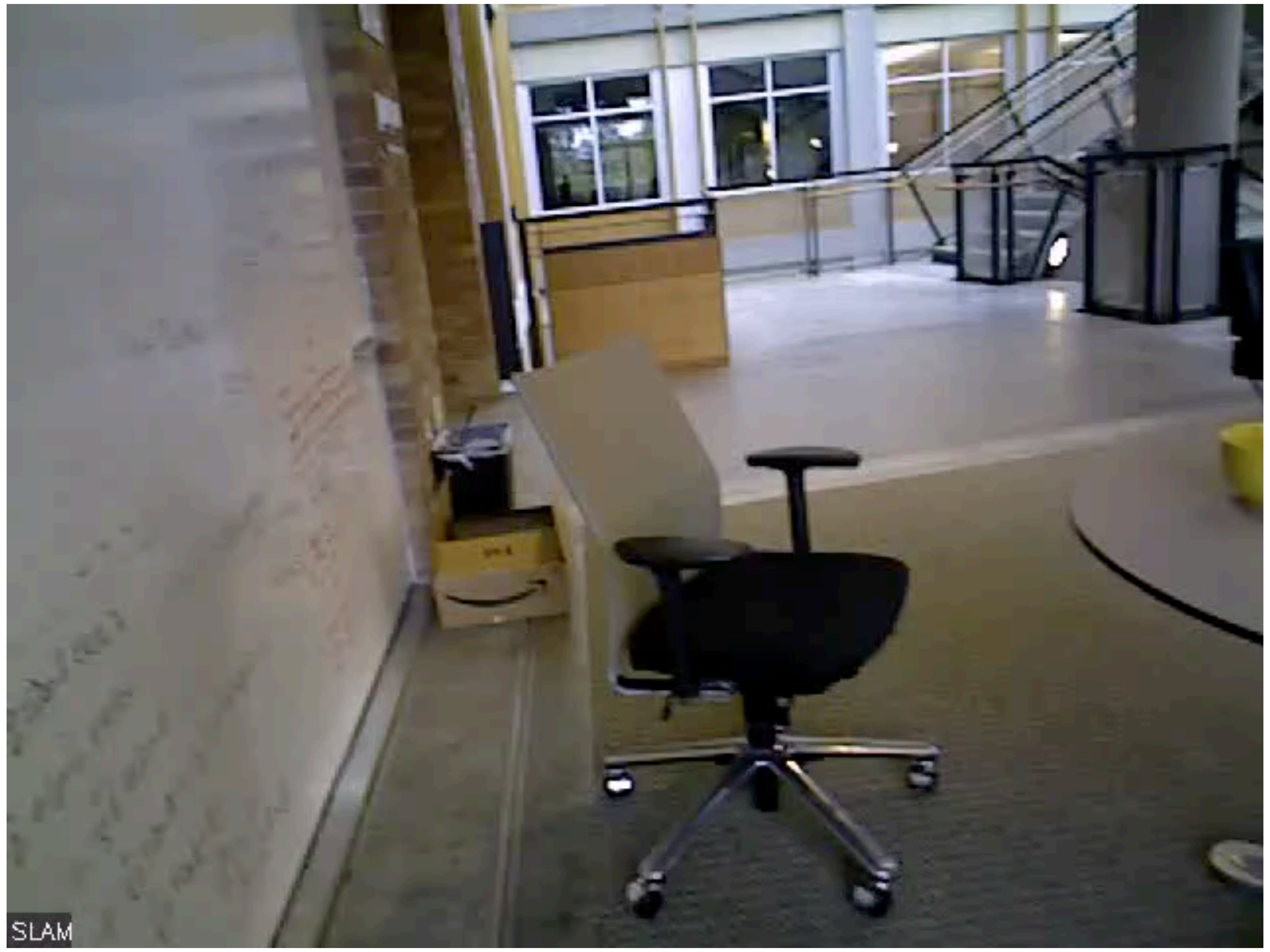


SLAM-aware Object Proposals with Fast-RCNN

Fast R-CNN, *Girshick 2015*

OBJECT RECOGNITION **with** SLAM

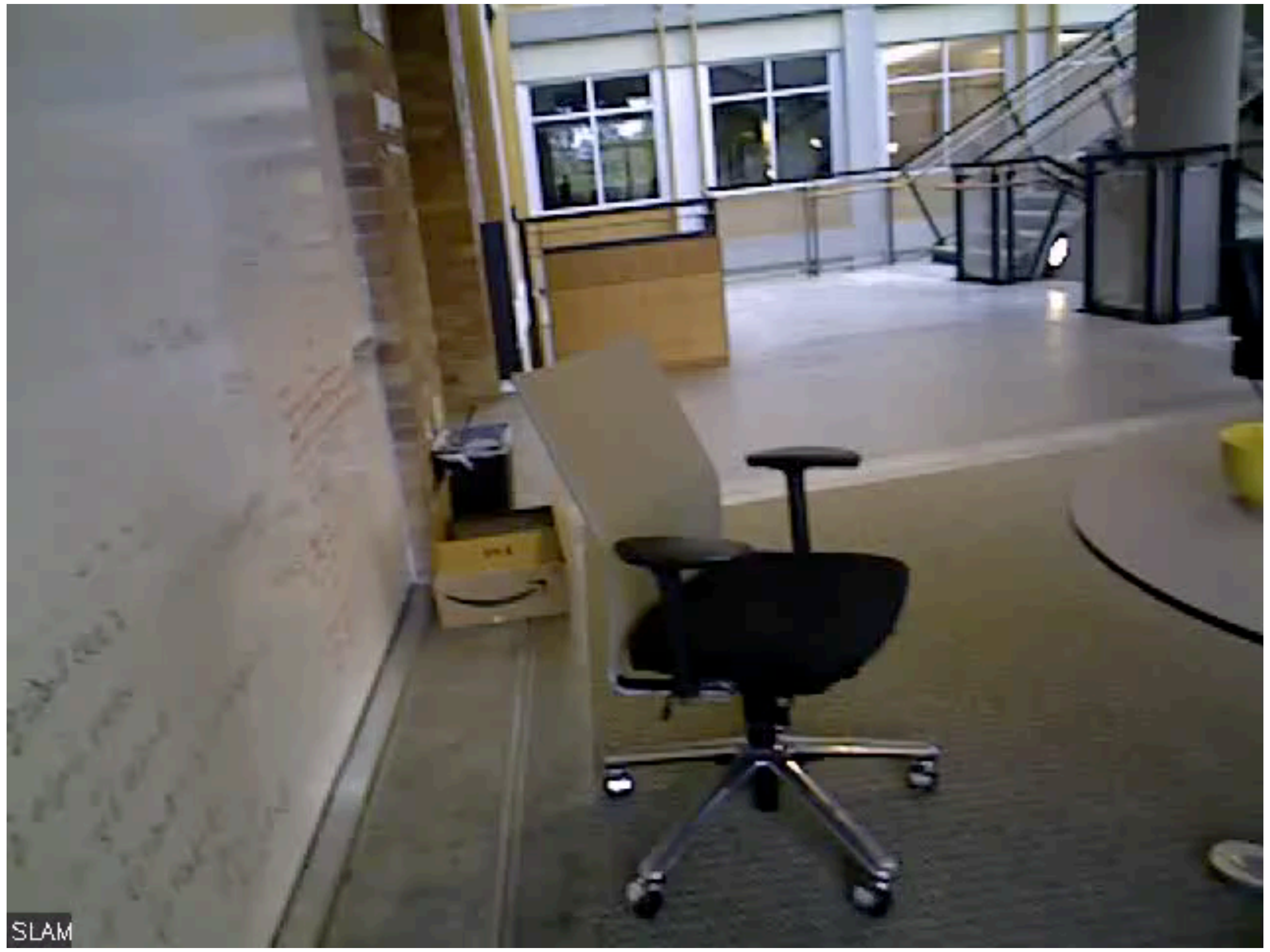
- **SLAM-aware object recognition**
  - Strong overall recognition performance
  - Better viewpoint, lighting invariance
  - Provides spatial context and scene knowledge
  - No spurious false positives
  - Occlusion handling



OBJECT RECOGNITION **with** SLAM

- **SLAM-aware object recognition**
  - Strong overall recognition performance
  - Better viewpoint, lighting invariance
  - Provides spatial context and scene knowledge
  - No spurious false positives
  - Occlusion handling

SLAM as a correspondence-engine for spatially-consistent object proposals

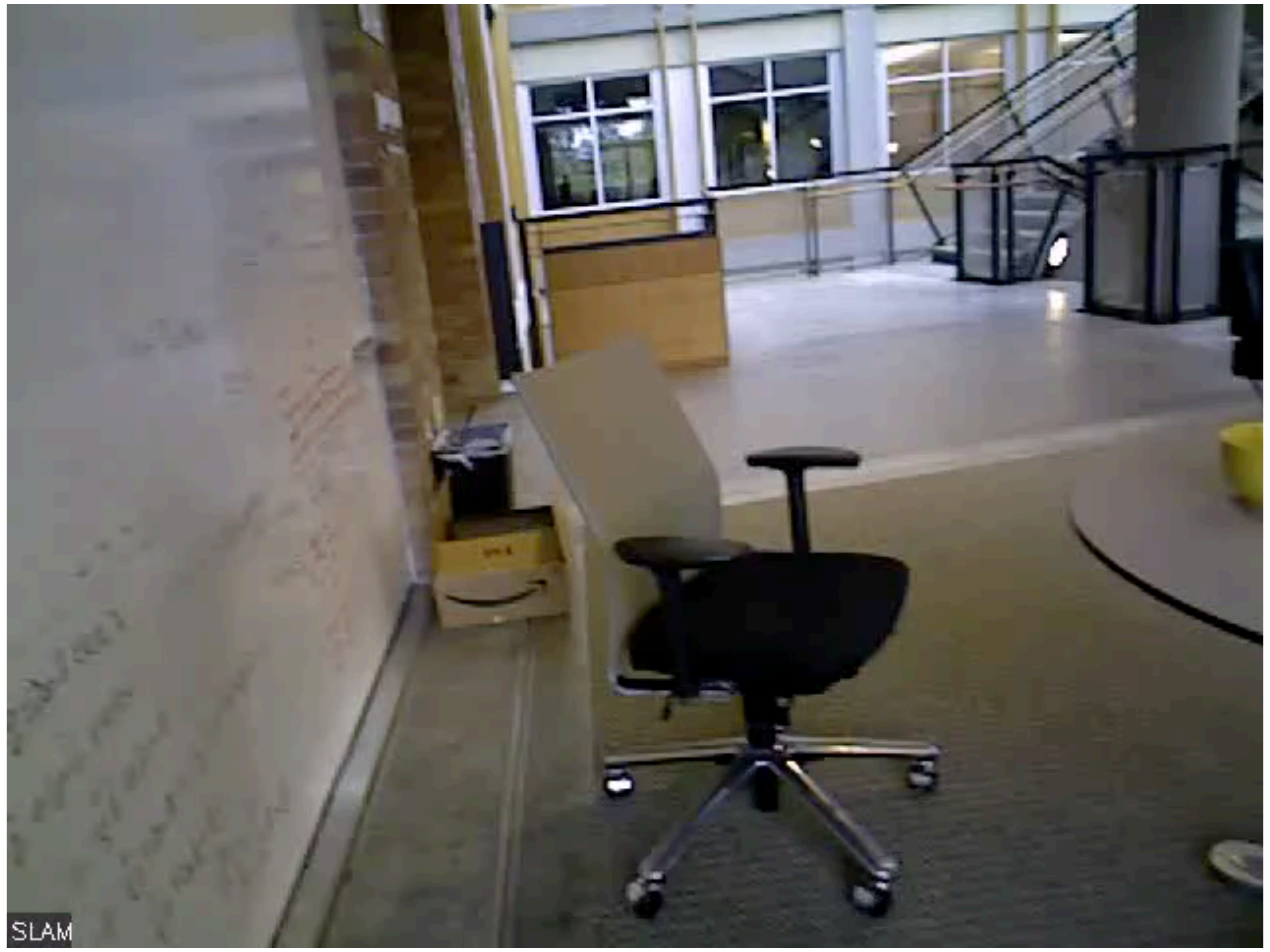


OBJECT RECOGNITION **with** SLAM

- **SLAM-aware object recognition**
  - Strong overall recognition performance
  - Better viewpoint, lighting invariance
  - Provides spatial context and scene knowledge
  - No spurious false positives
  - Occlusion handling

SLAM as a correspondence-engine for spatially-consistent object proposals

Object evidence is aggregated across all views, as enabled by the SLAM-aware system



SLAM-aware object proposal and evidence aggregation with Fast-RCNN
Fast R-CNN, *Girshick 2015*

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

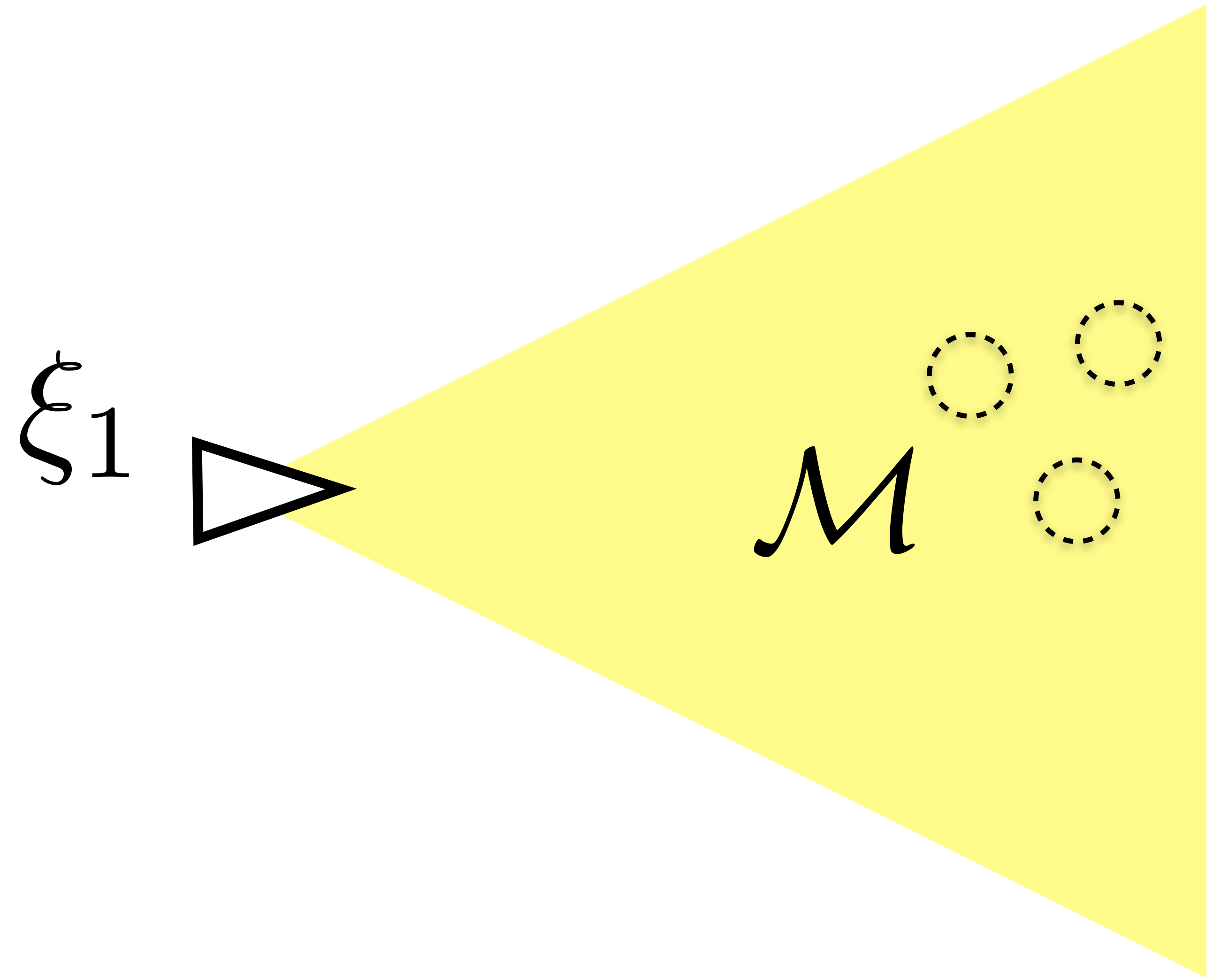
KEY CONCEPT

SLAM-aware

$$\{\xi, \mathcal{M}\}$$

ξ Keyframes

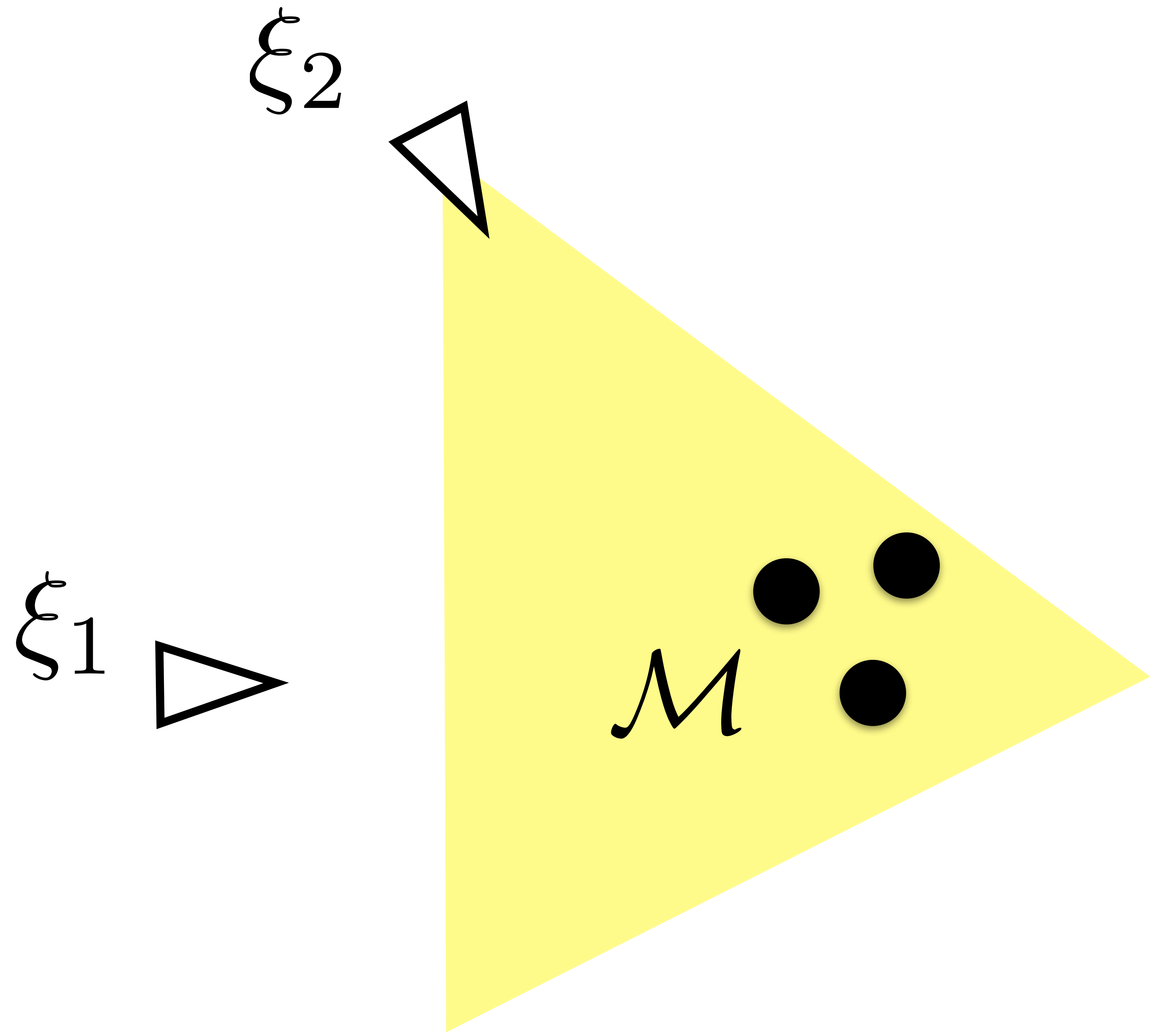
$\mathcal{M}$ Map



Initialization

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT



SLAM-aware

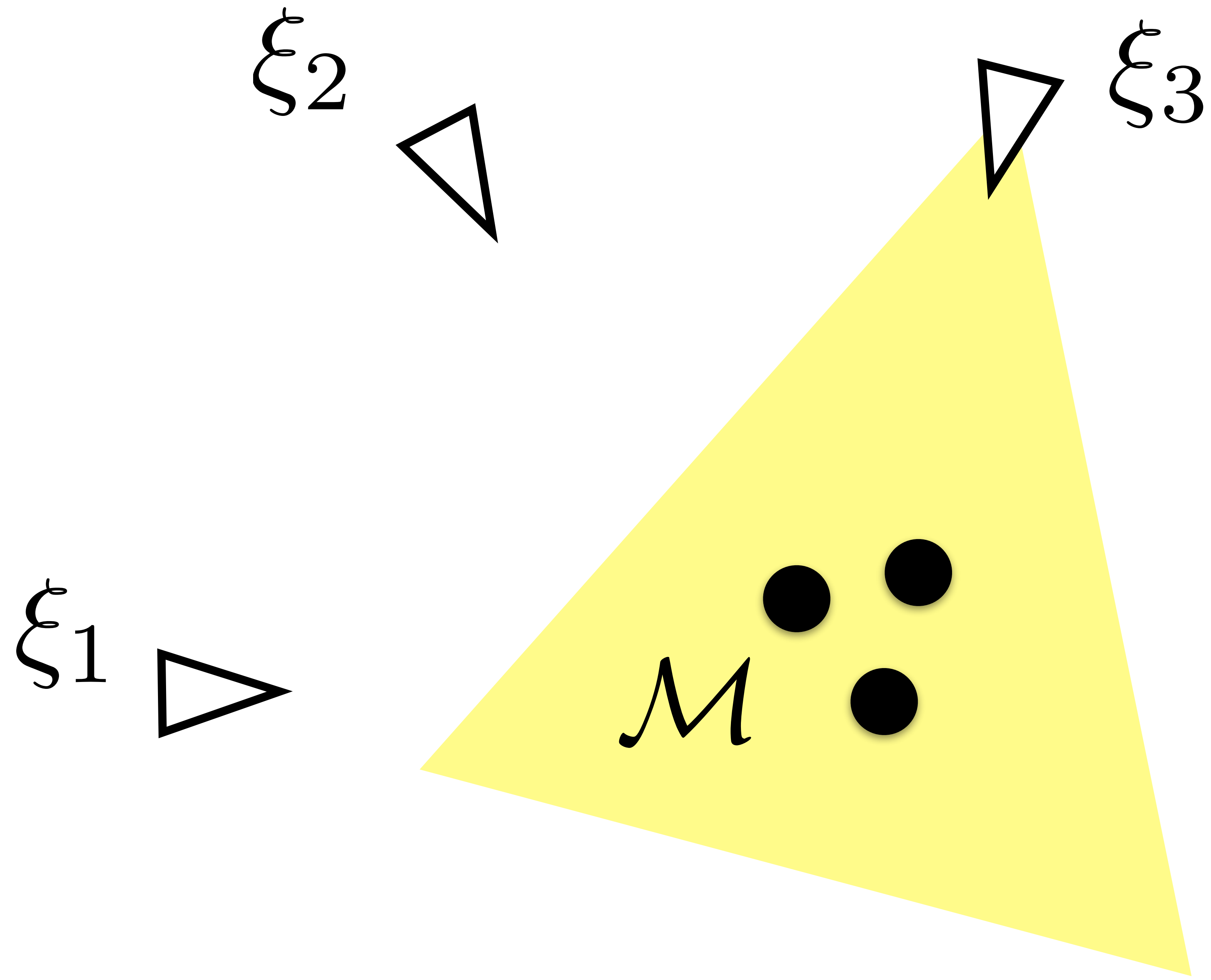
$$\{\xi, \mathcal{M}\}$$

ξ	$\mathcal{M}$
Keyframes	Map

Reduced ambiguity and improved reconstruction with more views

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT



SLAM-aware

$$\{\xi, \mathcal{M}\}$$

ξ	$\mathcal{M}$
Keyframes	Map

Distinct object views for classification via keyframe selection

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

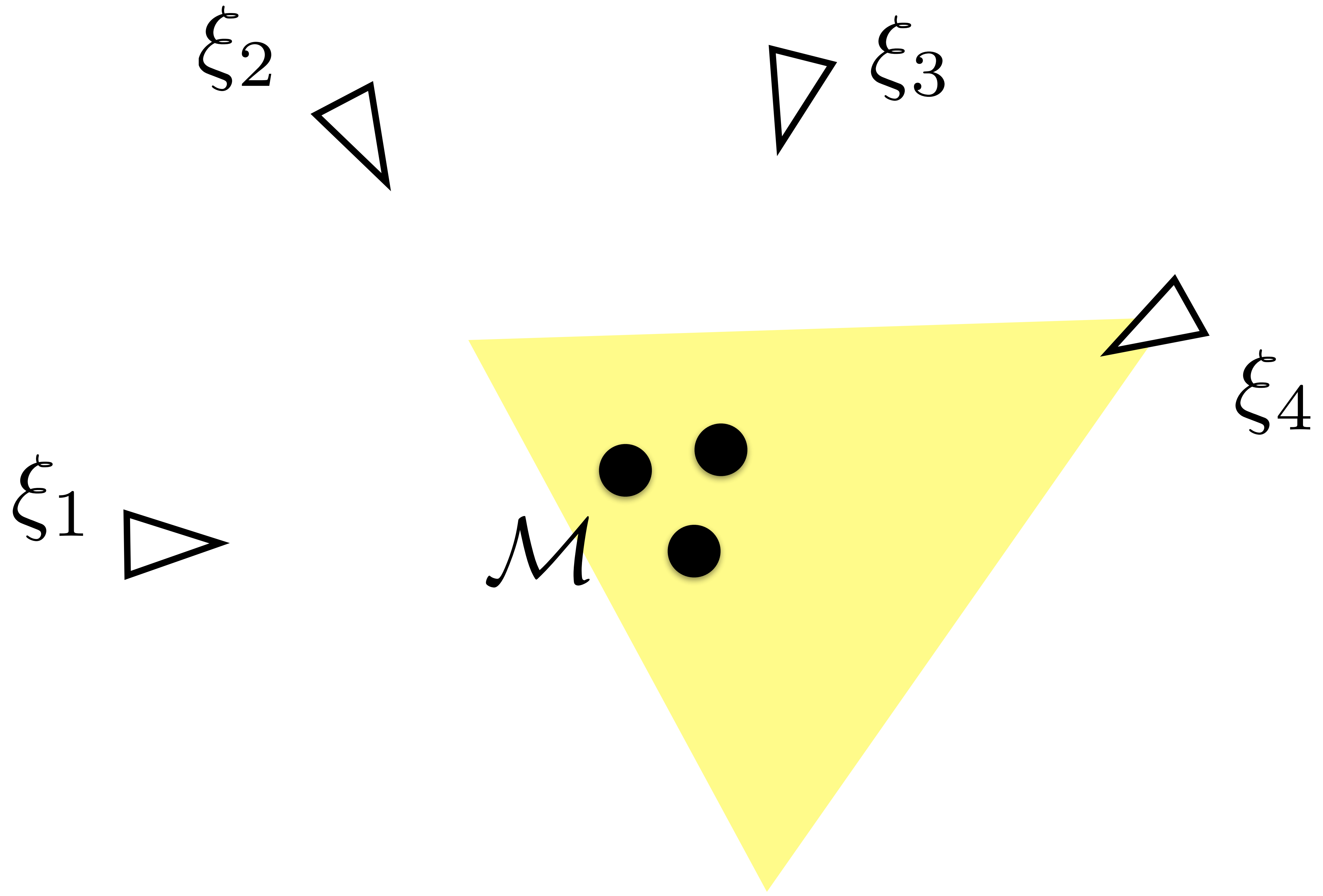
KEY CONCEPT

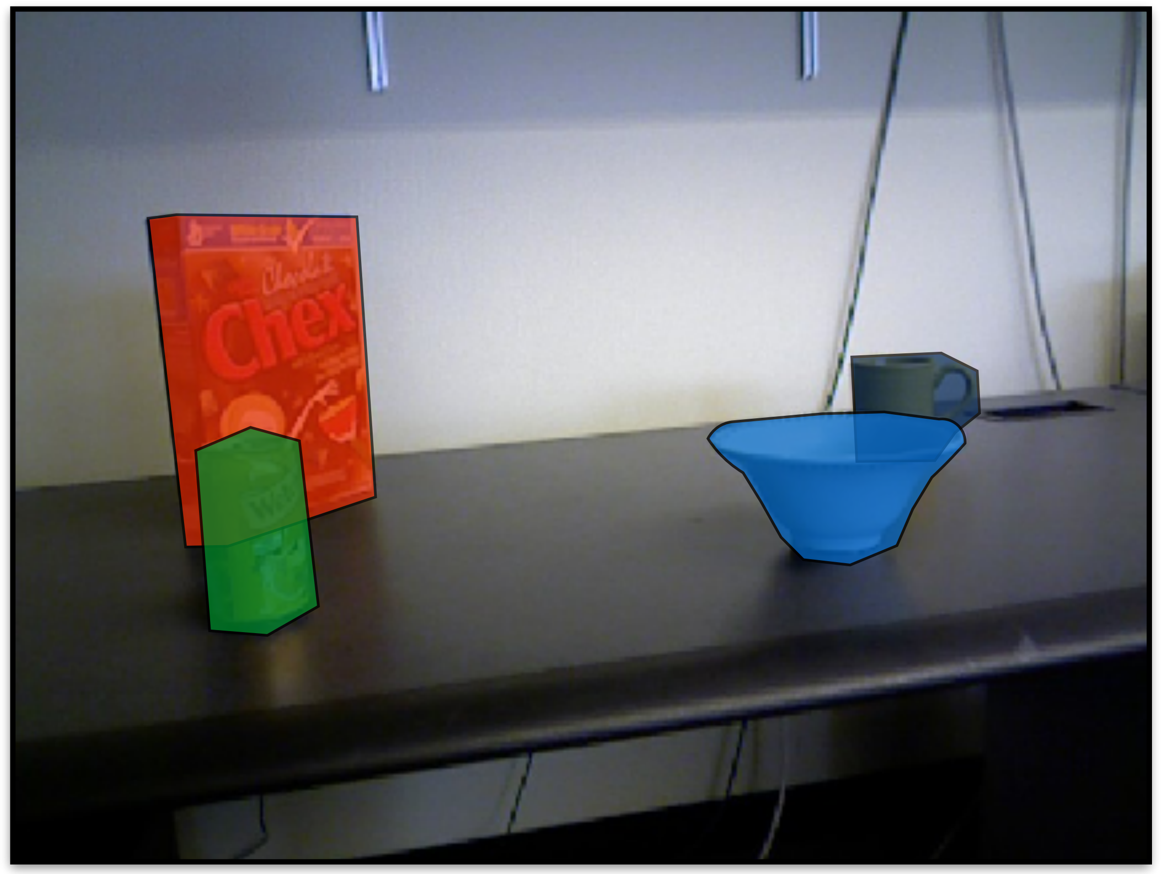
SLAM-aware

$$\{\xi, \mathcal{M}\}$$

ξ Keyframes

$\mathcal{M}$ Map





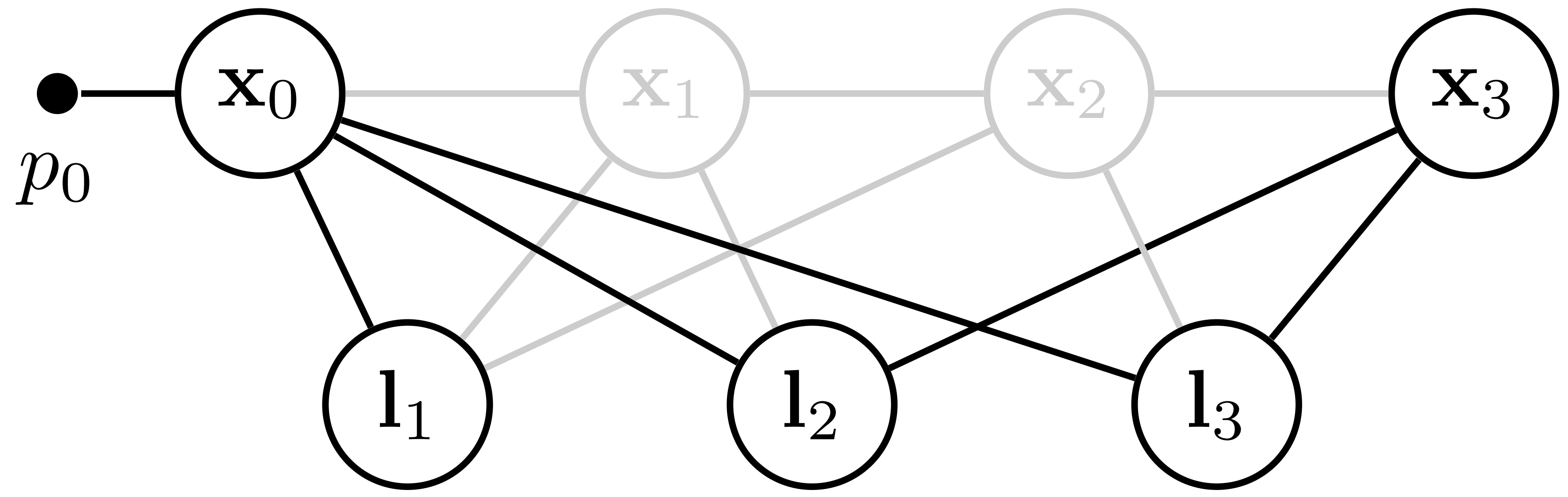
Occlusions require special treatment

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT

Keyframe-based Visual-SLAM

Bundle Adjustment



$$\mathbf{X}^*, \mathbf{L}^* = \arg\max_{\mathbf{X},\mathbf{L}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}_\mathbf{l})$$

$$= \arg\min_{\mathbf{X}} \sum_{k=1}^{K} \|h_k(\mathbf{x}_{ik}, \mathbf{l}_{jk}) - \mathbf{z}_k\|^2_{\Sigma_k}$$

SLAM-aware

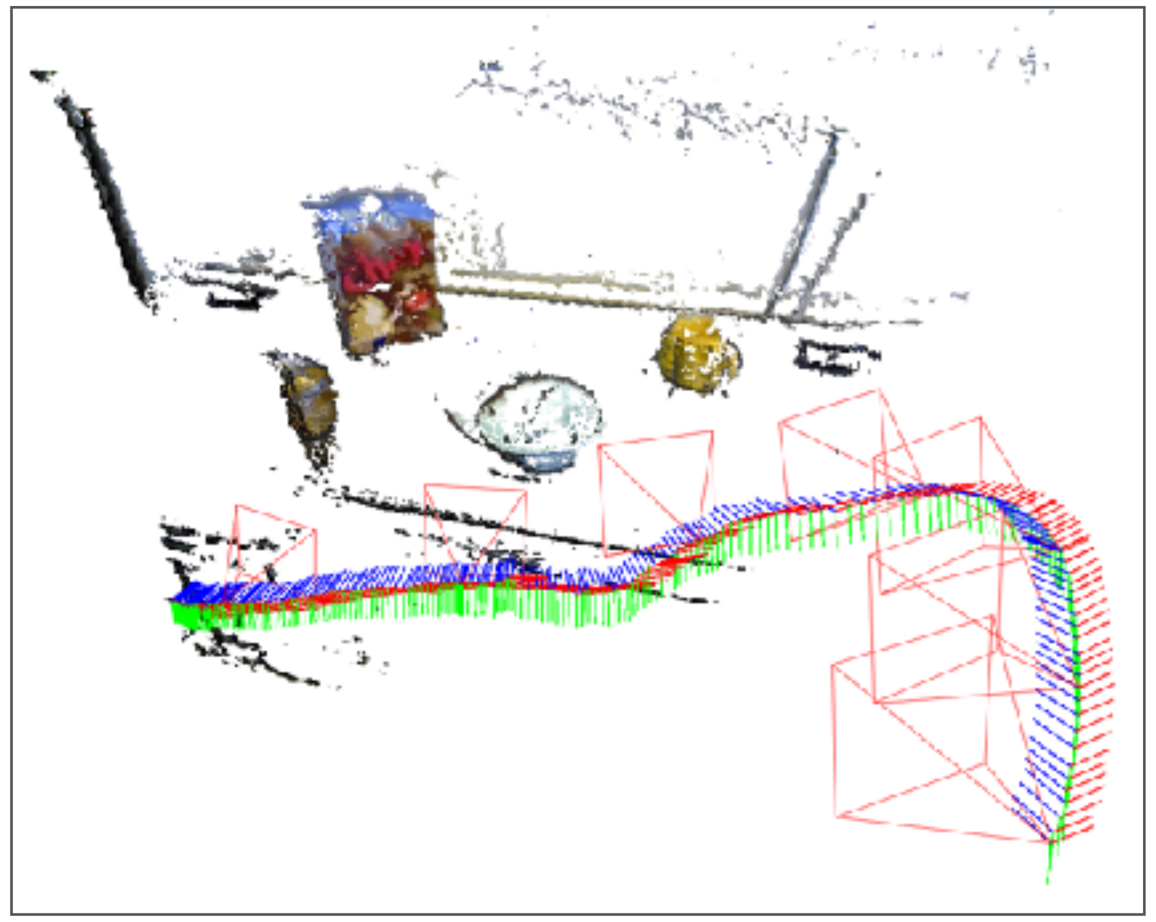
$$\{\xi, \mathcal{M}\}$$

ξ Keyframes

$\mathcal{M}$ Map

ORB-SLAM, *Mur-Artal et al 2015*

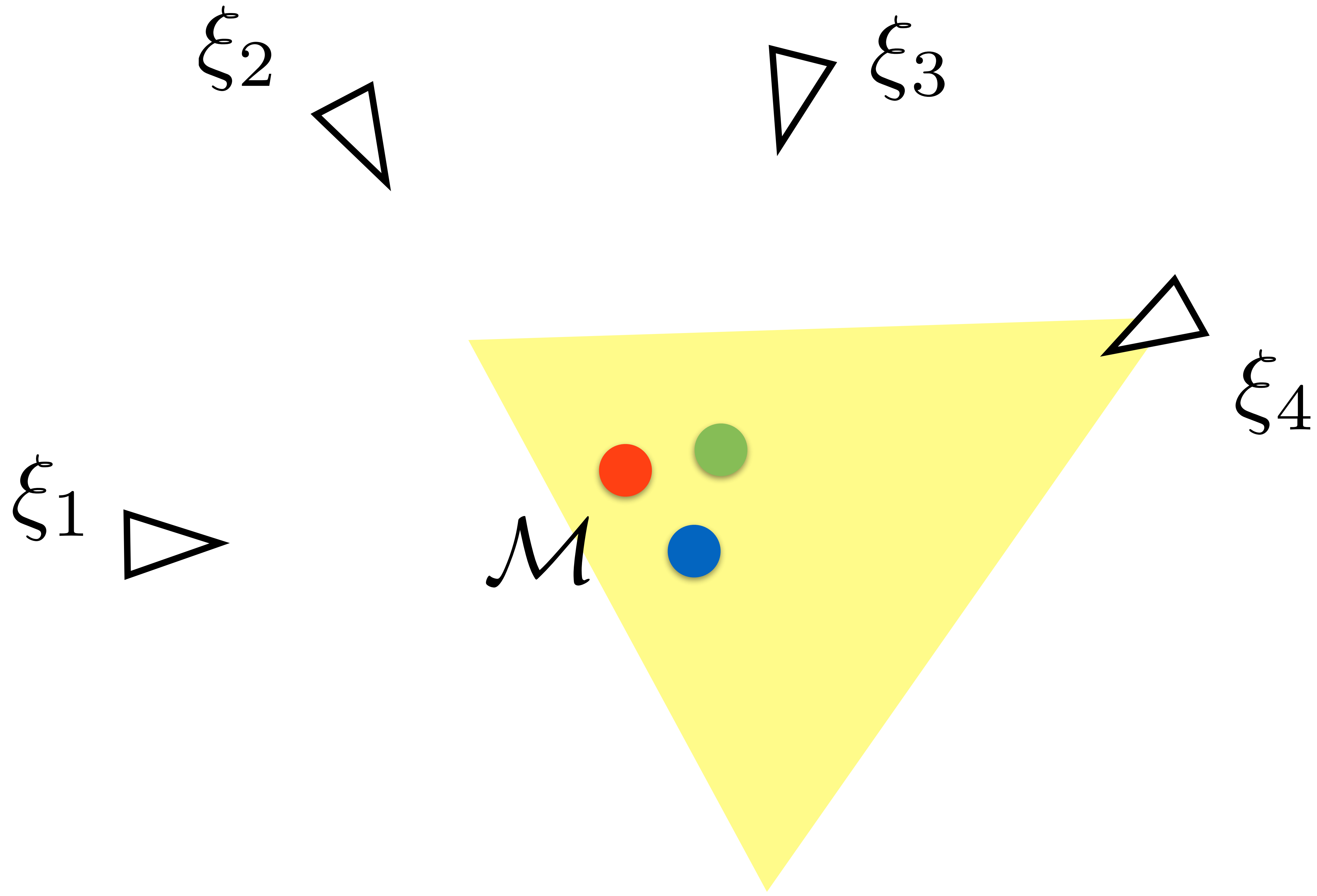
SVO: Depth Filter, *Forster et al 2014*



Semi-dense reconstruction with keyframes

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

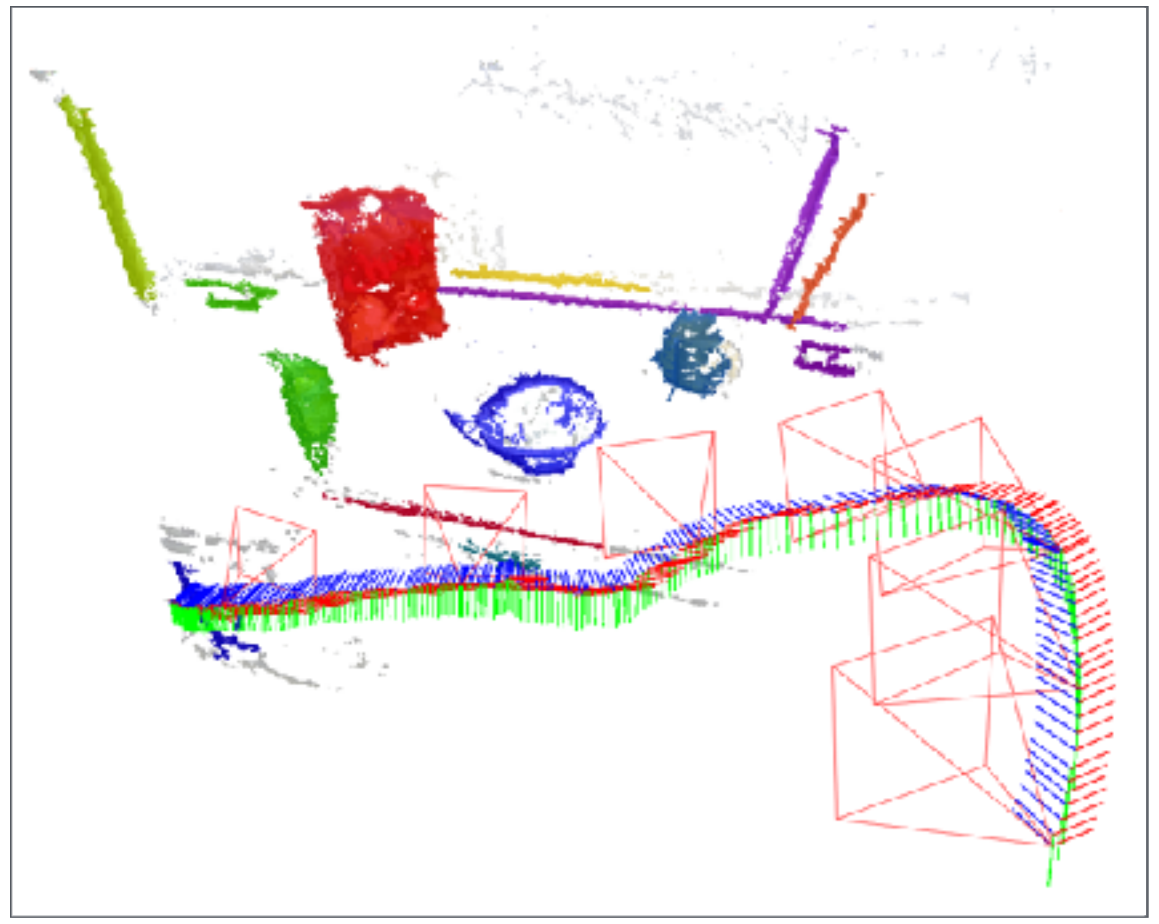
KEY CONCEPT



SLAM-aware

$$\{\xi, \mathcal{M}\}$$

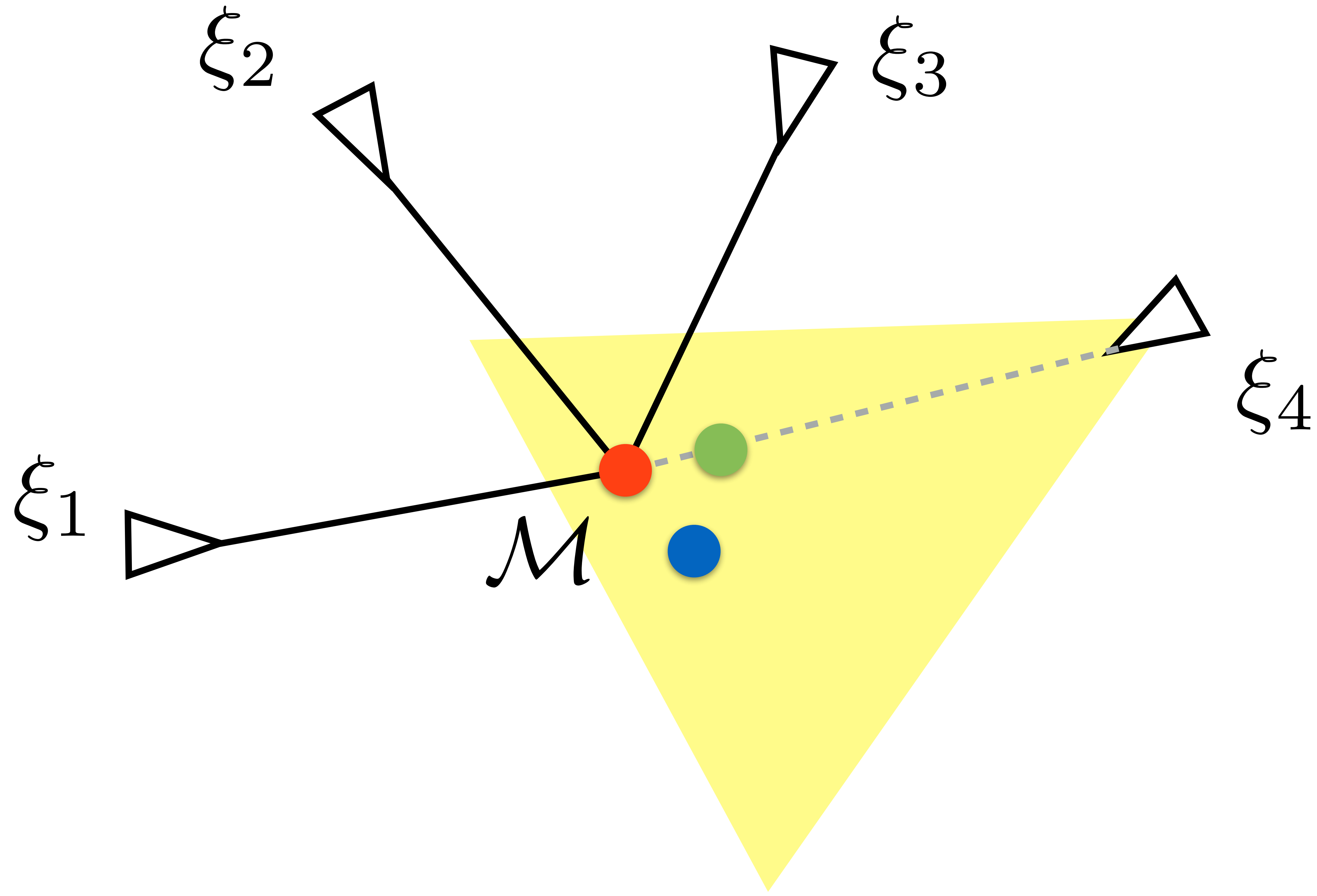
ξ	$\mathcal{M}$
Keyframes	Map



Semi-dense Reconstruction-driven Object Proposals

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT



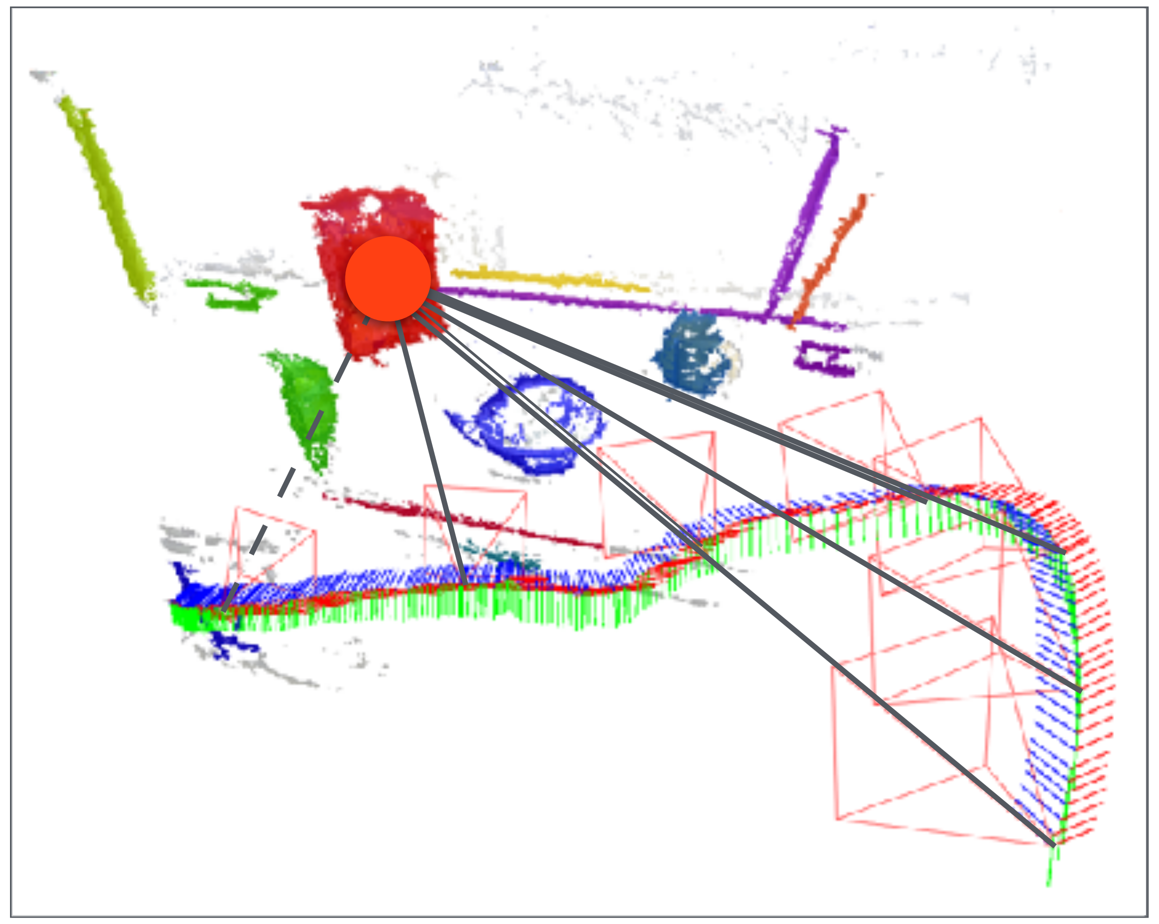
SLAM-aware

$$\{\xi, \mathcal{M}\}$$

ξ Keyframes

$\mathcal{M}$ Map

Goal: Determine most likely semantic label given all non-occluding views



Occlusion-aware Proposal Description

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT

For each object proposal

$$o^j \in \mathcal{O}$$

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT

We first compute the bounding box of the object proposal o^j onto the keyframe ξ_k

For each object proposal $o^j \in \mathcal{O}$

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT

We first compute the bounding box of
the object proposal $\mathcal{O}^j$ onto the keyframe ξ_k

For each object proposal
$\mathcal{O}^j \in \mathcal{O}$

RoI Pooling and Description using Fast R-CNN

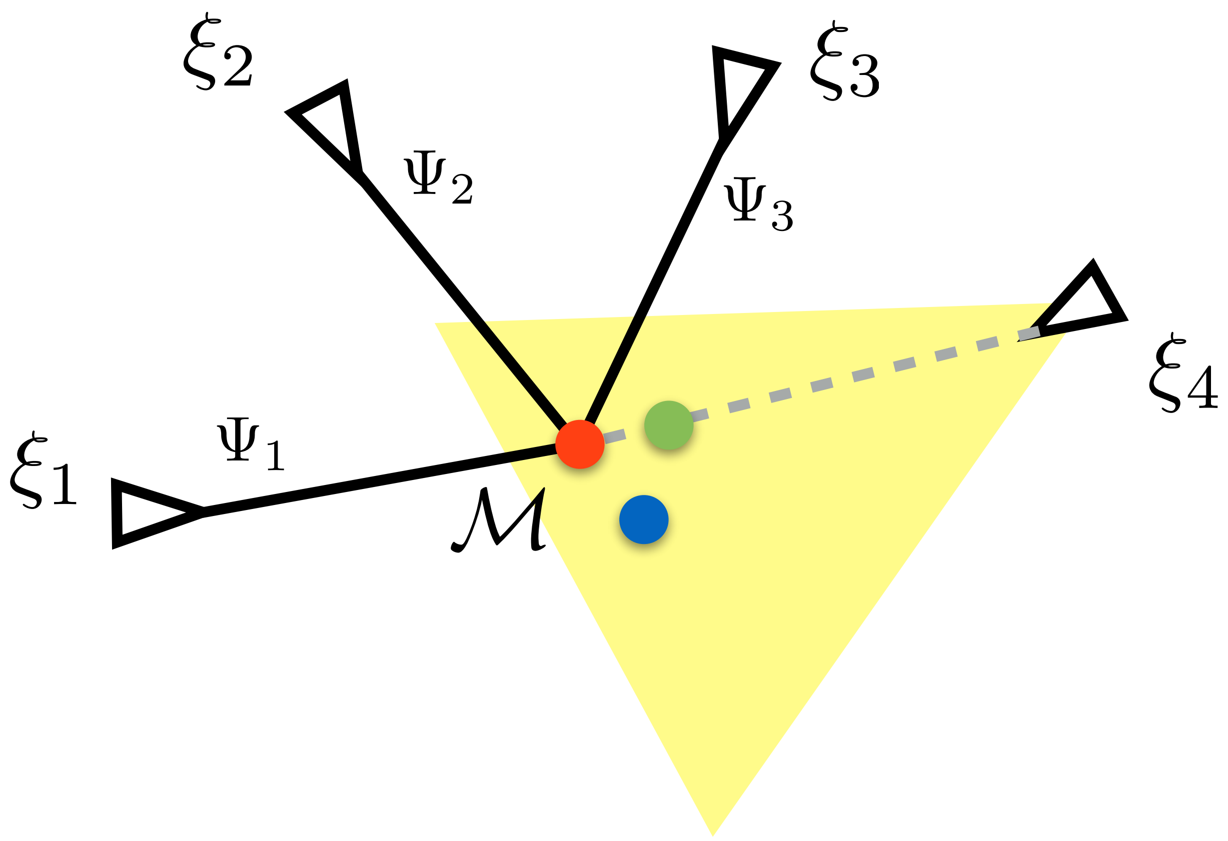
MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT

We first compute the bounding box of the object proposal $\mathcal{O}^j$ onto the keyframe ξ_k

For each object proposal
$\mathcal{O}^j \in \mathcal{O}$

RoI Pooling and Description using Fast R-CNN



MLE factorizes assuming features Ψ_k^j are conditional independent given class label $\mathbf{y}$

$$\hat{y}^j_{MLE} = \underset{y \in \{1,\ldots,C\}}{\operatorname{argmax}} \prod_{k \in \mathcal{V}^j} p(\Psi_k^j \mid Y = y)$$

$$= \underset{y \in \{1,\ldots,C\}}{\operatorname{argmax}} \sum_{k \in \mathcal{V}^j} \log p(\Psi_k^j \mid Y = y)$$

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

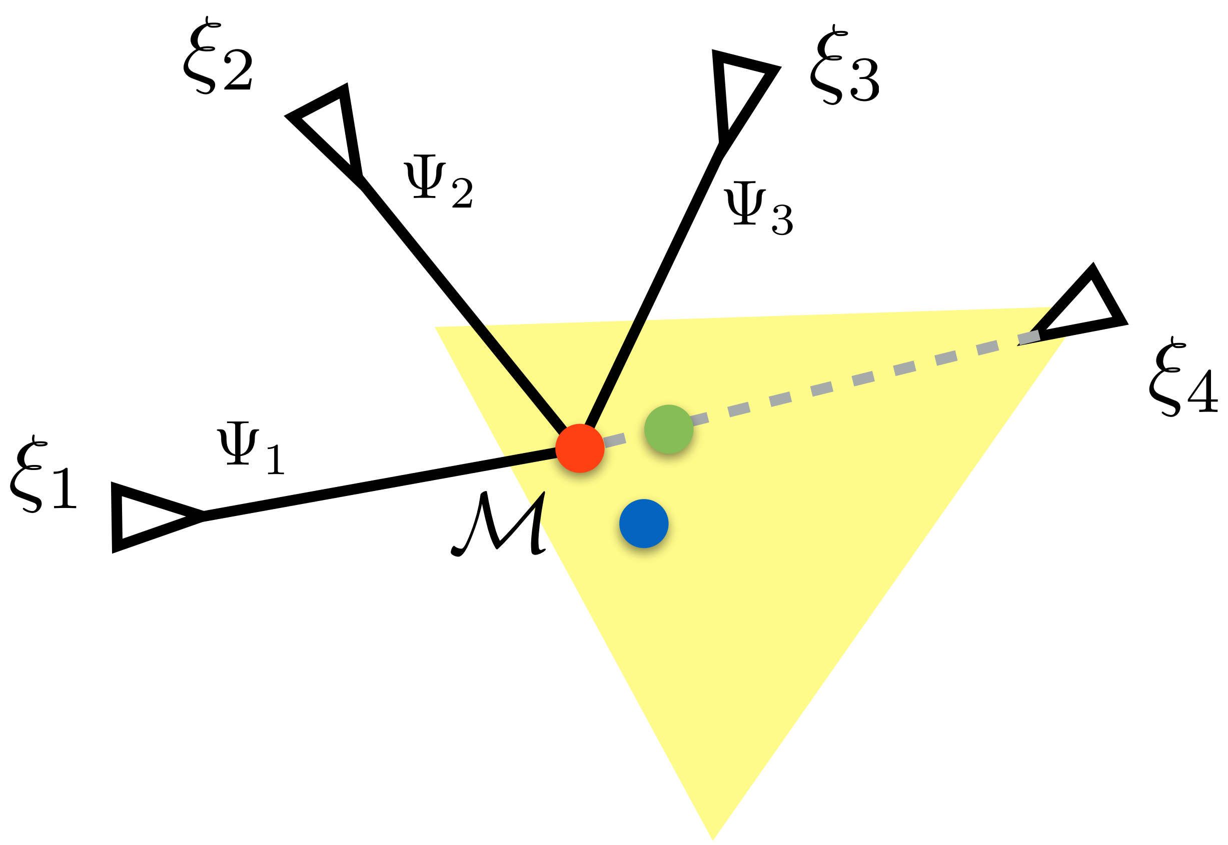
KEY CONCEPT

For each object proposal

$o^j \in \mathcal{O}$

We first compute the bounding box of
the object proposal o^j onto the keyframe ξ_k

RoI Pooling and Description using Fast R-CNN



MLE factorizes assuming features Ψ_k^j are conditional independent given class label $\mathbf{y}$

$$\hat{y}^j_{MLE} = \underset{y \in \{1,\ldots,C\}}{\operatorname{argmax}} \prod_{k \in \mathcal{V}^j} p(\Psi_k^j \mid Y = y)$$

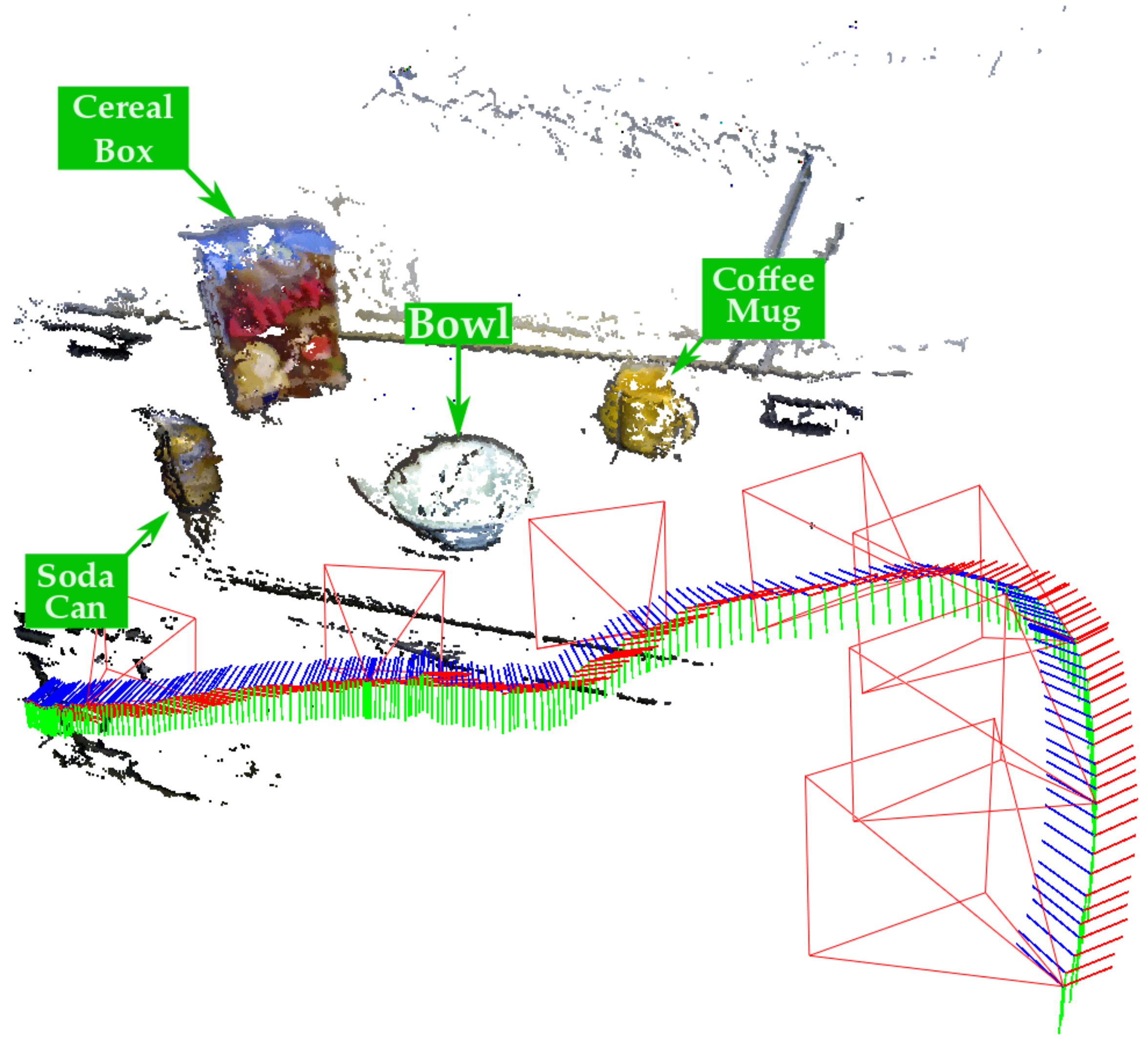
$$= \underset{y \in \{1,\ldots,C\}}{\operatorname{argmax}} \sum_{k \in \mathcal{V}^j} \log p(\Psi_k^j \mid Y = y)$$

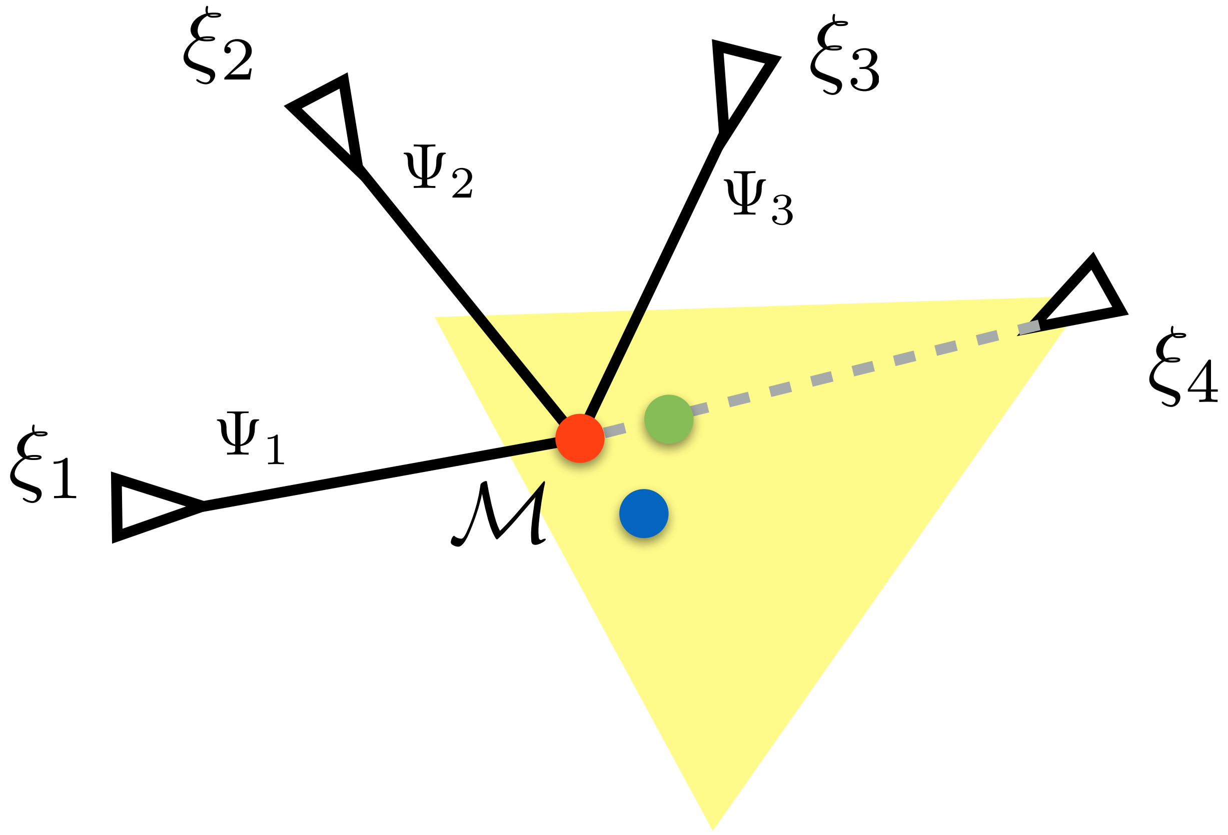
$$p(\Psi_k^j | Y = y)$$

Logistic regression on the features extracted from the object proposal $\mathbf{j}$ in view $\mathbf{k}$

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

KEY CONCEPT





MLE factorizes assuming features Ψ_k^j are conditional independent given class label $\mathbf{y}$

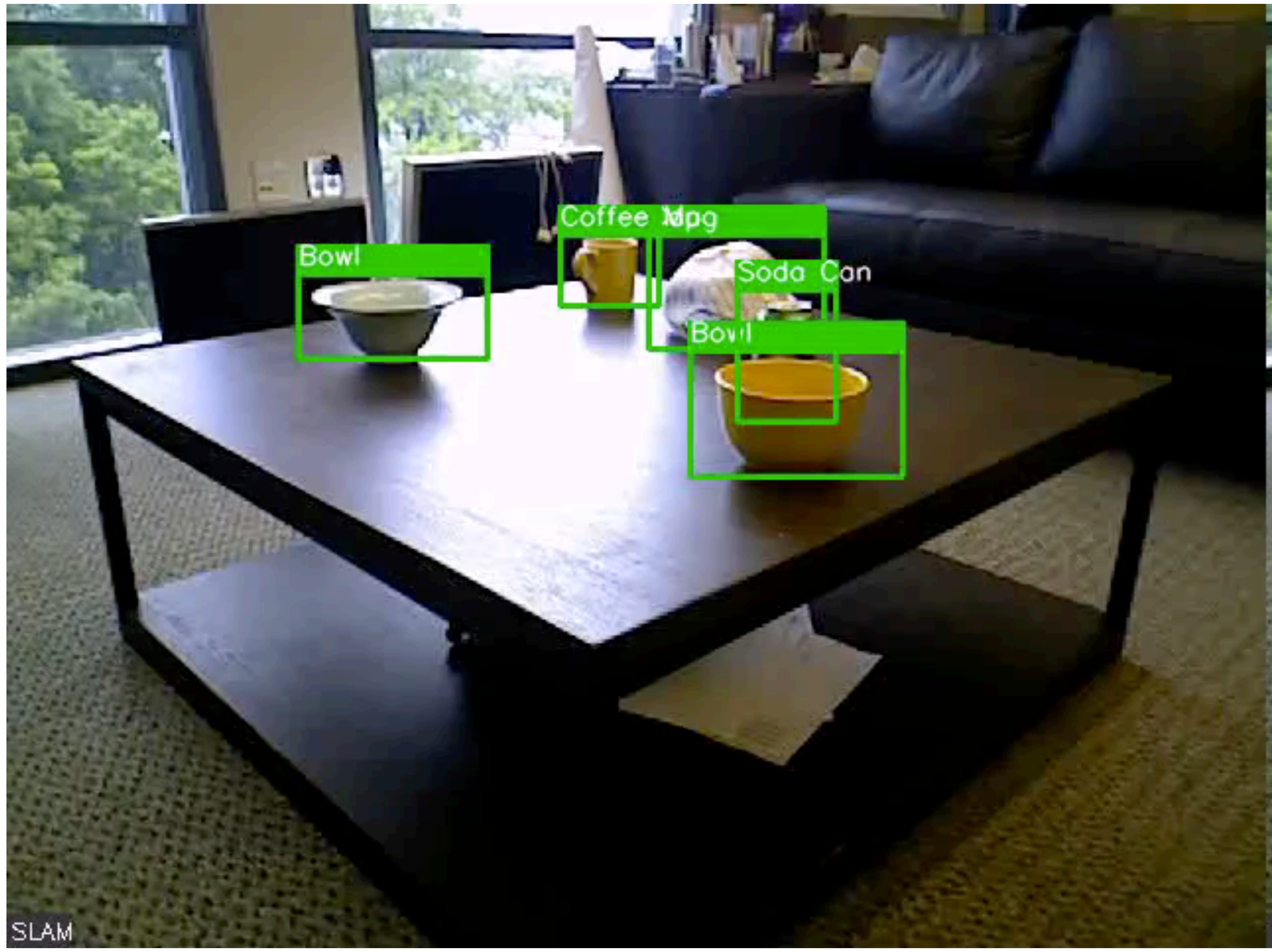
$$\hat{y}^j_{MLE} = \underset{y\in\{1,\dots,C\}}{\operatorname{argmax}} \prod_{k\in\mathcal{V}^j} p(\Psi_k^j \mid Y = y)$$

$$= \underset{y\in\{1,\dots,C\}}{\operatorname{argmax}} \sum_{k\in\mathcal{V}^j} \log p(\Psi_k^j \mid Y = y)$$

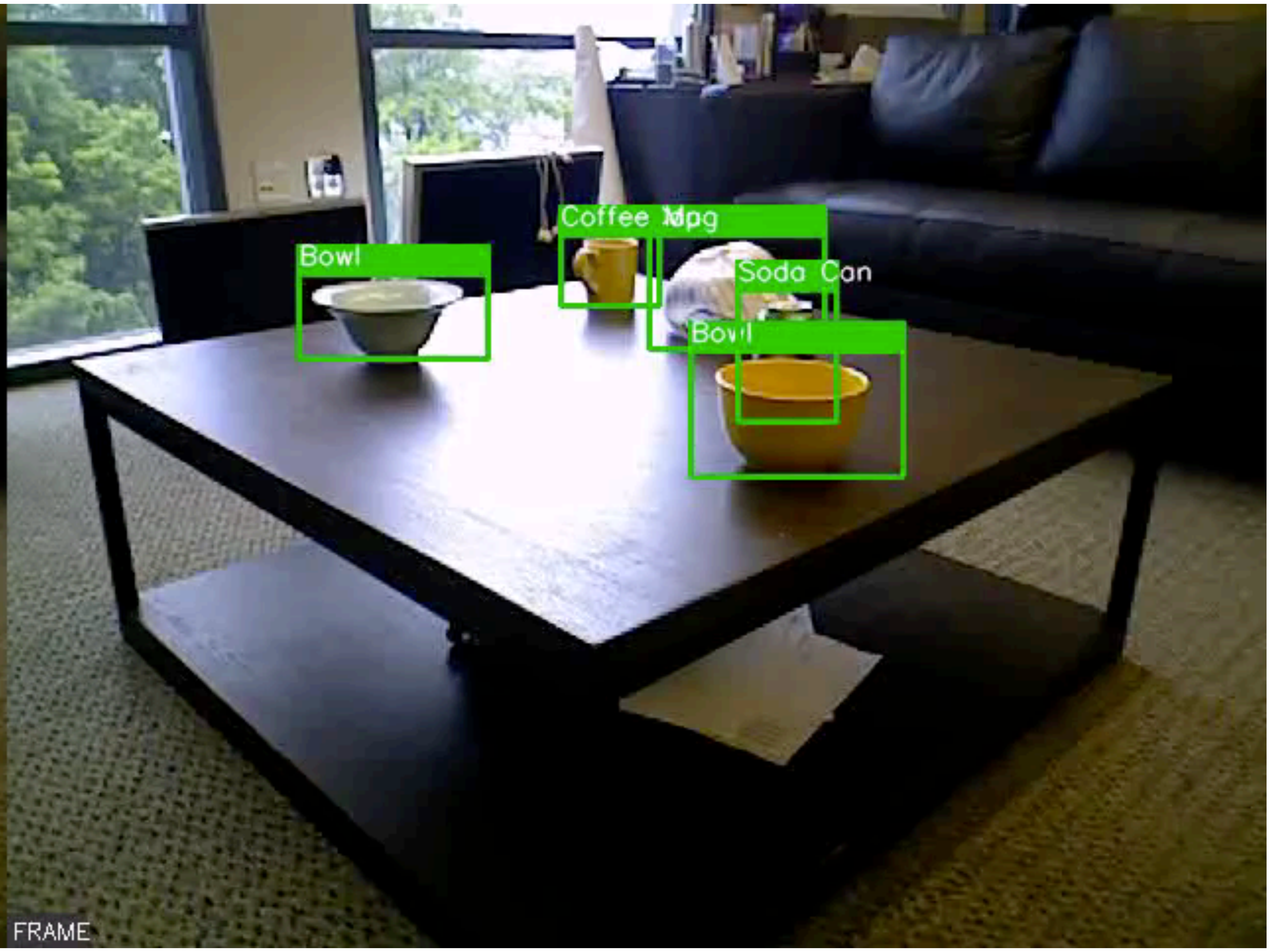
$$p(\Psi_k^j|Y = y)$$

Logistic regression on the features extracted from the object proposal $\mathbf{j}$ in view $\mathbf{k}$

SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION



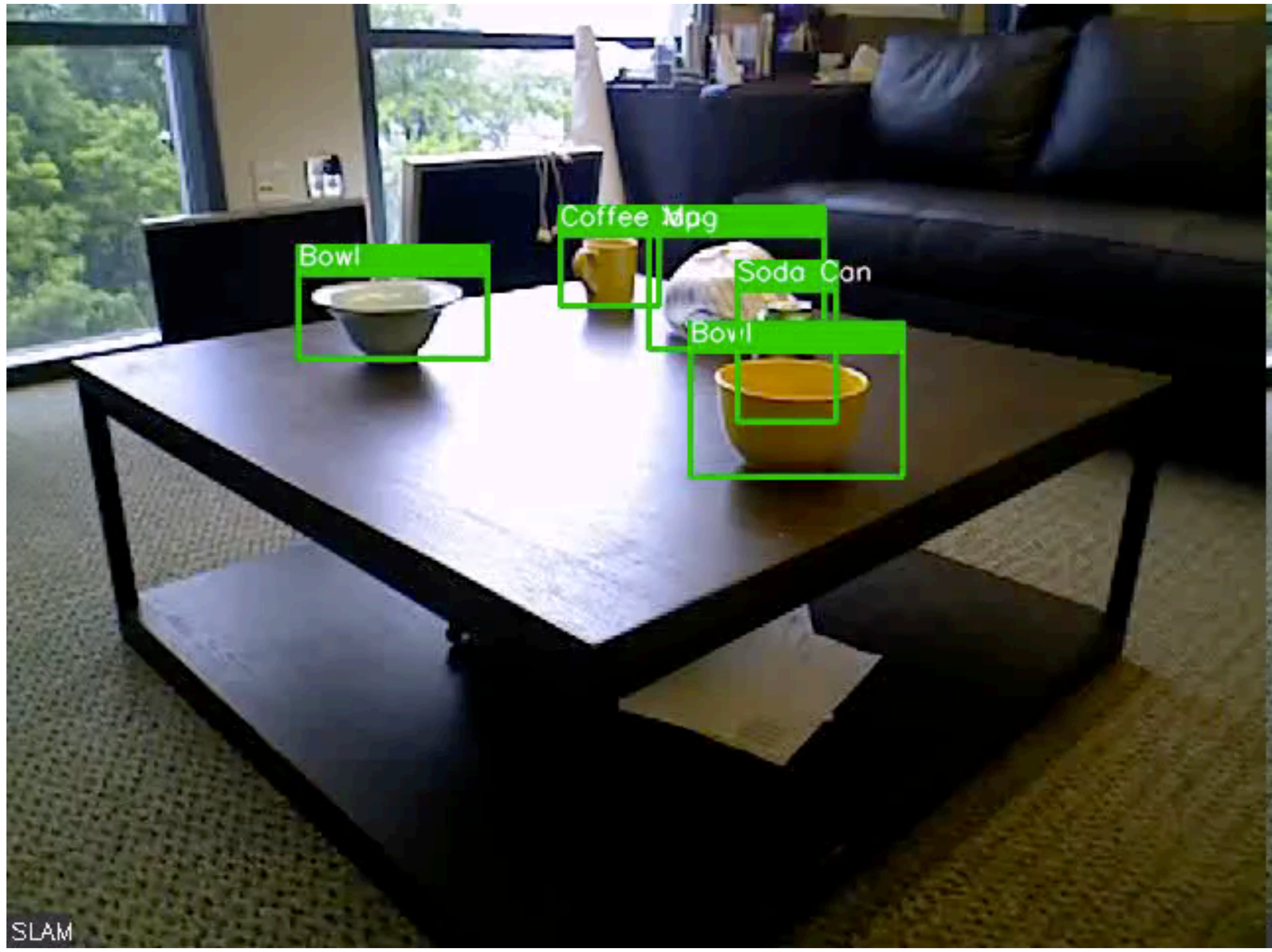
SLAM-Supported Recognition
(Ours)



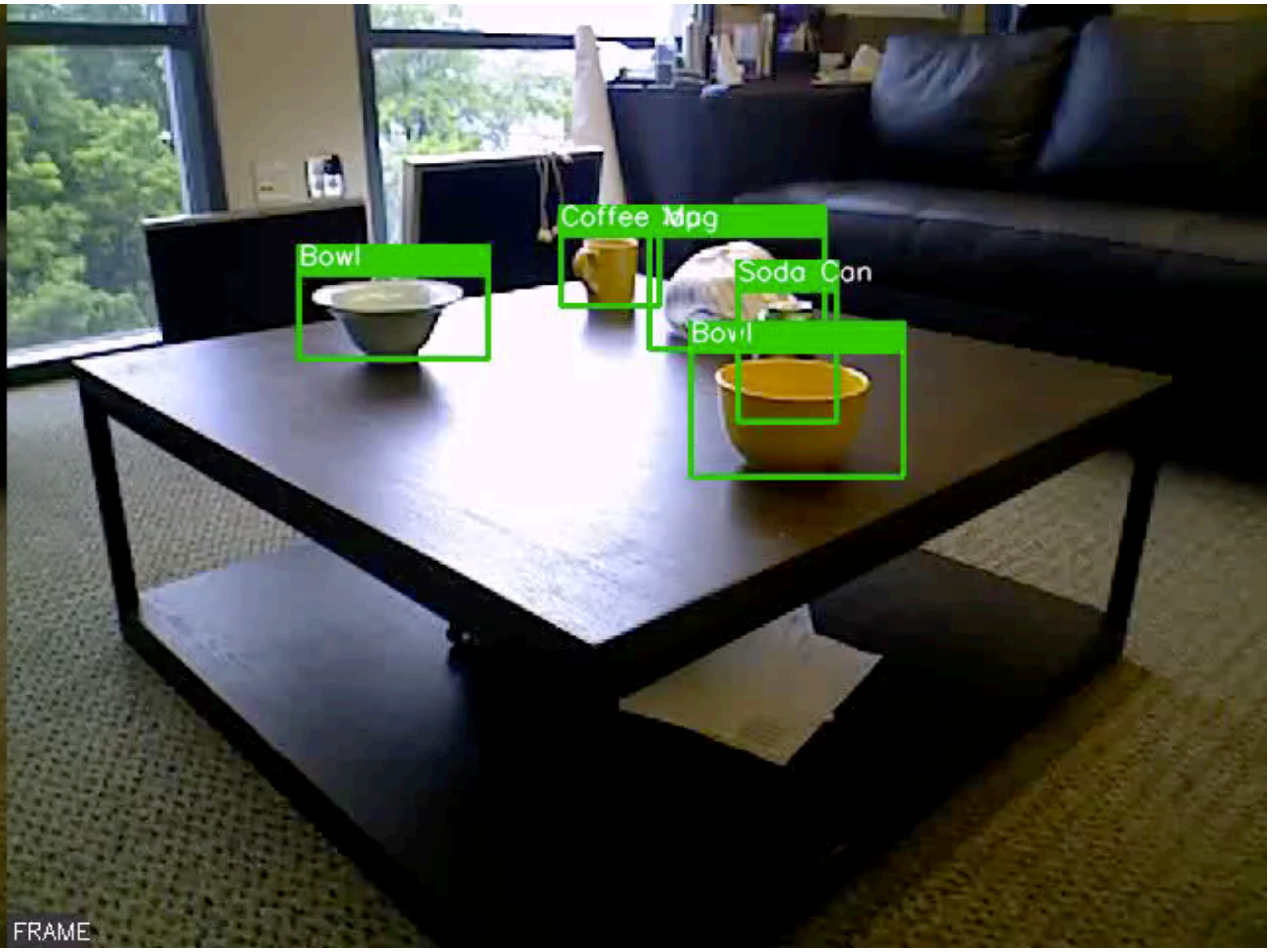
Frame-based Recognition
(Classical approach)

CORRECT PREDICTIONS INCORRECT PREDICTIONS

SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION



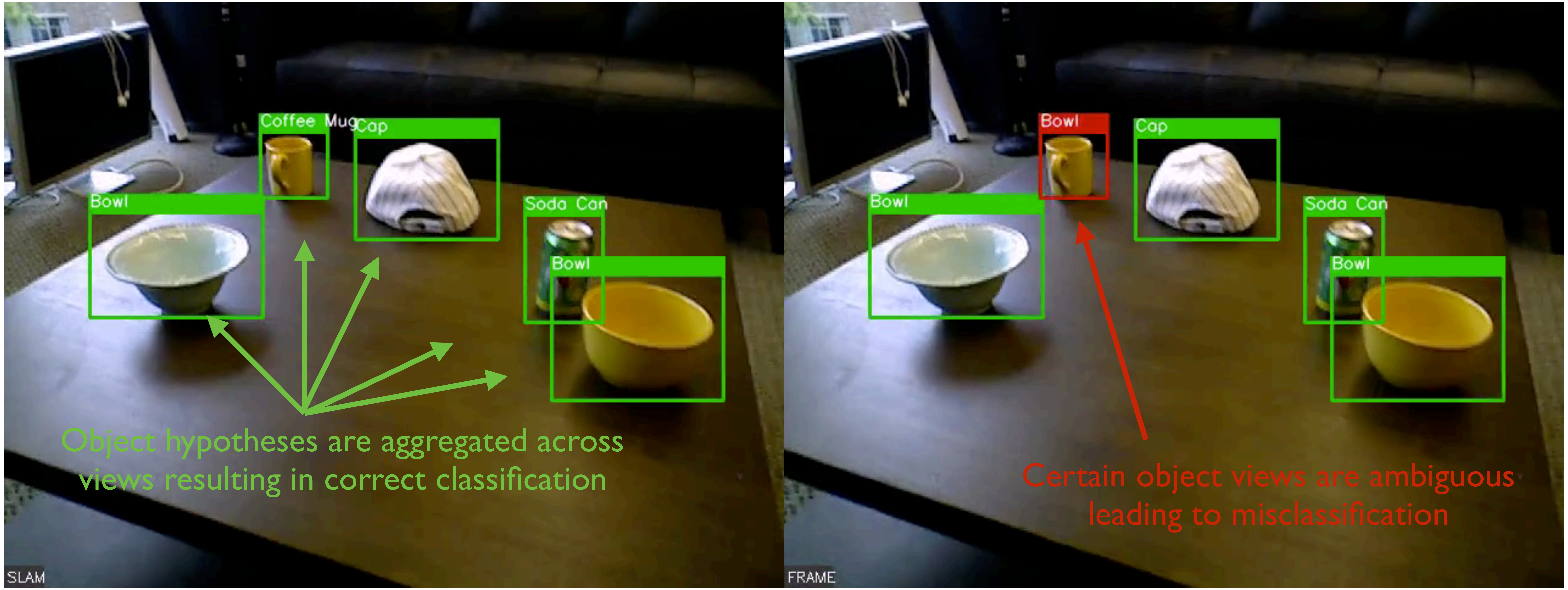
SLAM-Supported Recognition
(Ours)



Frame-based Recognition
(Classical approach)

CORRECT PREDICTIONS INCORRECT PREDICTIONS

SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION

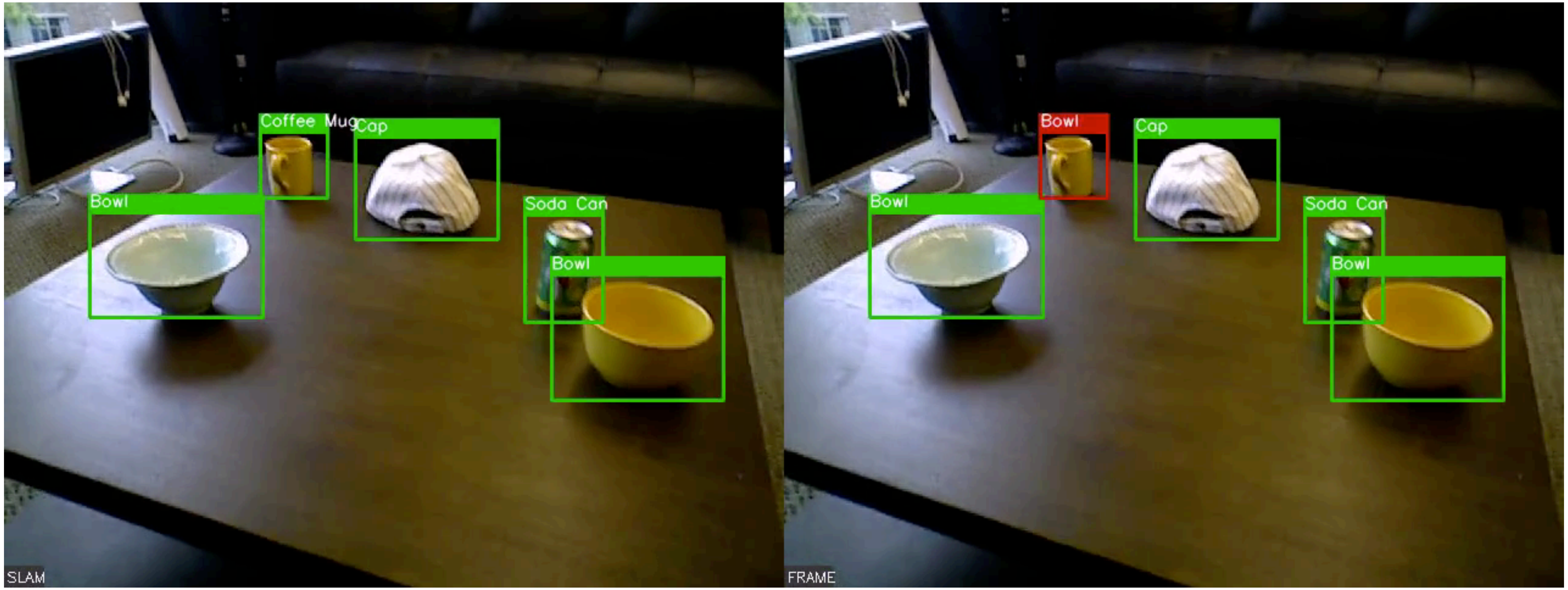


SLAM-Supported Recognition (Ours)

Frame-based Recognition (Classical approach)

CORRECT PREDICTIONS INCORRECT PREDICTIONS

SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION

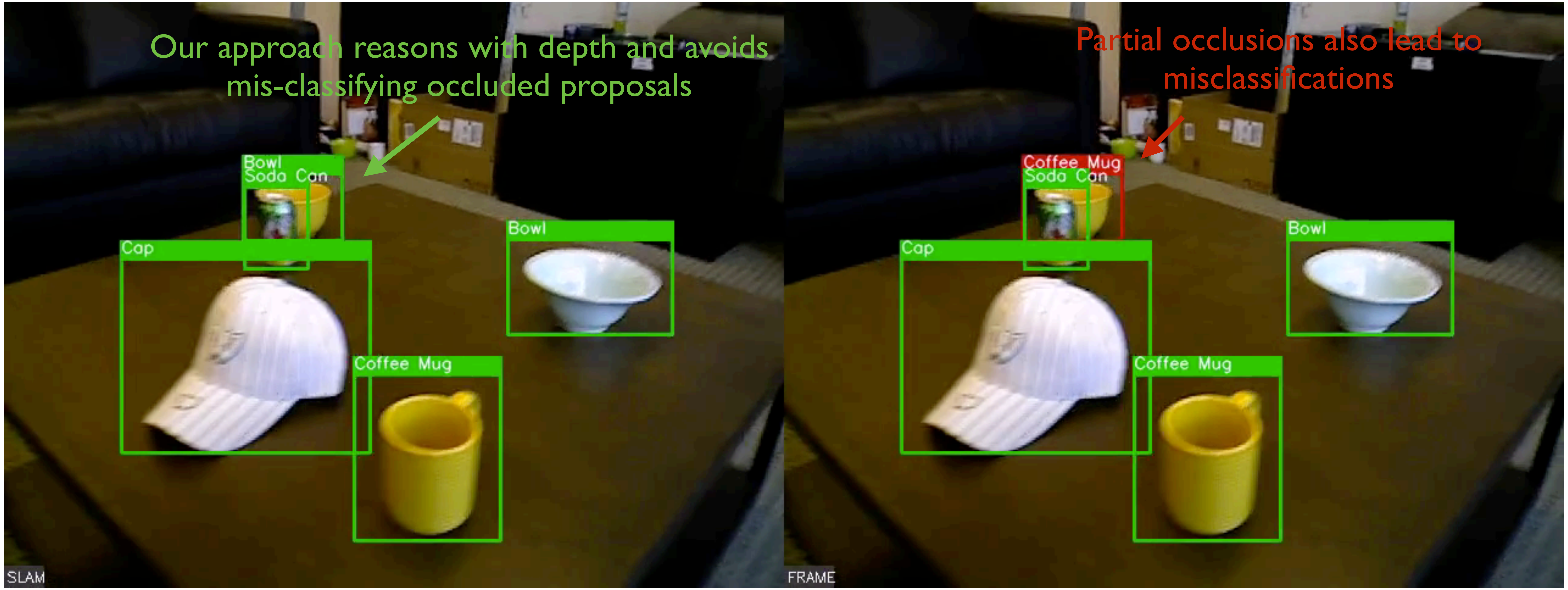


SLAM-Supported Recognition
(Ours)

Frame-based Recognition
(Classical approach)

CORRECT PREDICTIONS INCORRECT PREDICTIONS

SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION

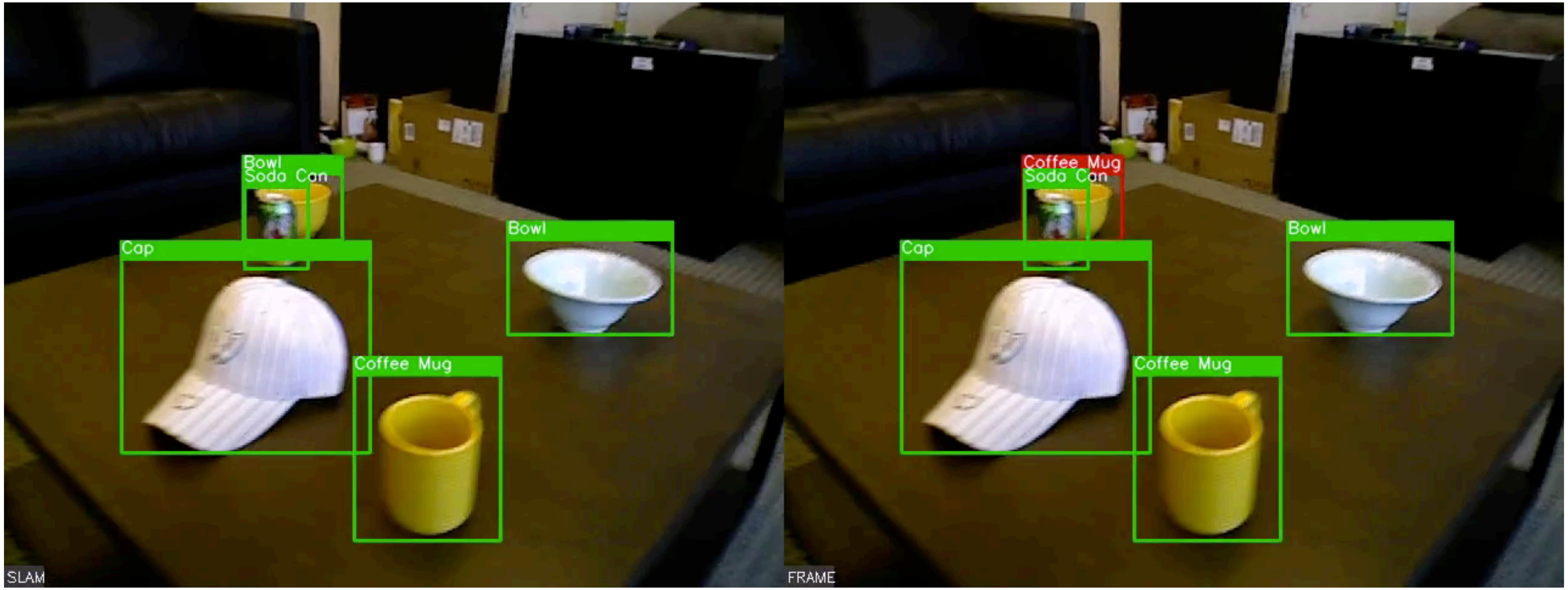


SLAM-Supported Recognition (Ours)

Frame-based Recognition (Classical approach)

CORRECT PREDICTIONS INCORRECT PREDICTIONS

SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION

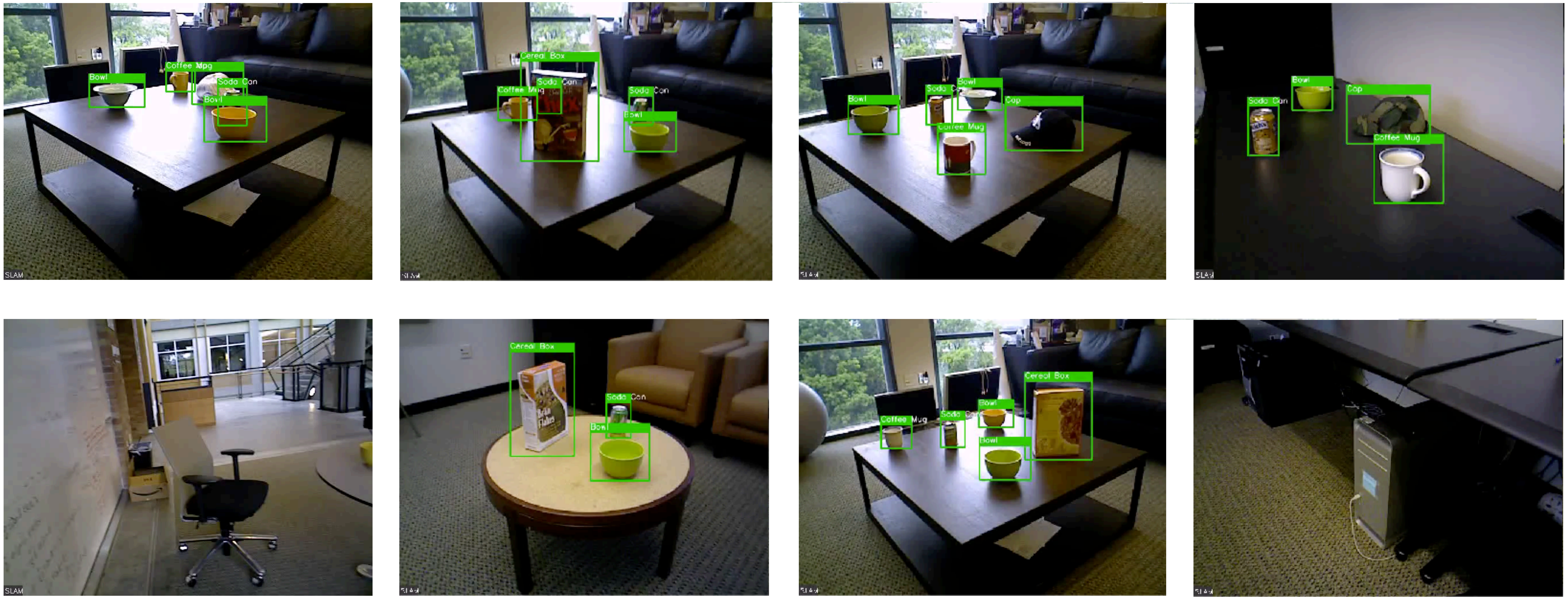


SLAM-Supported Recognition
(Ours)

Frame-based Recognition
(Classical approach)

CORRECT PREDICTIONS INCORRECT PREDICTIONS

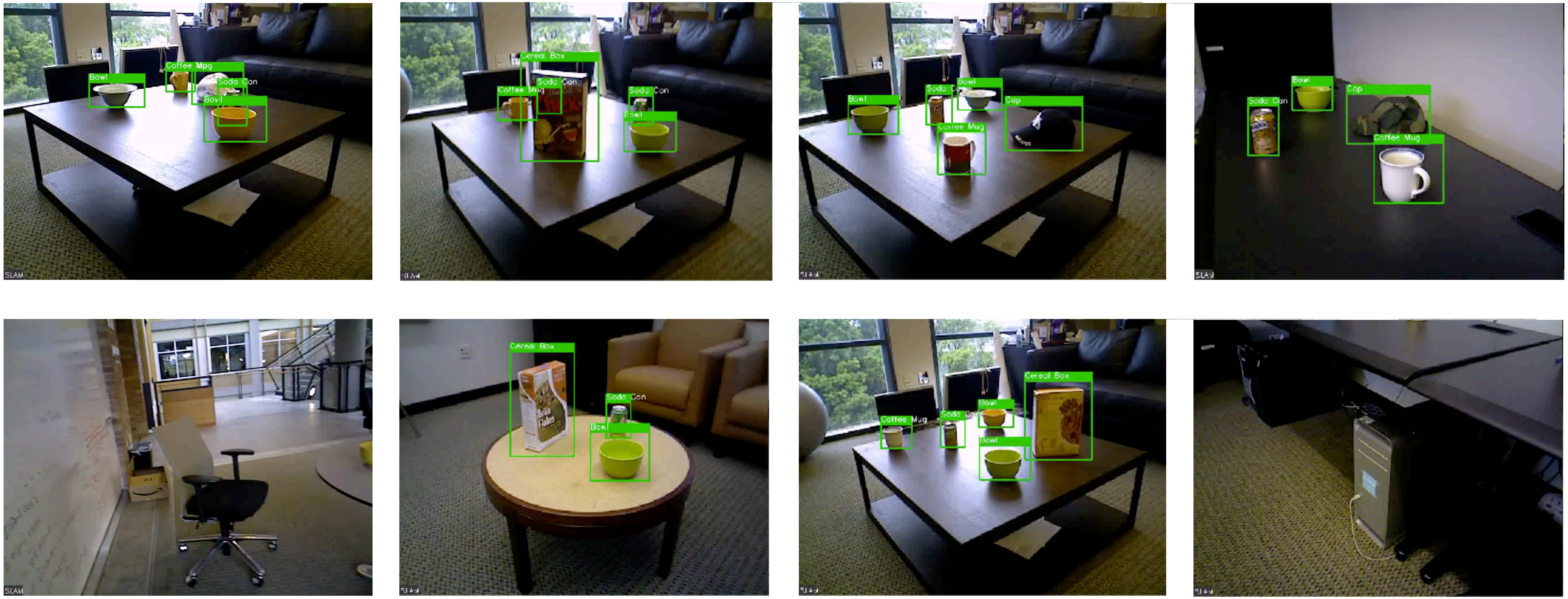
MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION



SLAM-Supported Recognition with Fast-RCNN

Occluded objects are also shown since they are visible from other viewpoints

MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

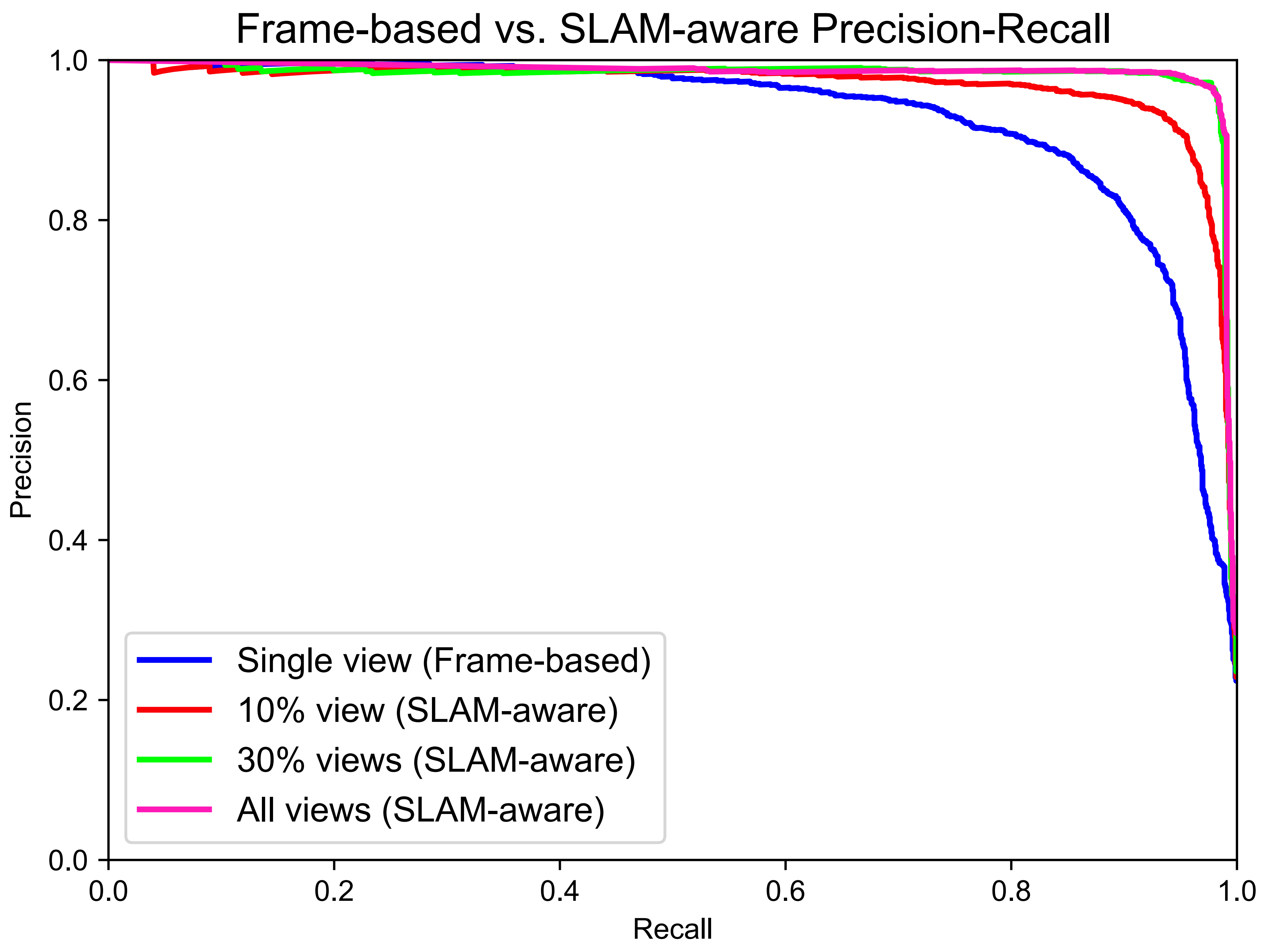


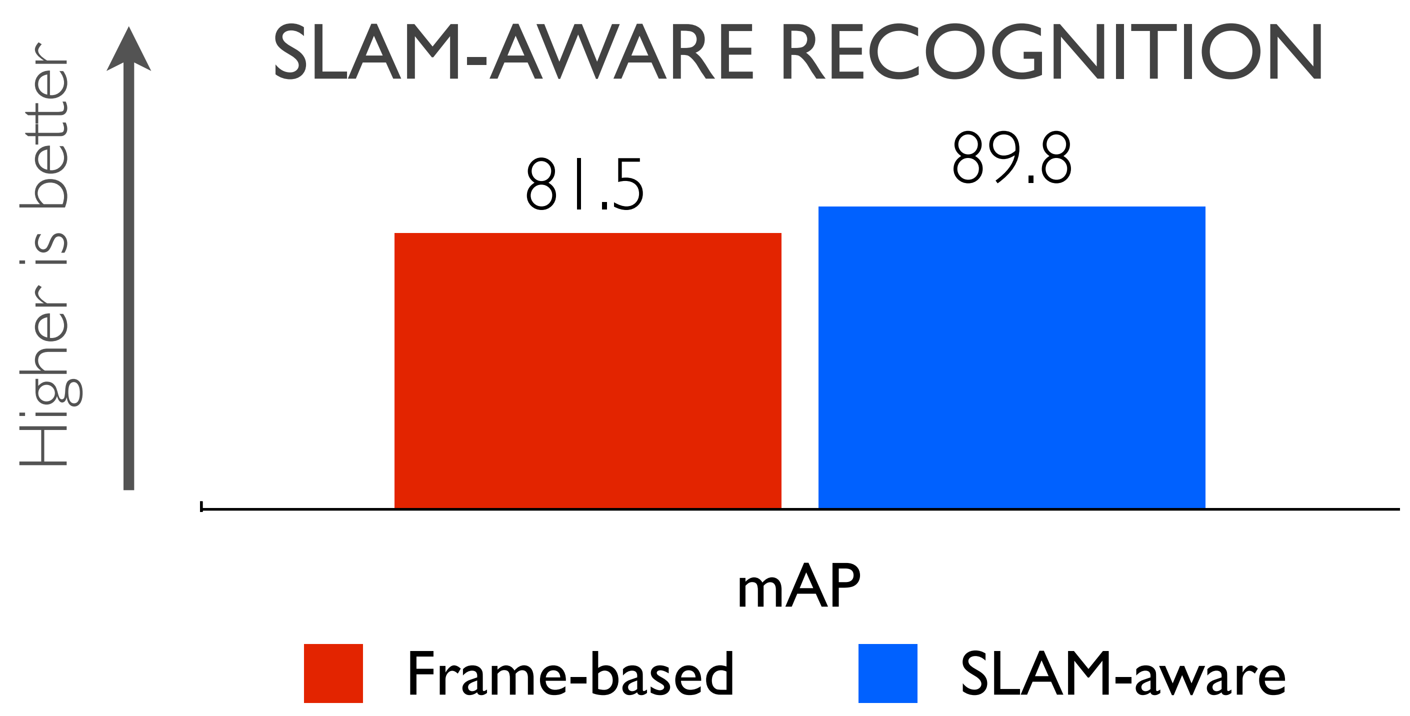
SLAM-Supported Recognition with Fast-RCNN

Occluded objects are also shown since they are visible from other viewpoints

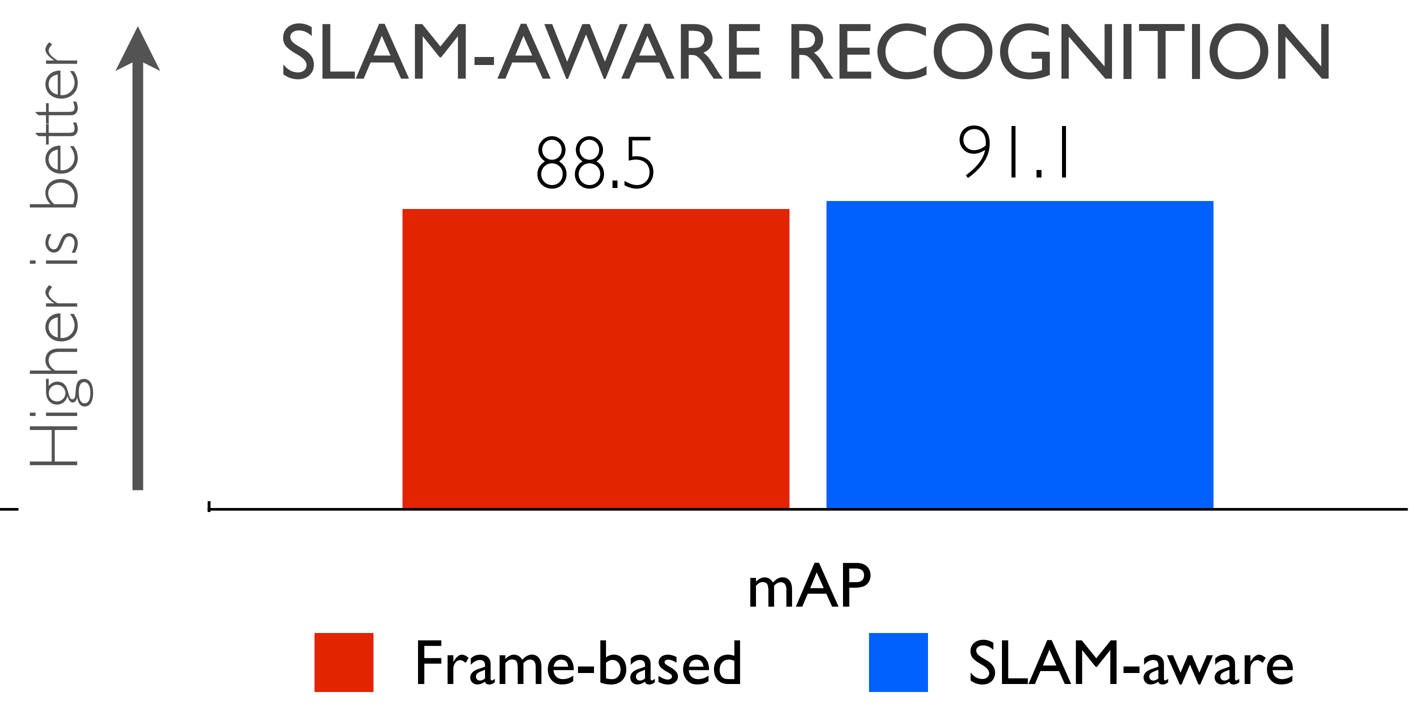
MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION PERFORMANCE

Comparing Frame-based Recognition with SLAM-aware Recognition
on UW RGB-D Scene Dataset (v2)





(VLAD-FLAIR encoding)



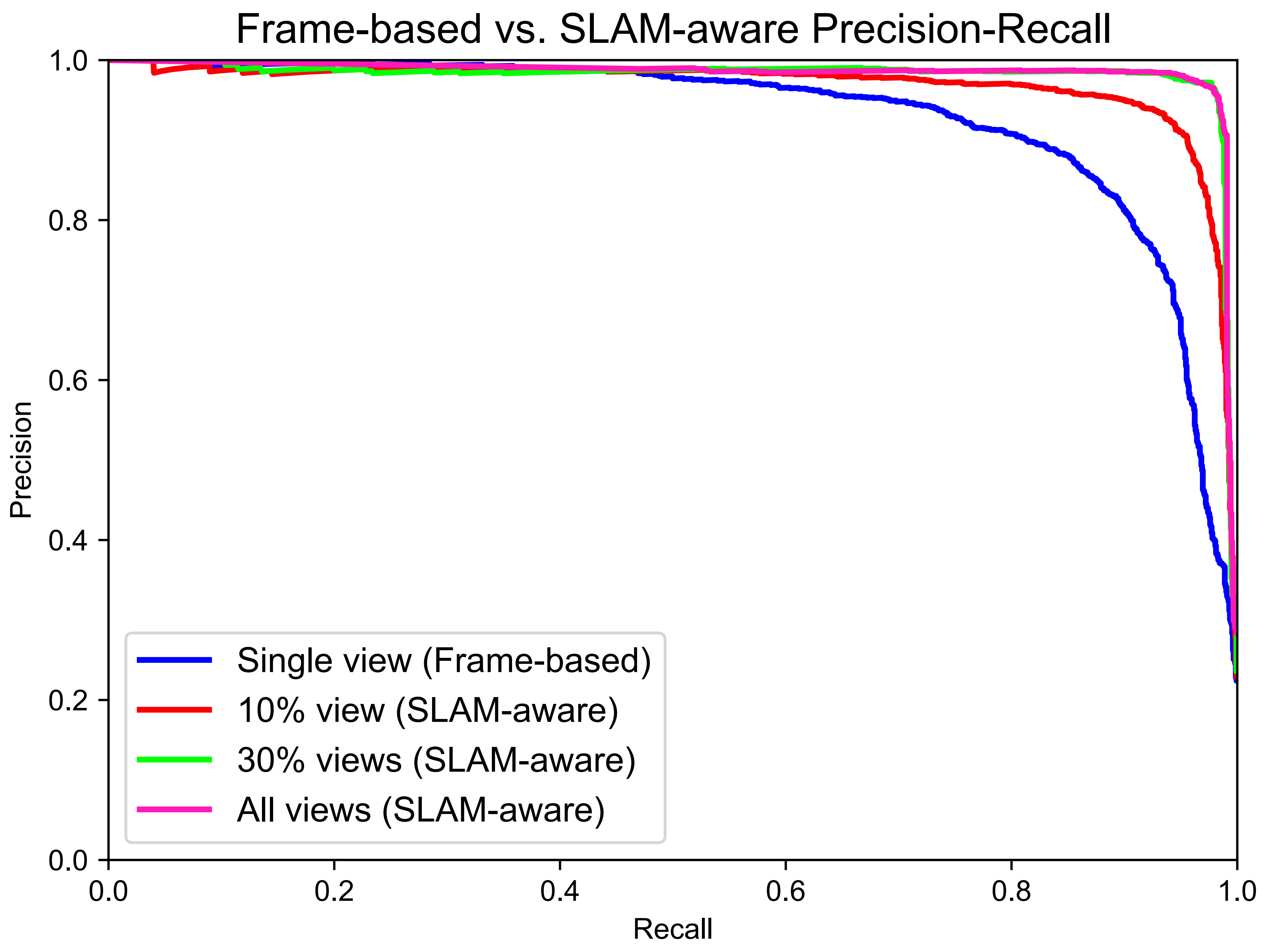
(Fast RCNN encoding)

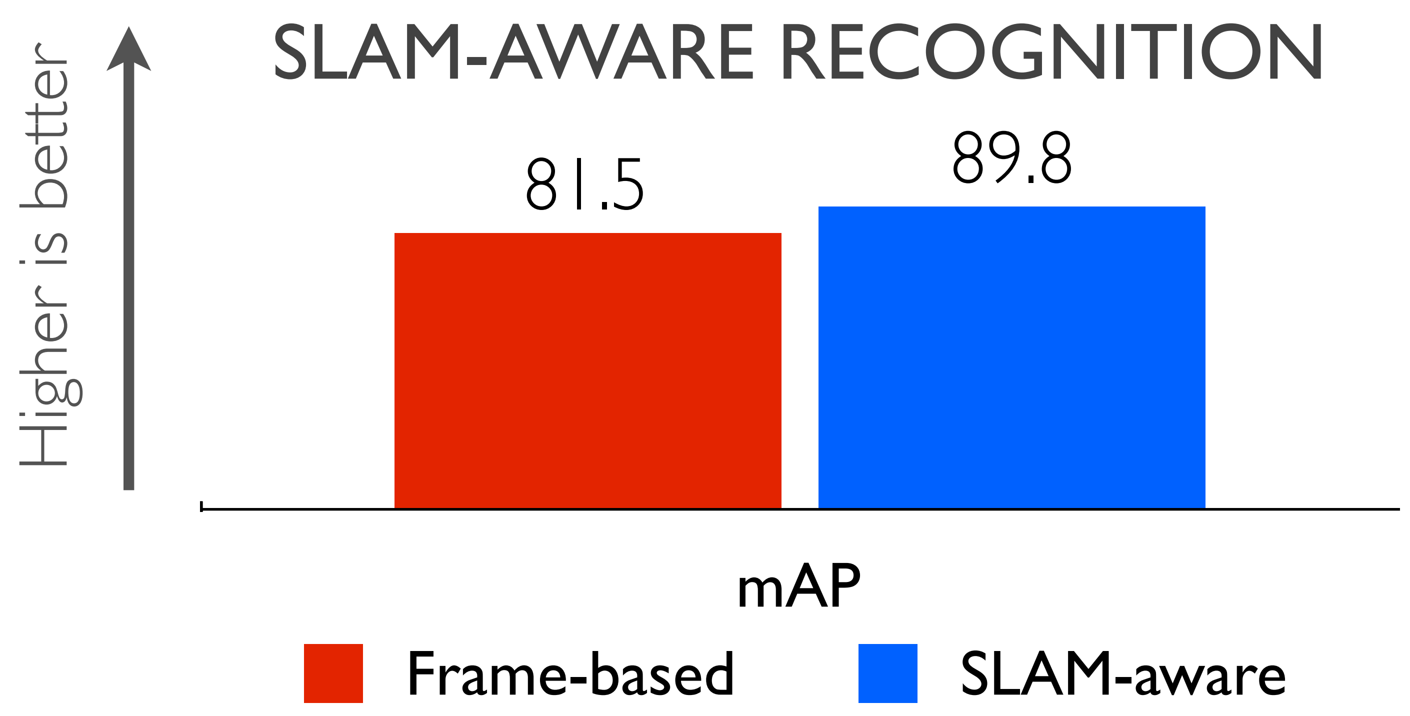
mAP - (mean Average Precision)

Only RGB channels are considered

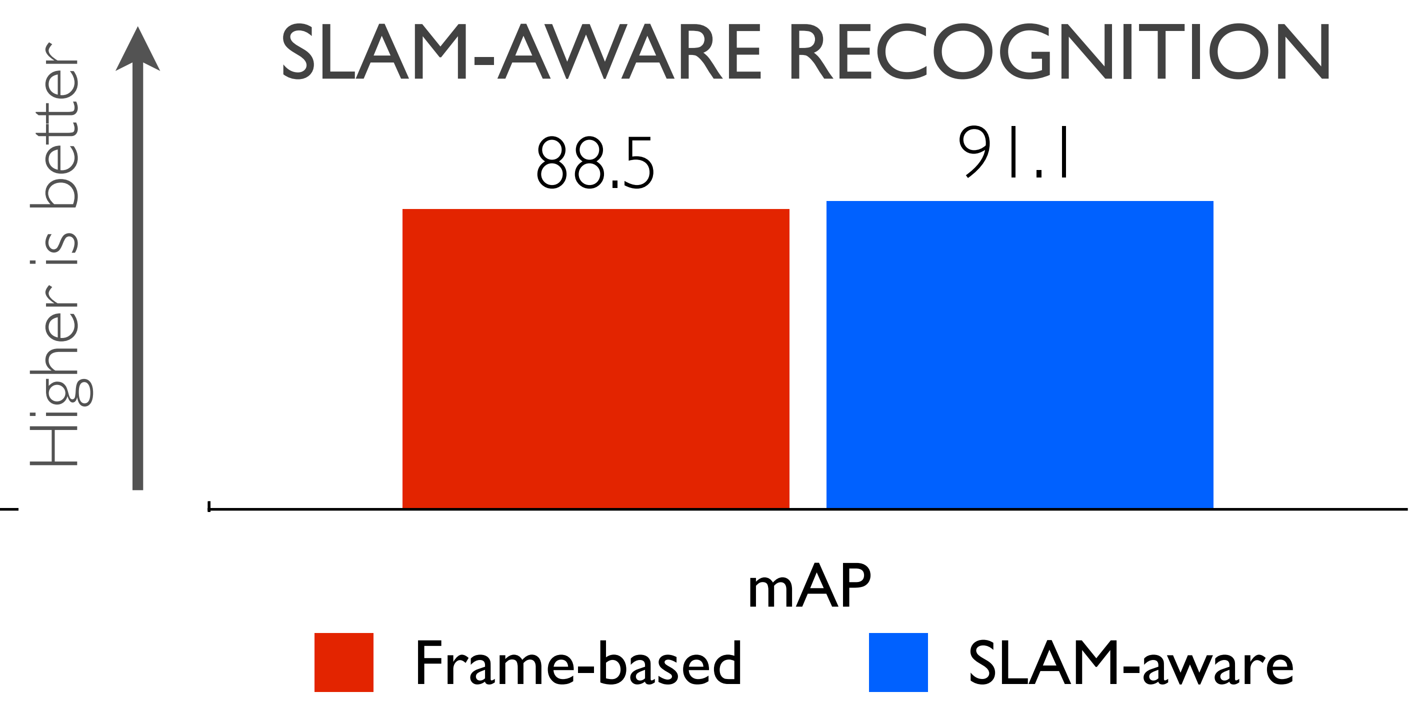
MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION PERFORMANCE

Comparing Frame-based Recognition with SLAM-aware Recognition
on UW RGB-D Scene Dataset (v2)





(VLAD-FLAIR encoding)



(Fast RCNN encoding)

mAP - (**m**ean **A**verage **P**recision)

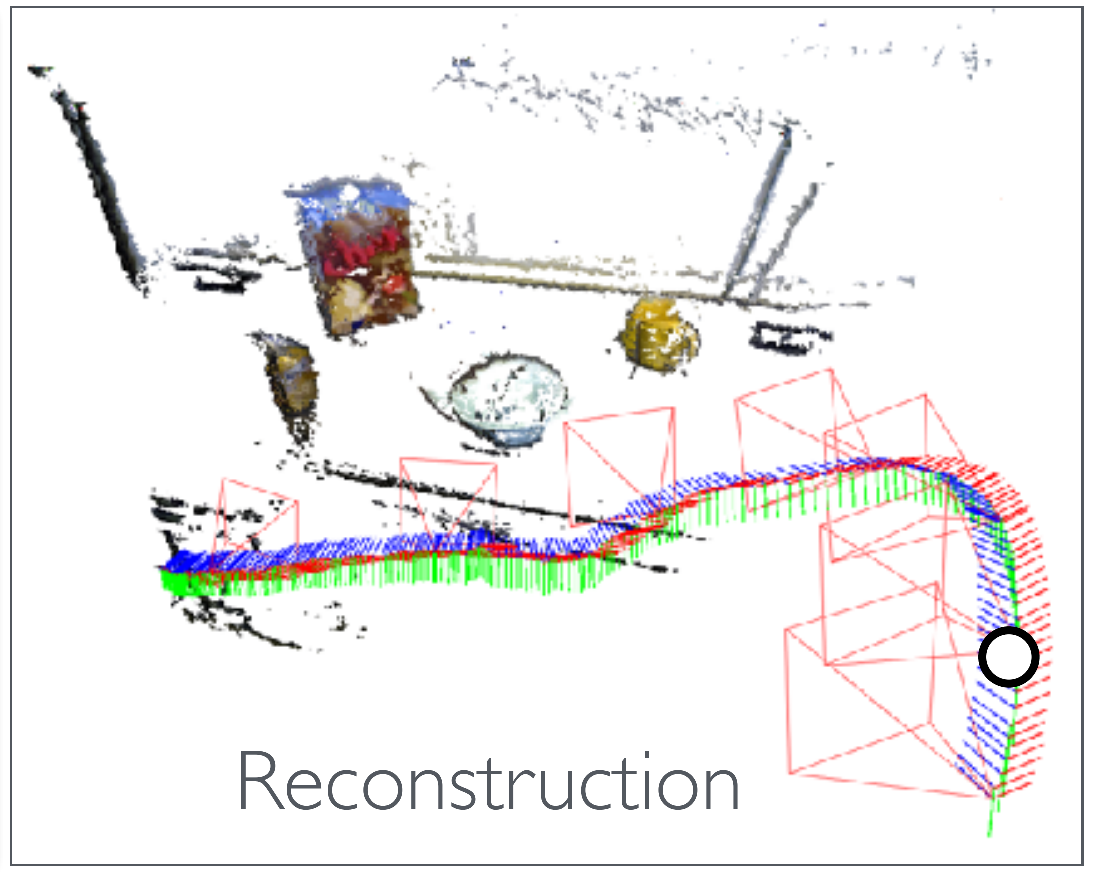
Only RGB channels are considered

Key Observation

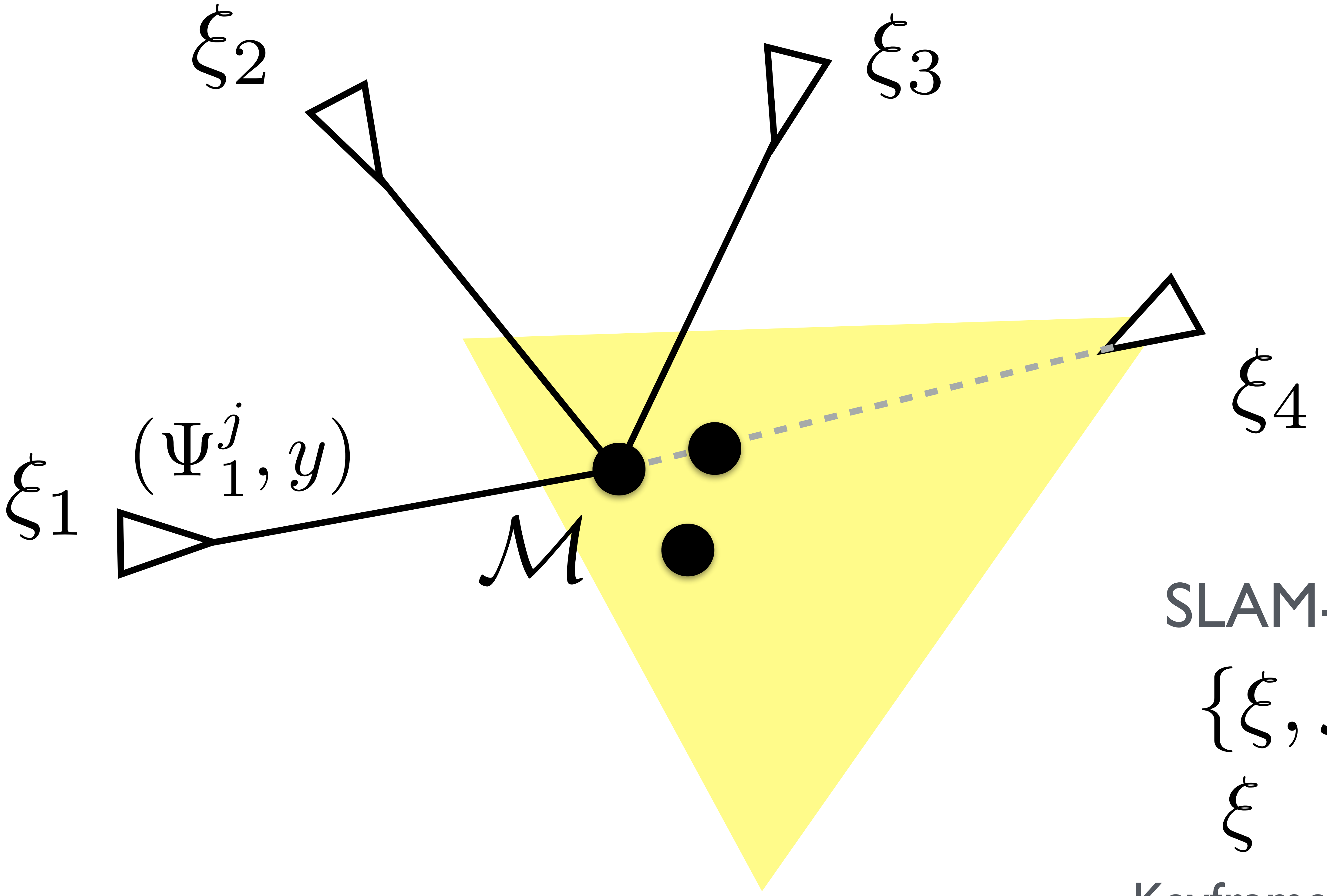
SLAM provides useful information for handling ambiguities in object labels, occlusion, and visibility understanding

SLAM-AWARE FEW-SHOT OBJECT LEARNING

- SLAM-aware few-shot object learning
  - Spatially-consistent proposals with occlusion-handling
  - Label drift mitigation via geometric consistency



SLAM as a correspondence-engine



SLAM-aware

$\{\xi, \mathcal{M}\}$

ξ $\mathcal{M}$

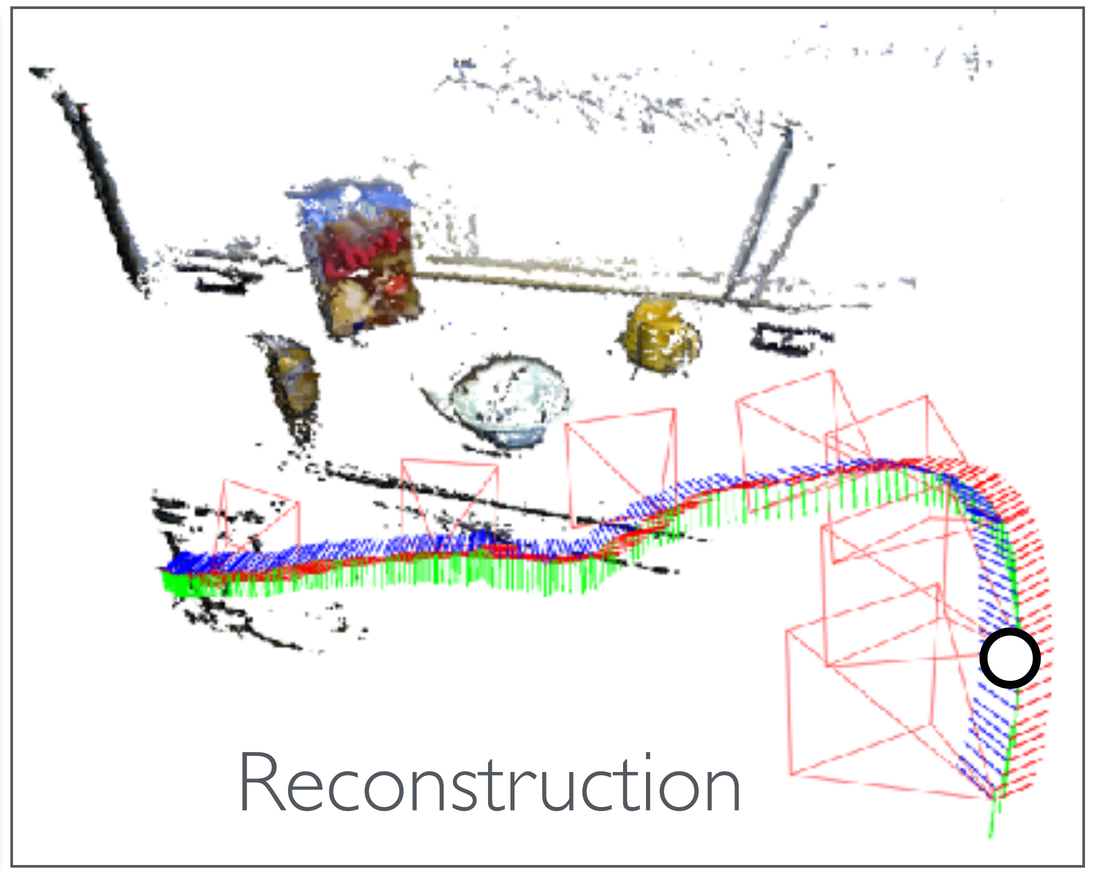
Keyframes Map

SLAM-aware label propagation

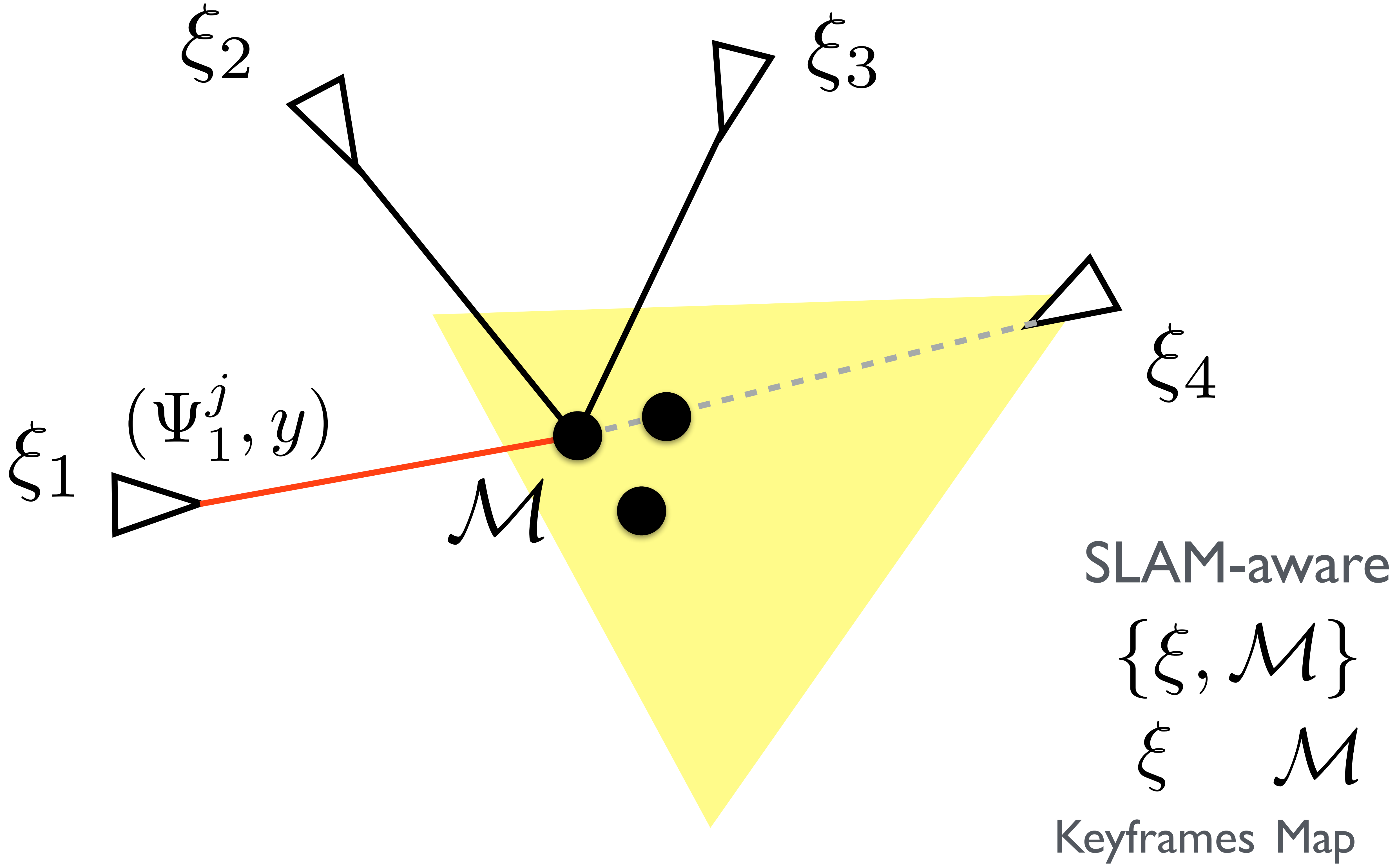
Occluded views are not propagated onto, avoiding any mis-labeling

SLAM-AWARE FEW-SHOT OBJECT LEARNING

- SLAM-aware few-shot object learning
  - Spatially-consistent proposals with occlusion-handling
  - Label drift mitigation via geometric consistency



SLAM as a correspondence-engine

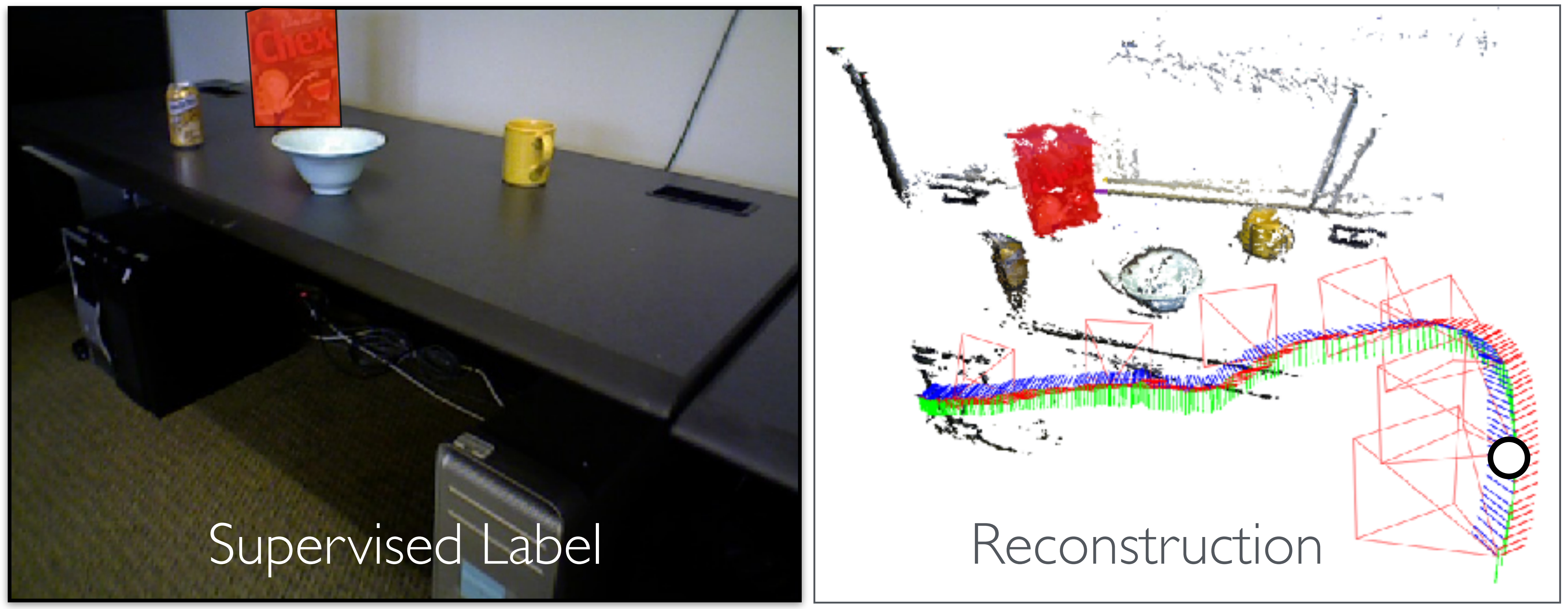


SLAM-aware label propagation

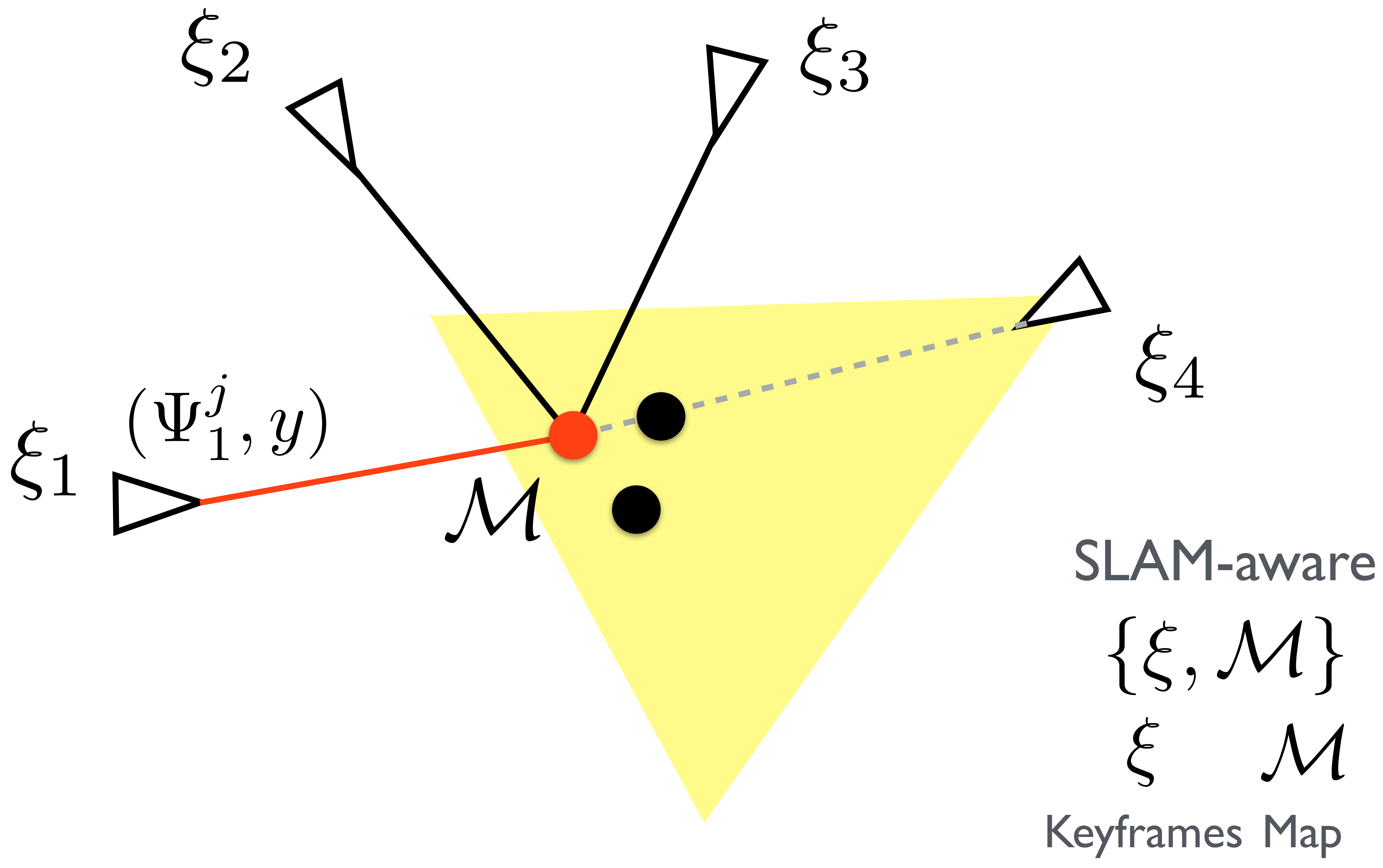
Occluded views are not propagated onto, avoiding any mis-labeling

SLAM-AWARE FEW-SHOT OBJECT LEARNING

- SLAM-aware few-shot object learning
  - Spatially-consistent proposals with occlusion-handling
  - Label drift mitigation via geometric consistency



SLAM as a correspondence-engine



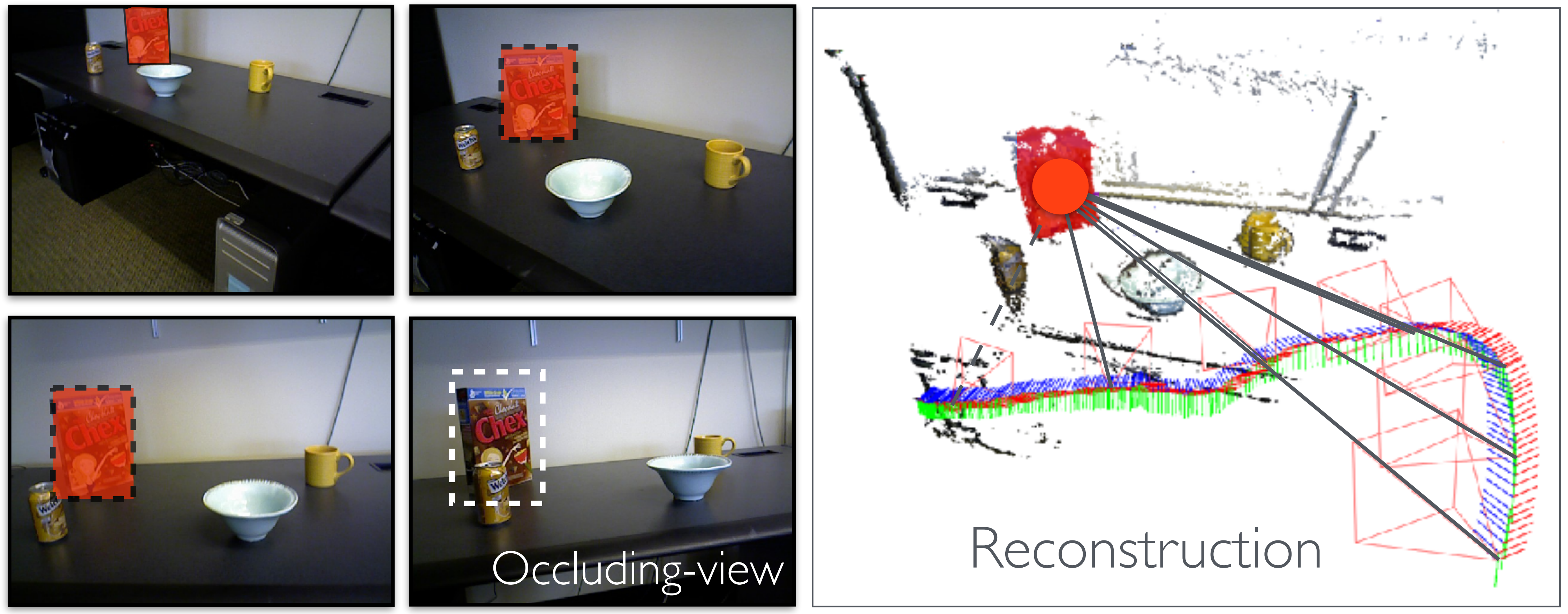
SLAM-aware label propagation

Occluded views are not propagated onto, avoiding any mis-labeling

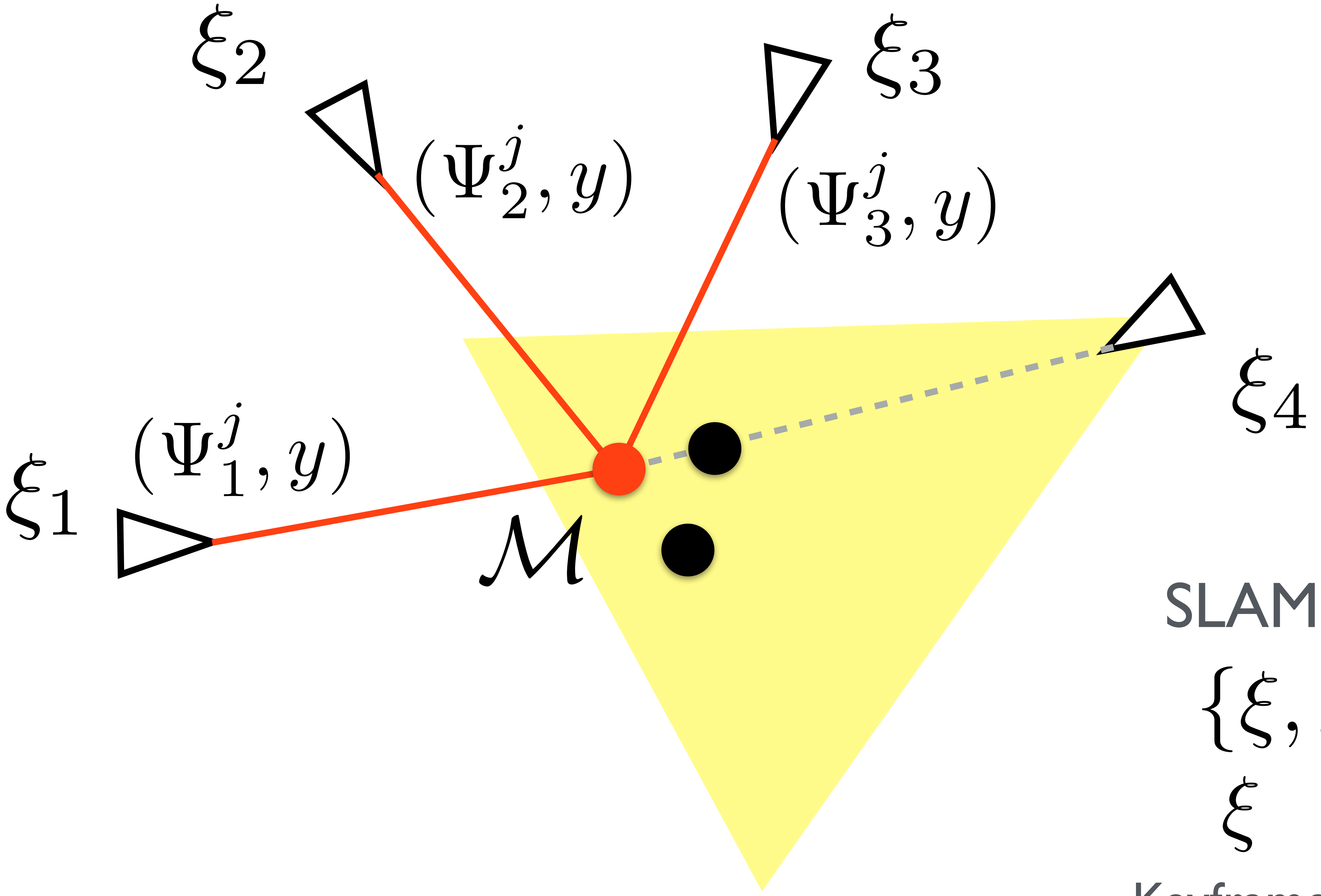
SLAM-AWARE FEW-SHOT OBJECT LEARNING

- SLAM-aware few-shot object learning
  - Spatially-consistent proposals with occlusion-handling
  - Label drift mitigation via geometric consistency

Propagated labels
& bounding boxes



SLAM as a correspondence-engine



SLAM-aware

$\{\xi, \mathcal{M}\}$

ξ $\mathcal{M}$

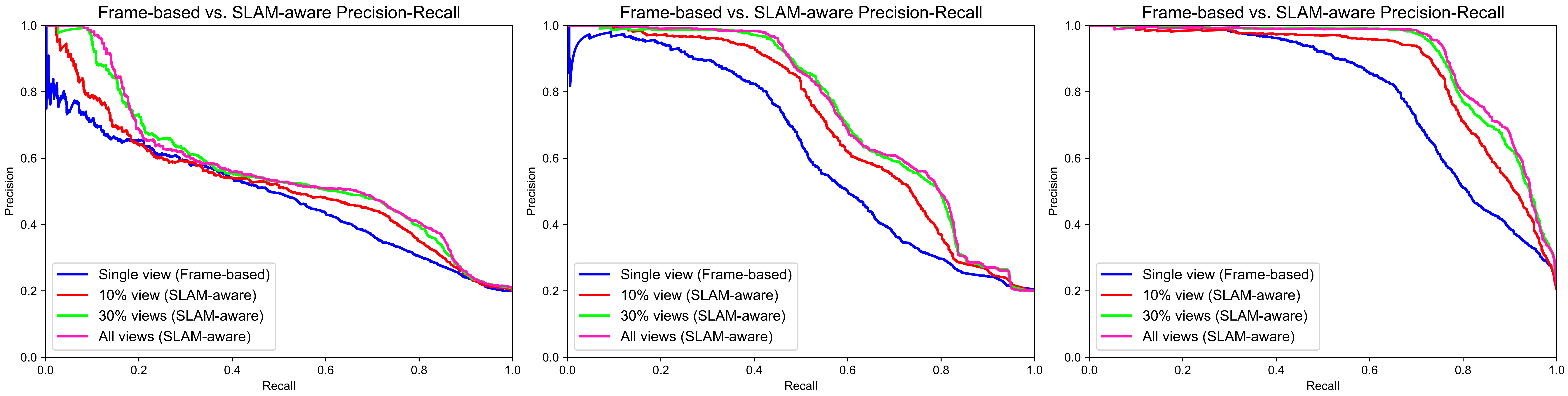
Keyframes Map

SLAM-aware label propagation

Occluded views are not propagated onto, avoiding any mis-labeling

SLAM-AWARE FEW-SHOT OBJECT LEARNING PERFORMANCE

- Randomized few-shot object learning
  - Randomly selected training information
  - Poorly trained classifiers can benefit from SLAM-aware recognition

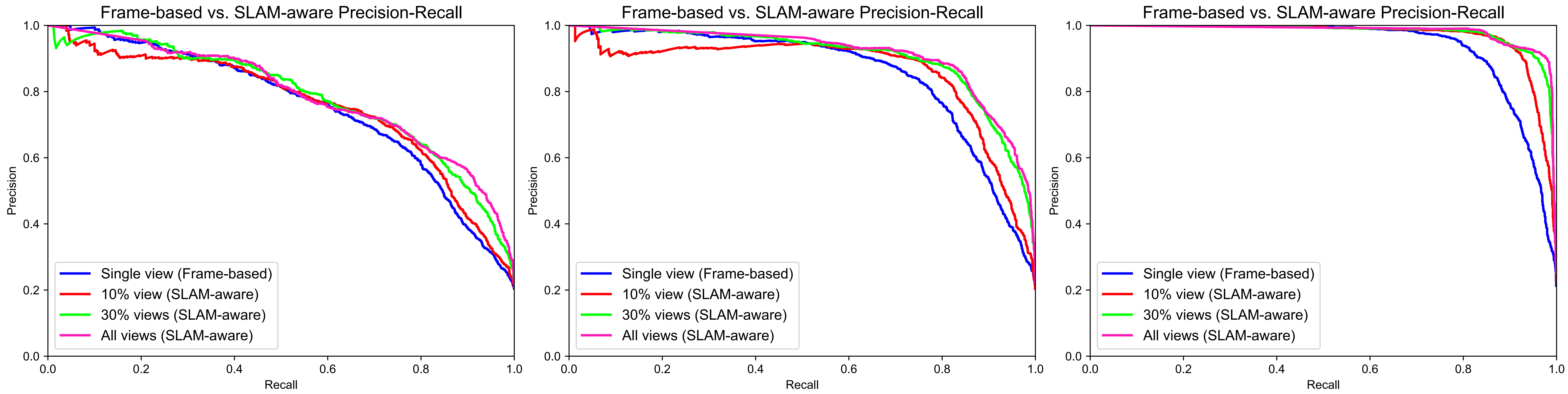


(a) 2-shot **(b) 5-shot** **(c) 10-shot**

Randomized Few-Shot Learning with SLAM-aware Recognition

SLAM-AWARE FEW-SHOT OBJECT LEARNING PERFORMANCE

- SLAM-aware few-shot object learning
  - Randomly selected training information with SLAM-aware label propagation
  - Despite minimal labels, trained classifiers are significantly more powerful



(a) 1-shot **(b) 2-shot** **(b) 4-shot**

SLAM-aware Few-Shot Learning with SLAM-aware Recognition

SLAM-AWARE FEW-SHOT OBJECT LEARNING PERFORMANCE

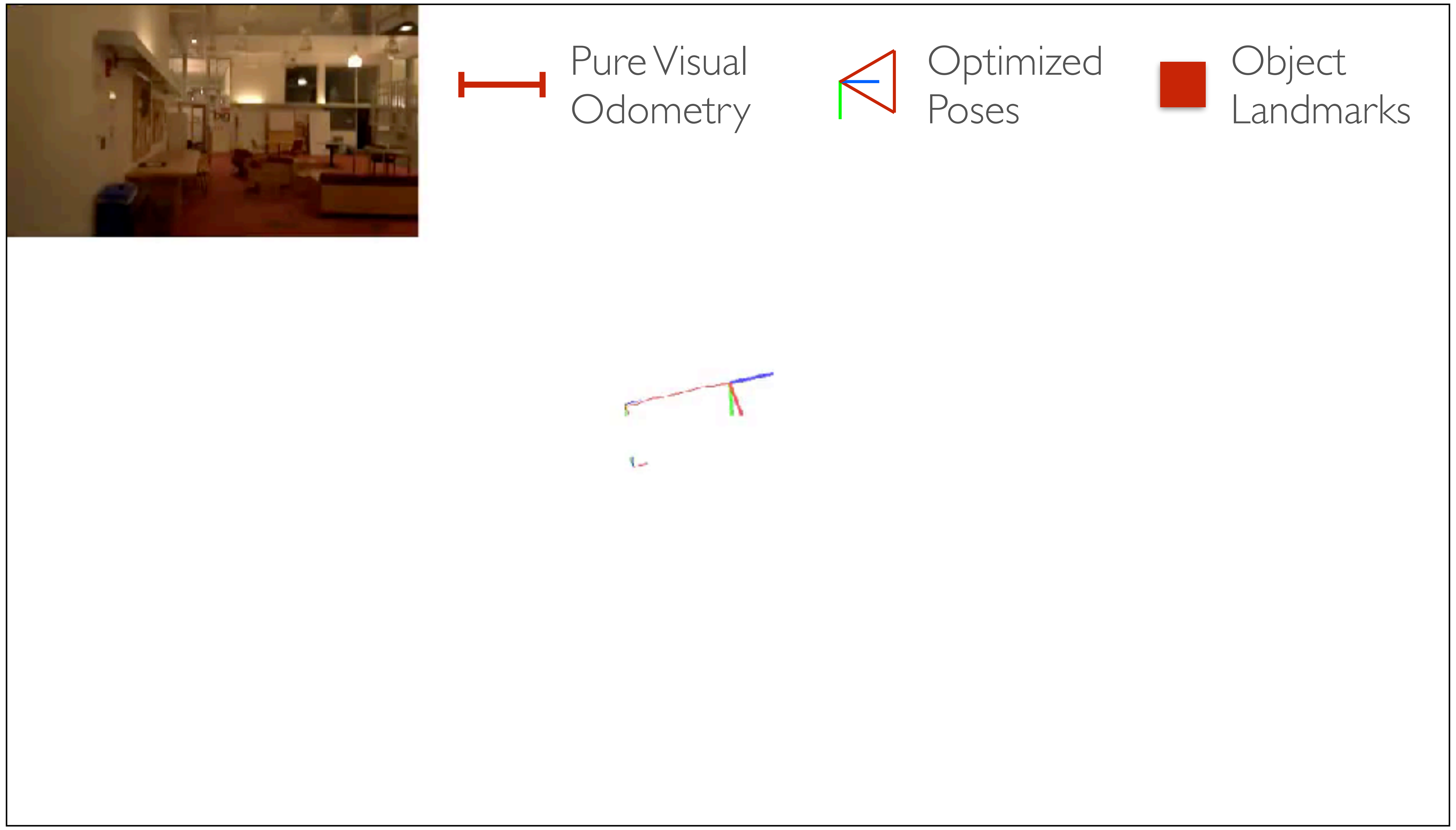
- SLAM-aware few-shot object learning
  - Randomly selected training labels with SLAM-aware label propagation
  - Despite fewer labels provided, SLAM-aware few-shot training can still achieve strong performance

Method	Frame-based Recognition mAP / Recall / F1-score	SLAM-aware Recognition mAP / Recall / F1-score
2-shot (Randomized)	80.5 / 63.4 / 69.7	83.1 / 74.8 / 77.1
5-shot (Randomized)	76.0 / 72.6 / 73.7	81.6 / 80.9 / 80.5
10-shot (Randomized)	79.6 / 74.5 / 76.0	81.6 / 82.2 / 81.5
20-shot (Randomized)	85.9 / 80.5 / 82.2	91.0 / 89.8 / 90.2
1-shot (SLAM-aware)	85.3 / 85.2 / 82.6	87.9 / 87.0 / 84.3
2-shot (SLAM-aware)	87.4 / 87.6 / 86.3	89.6 / 89.0 / 87.3
4-shot (SLAM-aware)	89.6 / 89.3 / 89.2	90.6 / 90.8 / 90.5

Comparison of SLAM-aware and randomized few-shot object learning

RECOGNITION-SUPPORTED SLAM

- **Object Recognition** as a front-end measurement for SLAM
  - Rich feature capacity
  - Scalable / Reduced complexity
  - Viewpoint, Lighting invariant
  - Pre-trained recognition models
  - Perceptual aliasing
  - Lack of contextual / scene knowledge



Future Work: Recognition-Supported SLAM

(Long-range loop-closure corrections with learned objects)

PRIOR ART

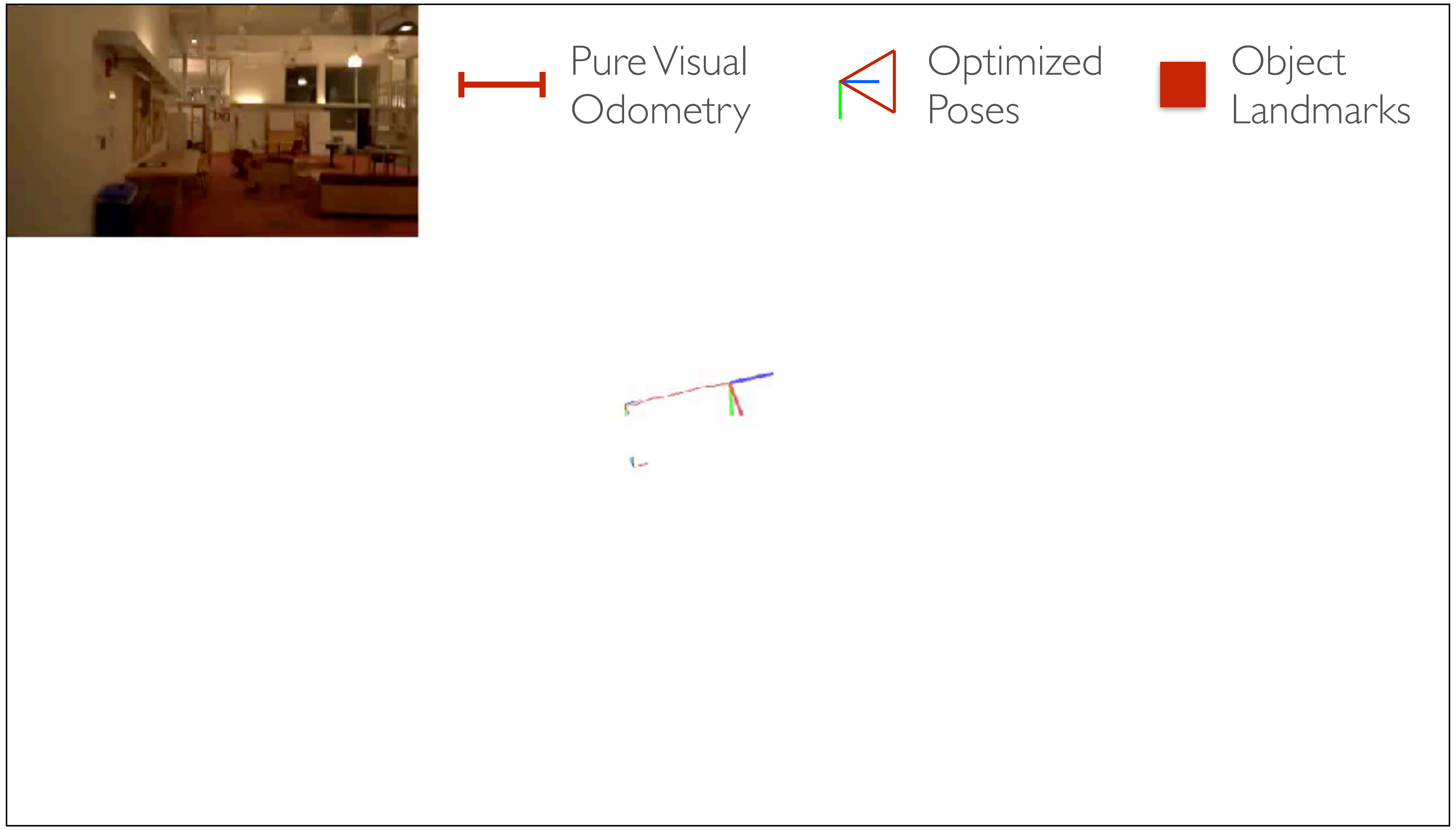
1. Object-based SLAM: SLAM++ *[Moreno et. al 2013]*
2. Semantic SFM *[Bao et. al 2011]*
3. Localization from Semantic Observations *[Antanasov et. al 2015]*

RECOGNITION-SUPPORTED SLAM

- **Object Recognition** as a front-end measurement for SLAM
  - Rich feature capacity
  - Scalable / Reduced complexity
  - Viewpoint, Lighting invariant
  - Pre-trained recognition models
  - Perceptual aliasing
  - Lack of contextual / scene knowledge

PRIOR ART

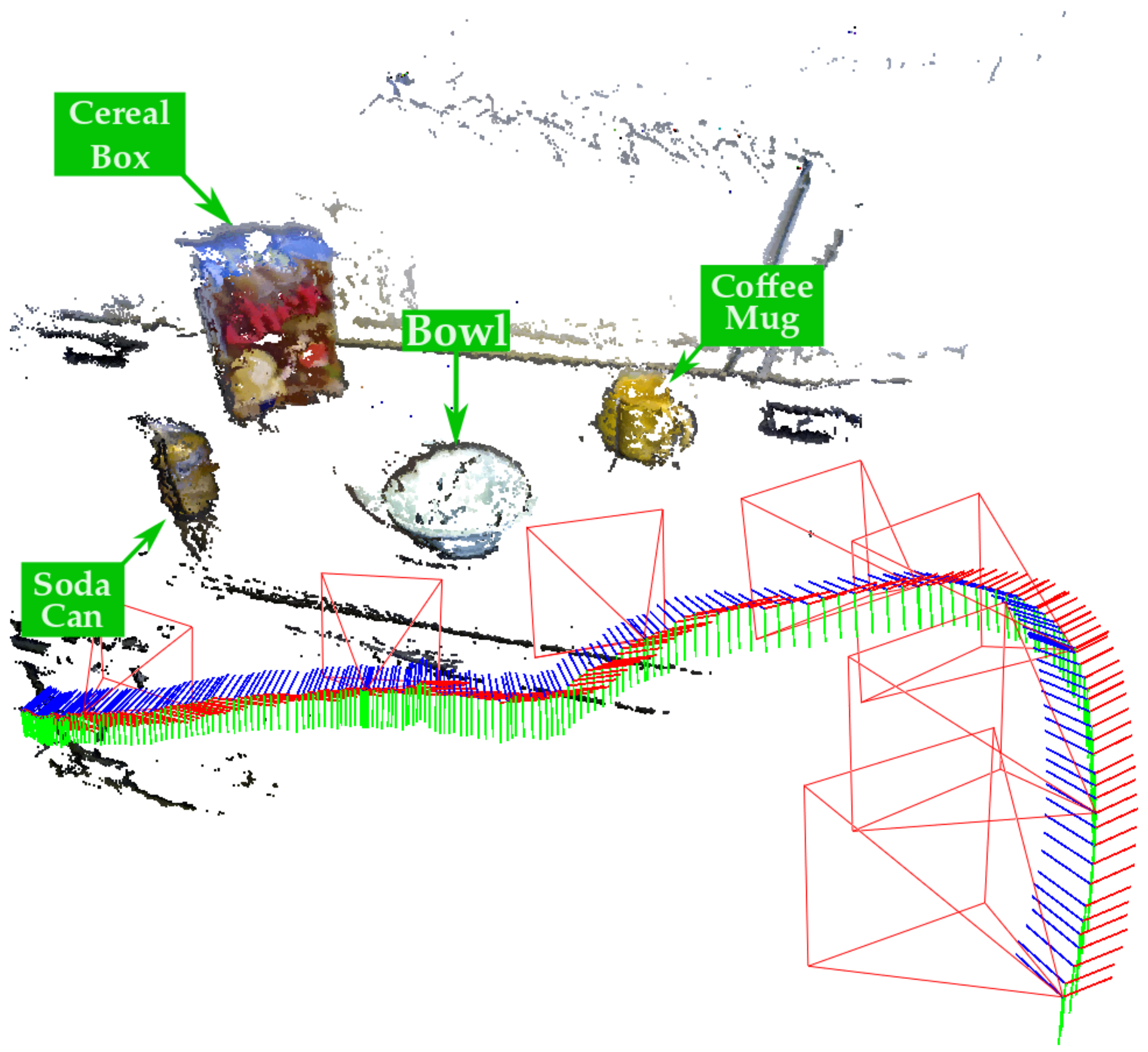
1. Object-based SLAM: SLAM++ *[Moreno et. al 2013]*
2. Semantic SFM *[Bao et. al 2011]*
3. Localization from Semantic Observations *[Antanasov et. al 2015]*



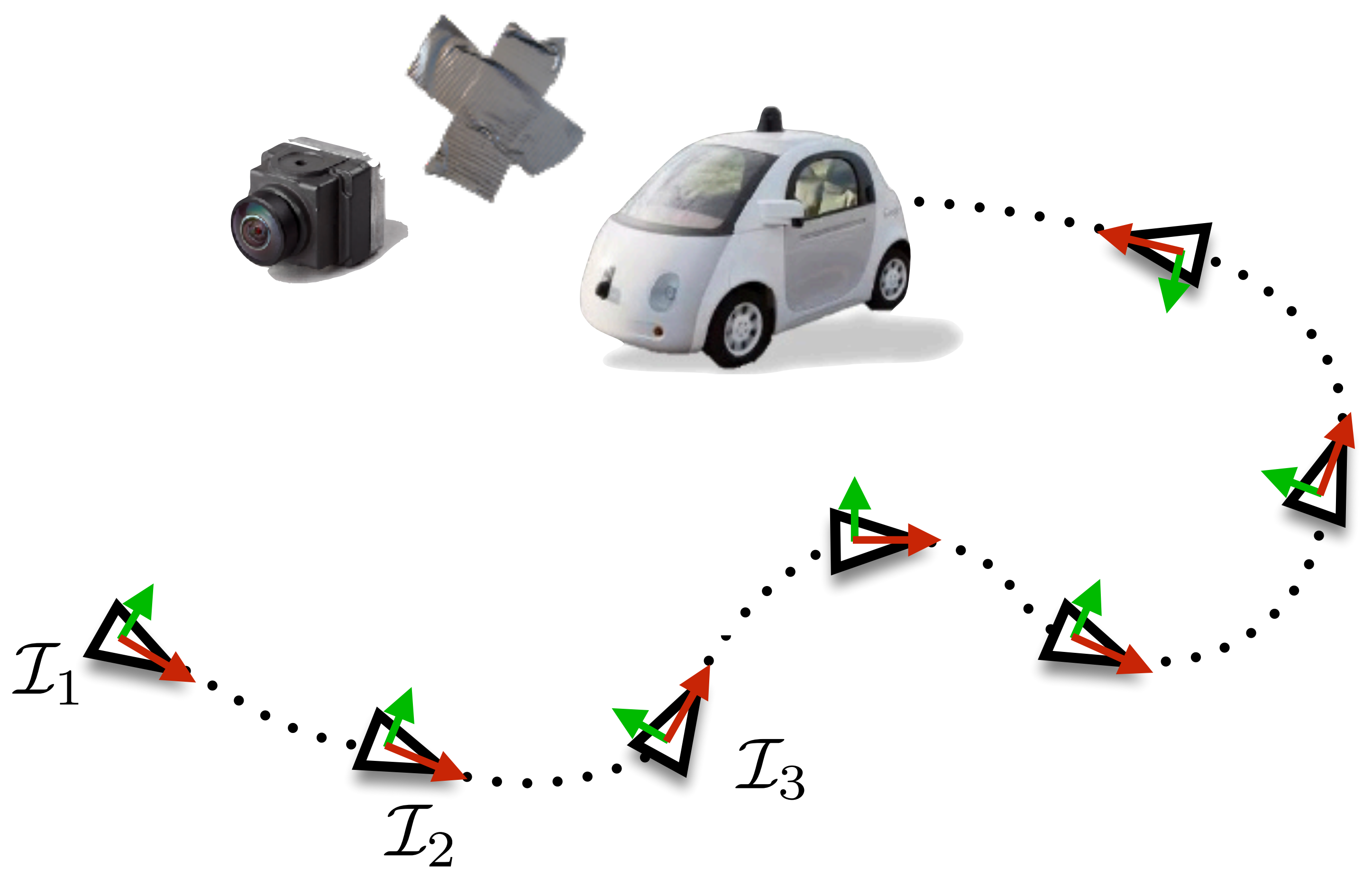
Future Work: Recognition-Supported SLAM
(Long-range loop-closure corrections with learned objects)

SLAM AS A SUPERVISORY SIGNAL

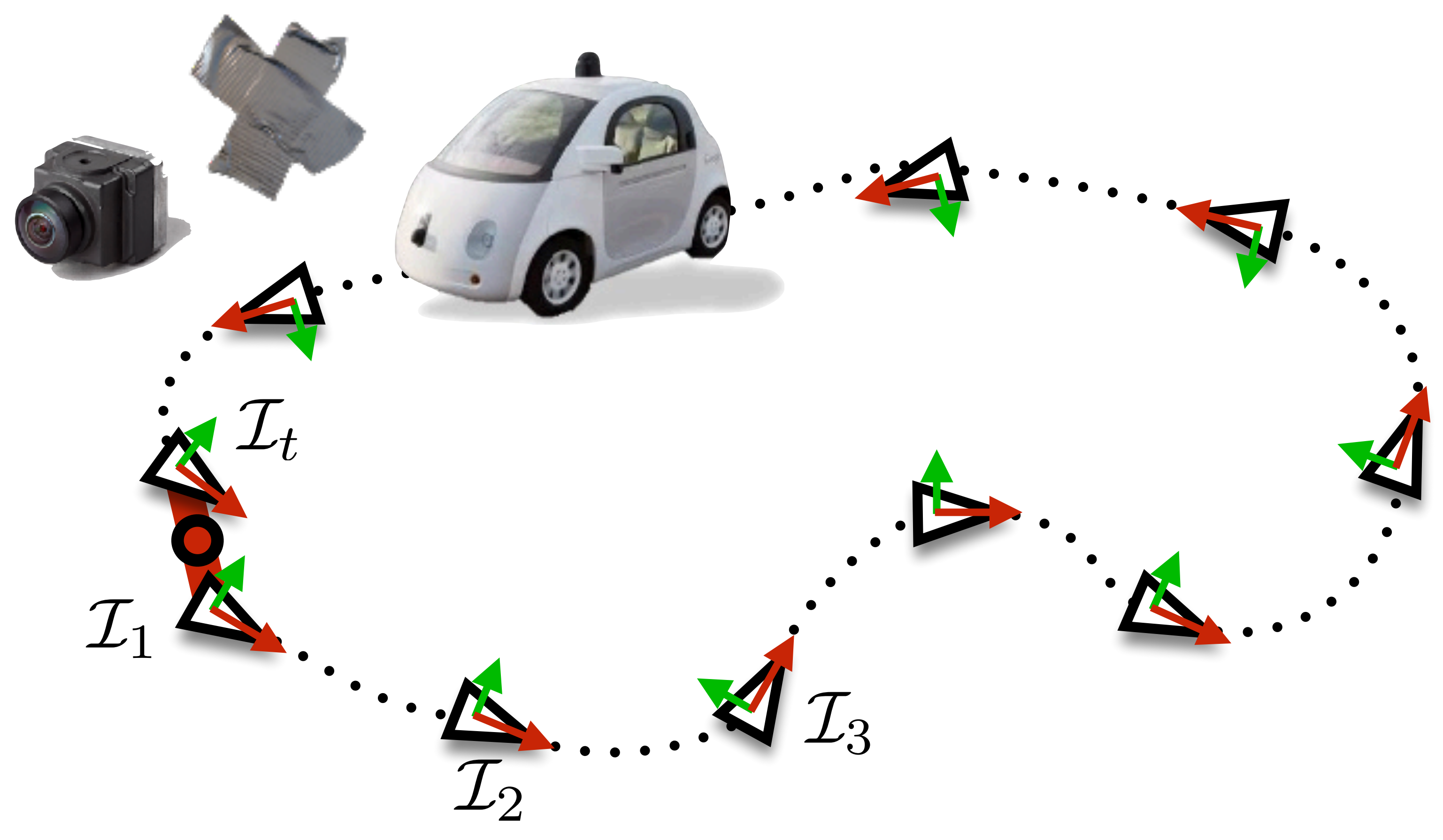
Monocular SLAM-Supported Object Recognition



Self-Supervised Visual Ego-motion Learning



Self-Supervised Visual Place Recognition Learning



SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

Correspondence Engine
(Geometric data association)

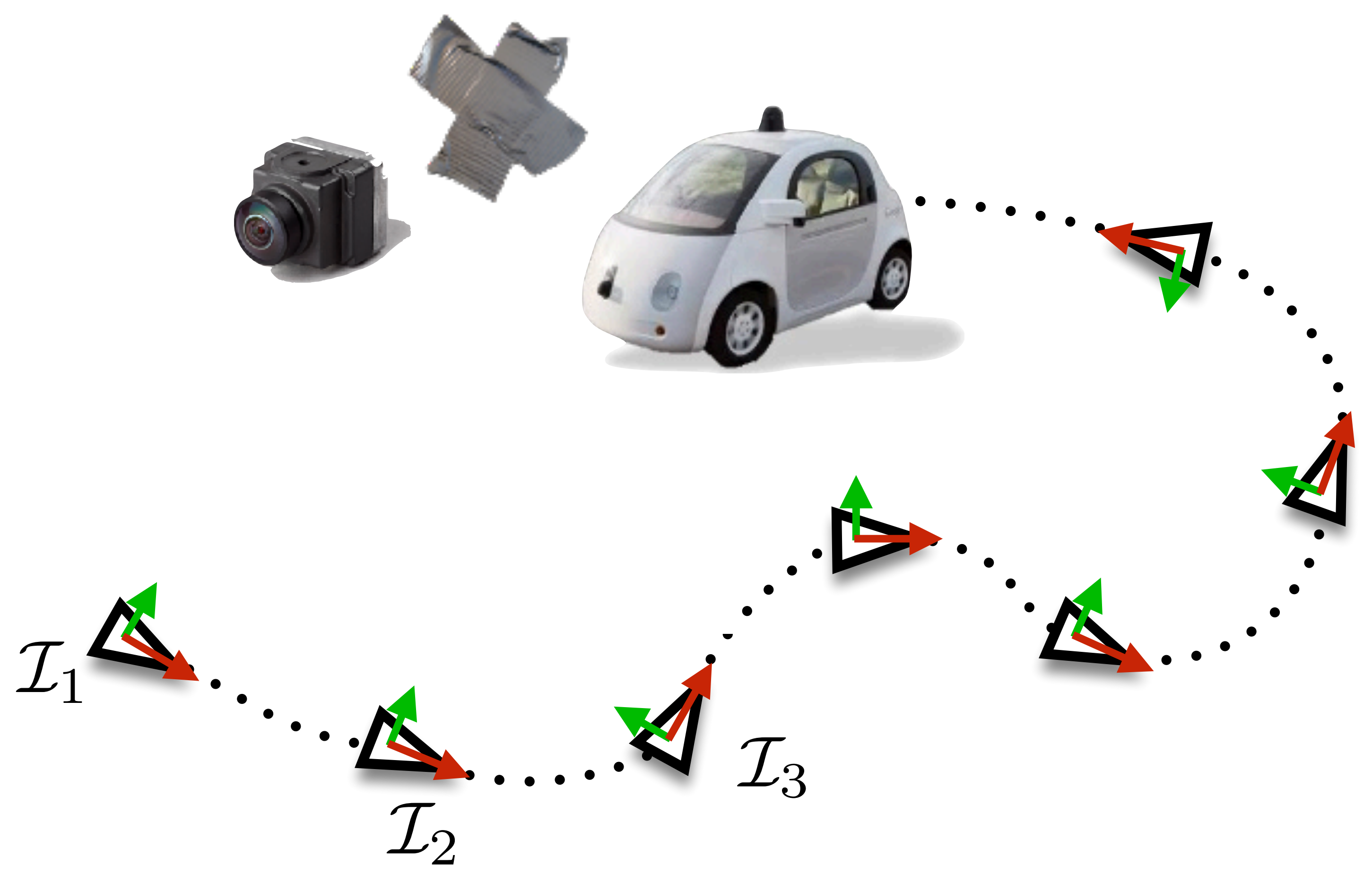
Self-Supervision
(SLAM-aided supervision)

Knowledge Transfer
(Bootstrapping)

SLAM AS A SUPERVISORY SIGNAL

Monocular SLAM-Supported Object Recognition

Self-Supervised Visual Ego-motion Learning



Self-Supervised Visual Place Recognition Learning

SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

Correspondence Engine
(Geometric data association)

Self-Supervision
(SLAM-aided supervision)

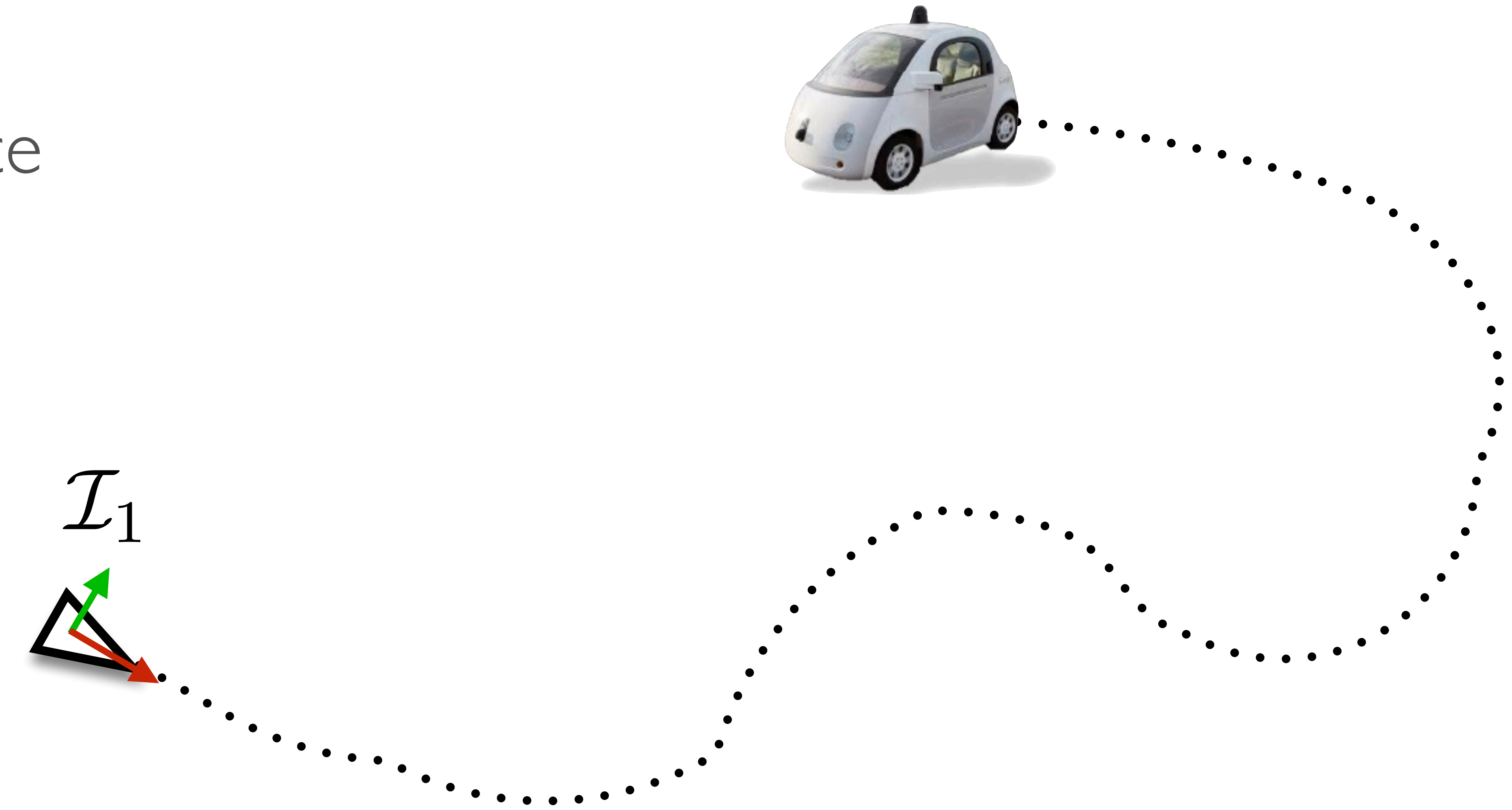
Knowledge Transfer
(Bootstrapping)

VISUAL EGO-MOTION

- **Visual Ego-motion / Visual Odometry**
  - Trace the trajectory of the camera given a continuous image sequence

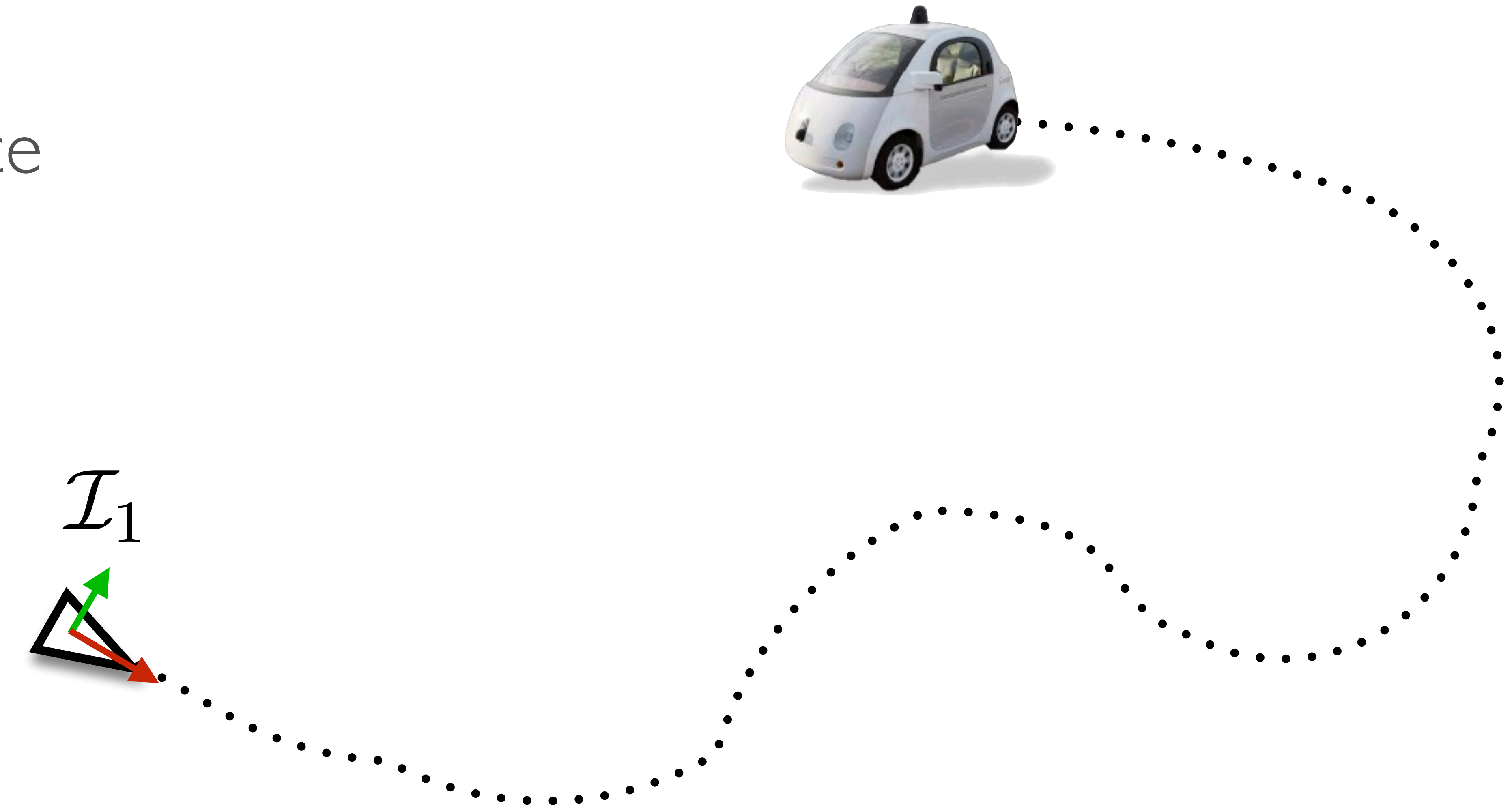
VISUAL EGO-MOTION

- **Visual Ego-motion / Visual Odometry**
  - Trace the trajectory of the camera given a continuous image sequence



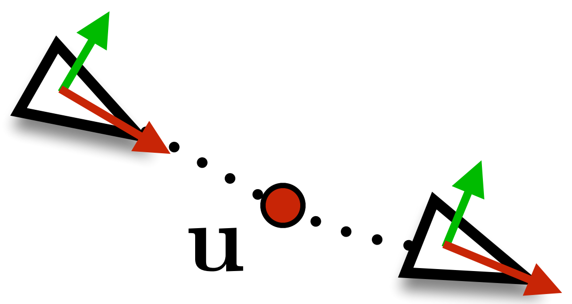
VISUAL EGO-MOTION

- **Visual Ego-motion / Visual Odometry**
  - Trace the trajectory of the camera given a continuous image sequence



DETERMINE f such that

$$f(\mathcal{I}_{t-1}, \mathcal{I}_t) \mapsto$$

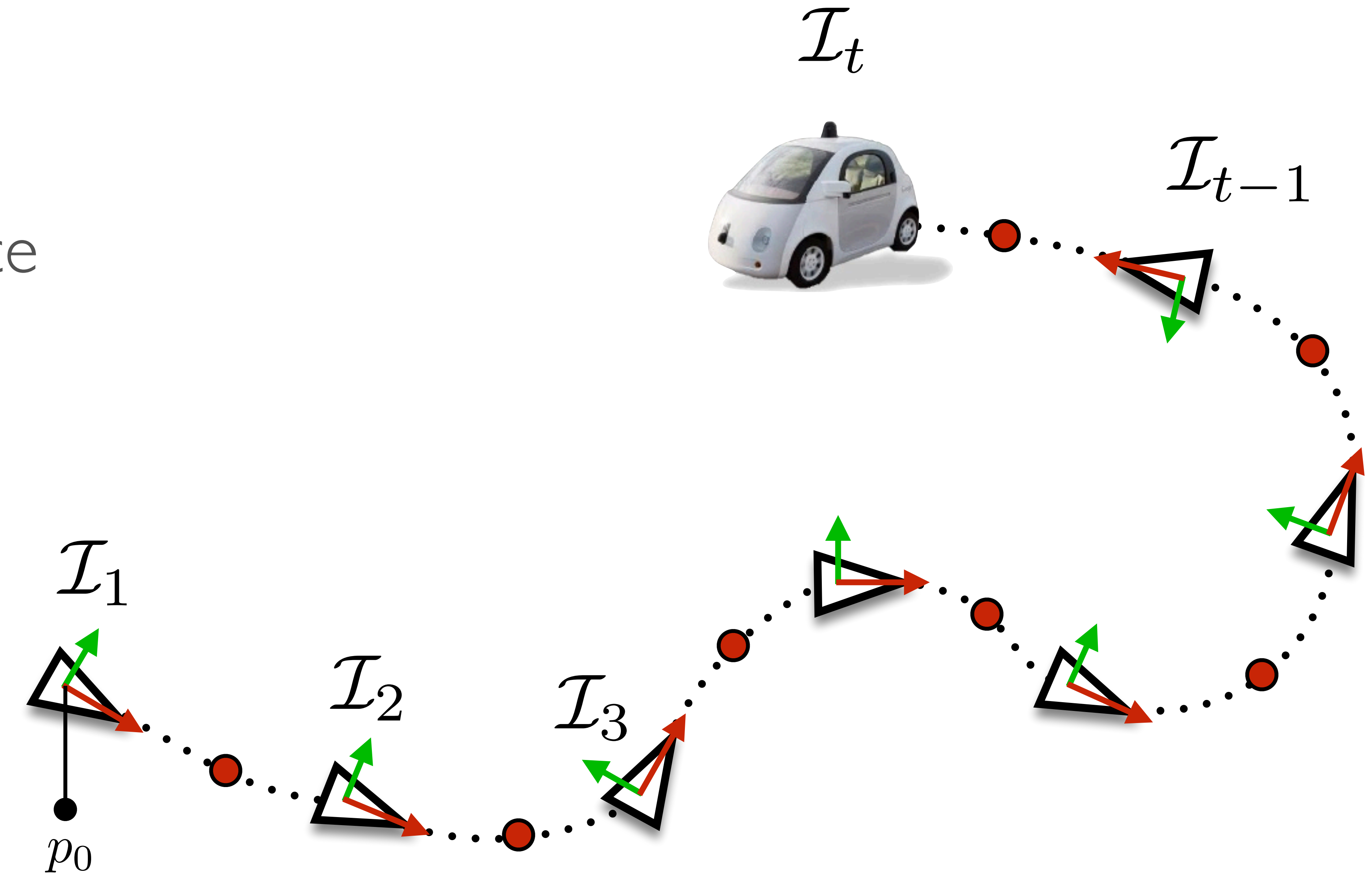


Subsequent Images

Odometry
(Relative motion)

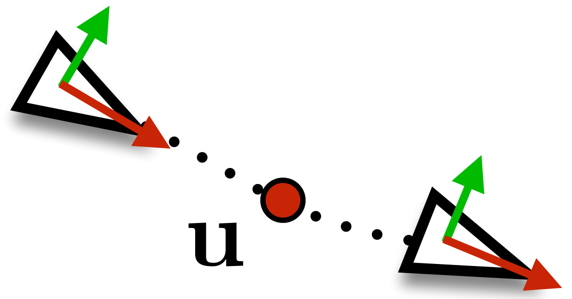
VISUAL EGO-MOTION

- **Visual Ego-motion / Visual Odometry**
  - Trace the trajectory of the camera given a continuous image sequence



DETERMINE f such that

$$f(\mathcal{I}_{t-1}, \mathcal{I}_t) \mapsto$$

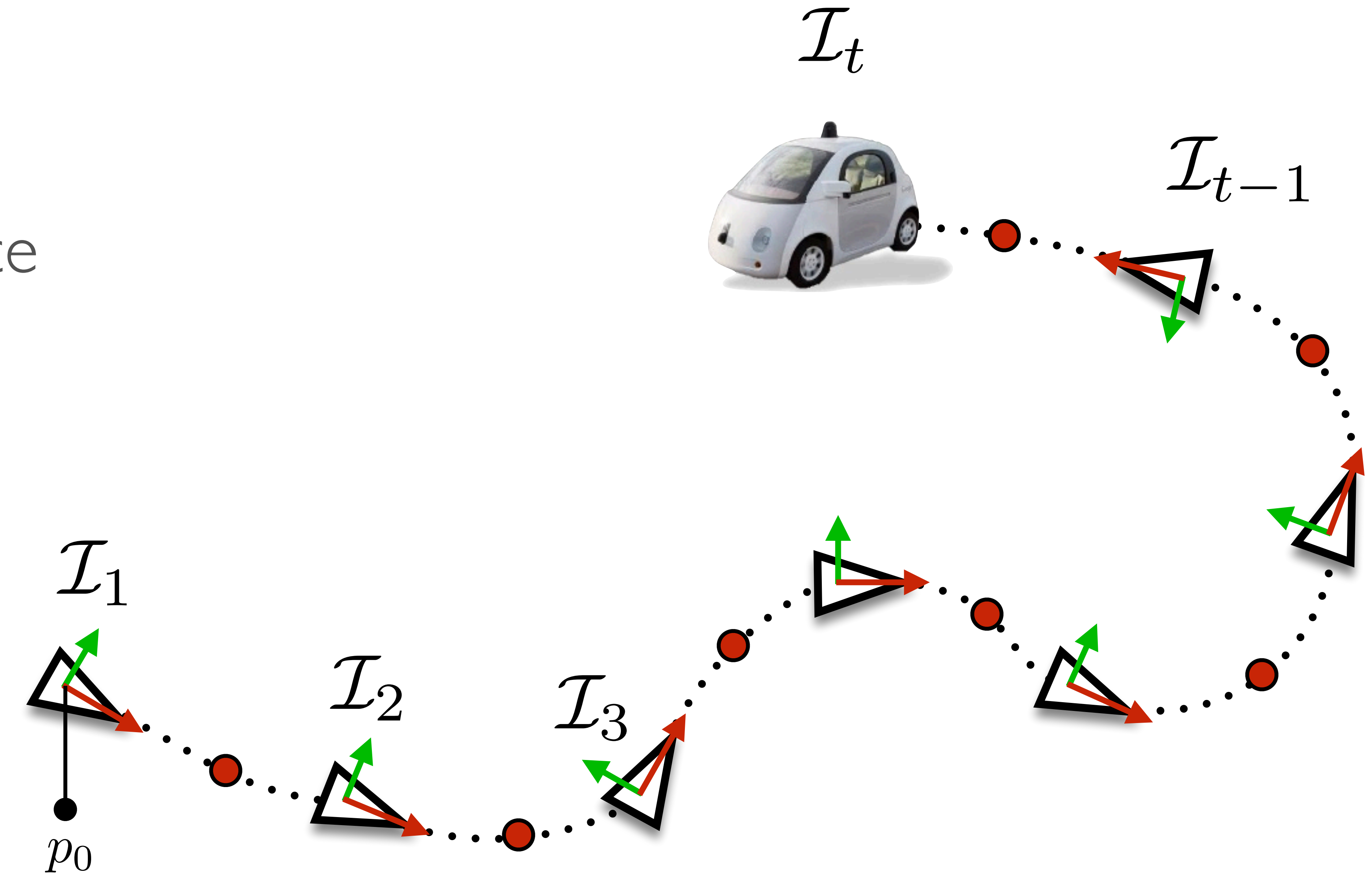


Subsequent Images

Odometry
(Relative motion)

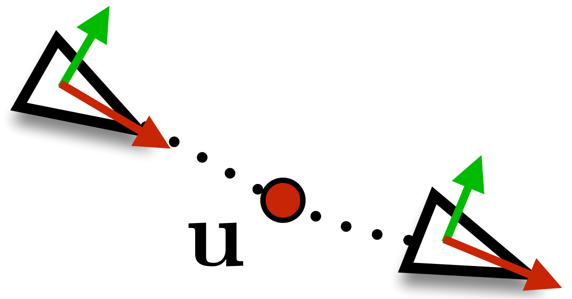
VISUAL EGO-MOTION

- **Visual Ego-motion / Visual Odometry**
  - Trace the trajectory of the camera given a continuous image sequence



DETERMINE f such that

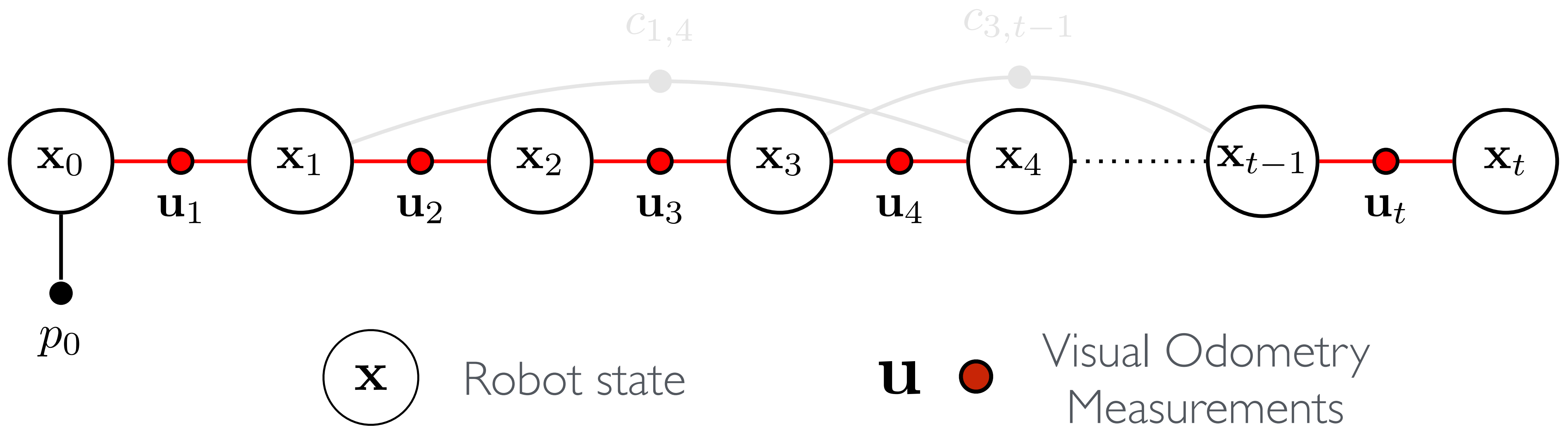
$$f(\mathcal{I}_{t-1}, \mathcal{I}_t) \mapsto$$



Subsequent Images

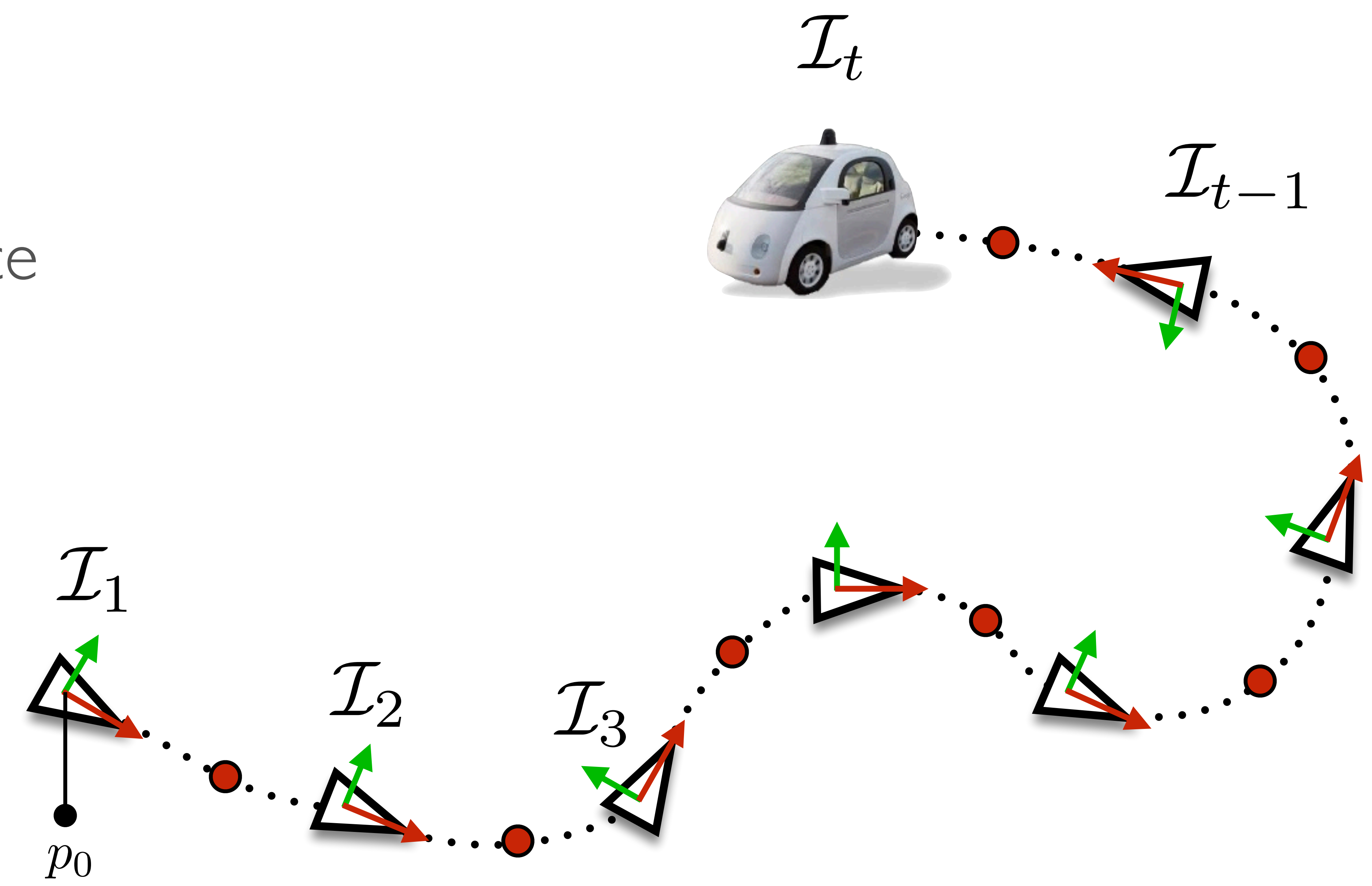
Odometry (Relative motion)

Factor Graph for Vision-based Pose-Graph SLAM



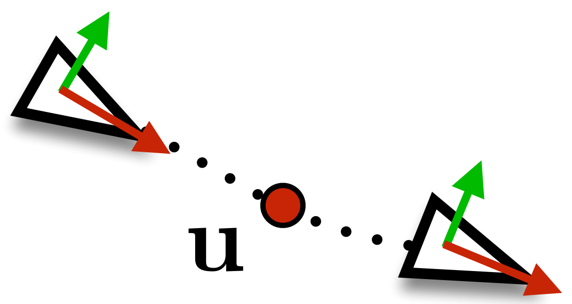
VISUAL EGO-MOTION

- **Visual Ego-motion / Visual Odometry**
  - Trace the trajectory of the camera given a continuous image sequence



DETERMINE f such that

$$f(\mathcal{I}_{t-1}, \mathcal{I}_t) \mapsto$$



Subsequent Images

Odometry (Relative motion)

Factor Graph for Vision-based Pose-Graph SLAM

$$\begin{aligned}\mathbf{X}^* &= \arg\max_{\mathbf{X}} p(\mathbf{X} \mid \mathbf{U}, \mathbf{Z_c}) \\ &= \arg\min_{\mathbf{X}} \Big\{ \underbrace{\sum_{i=1}^{M} \|f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i\|^2_{\Sigma_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{(j,k)\in\mathcal{C}} \|h_c(\mathbf{x}_j, \mathbf{x}_k) - \mathbf{z}_{jk}\|^2_{\Sigma_c}}_{\text{Loop-Closure Constraint Factors}} \Big\}\end{aligned}$$

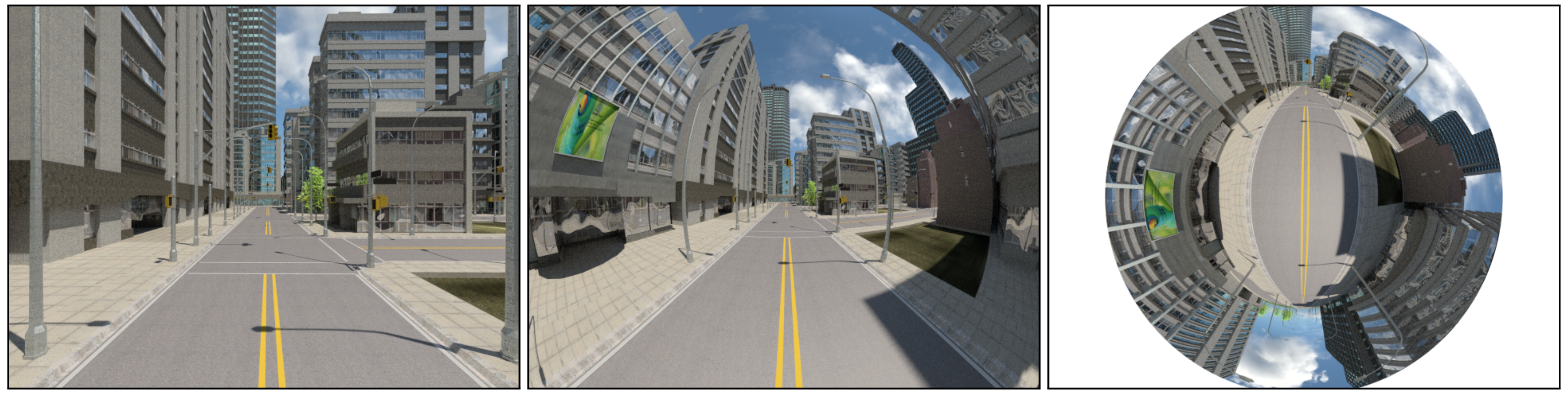
MOTIVATION

MOTIVATION

- Why learn Visual Ego-motion / Odometry?

MOTIVATION

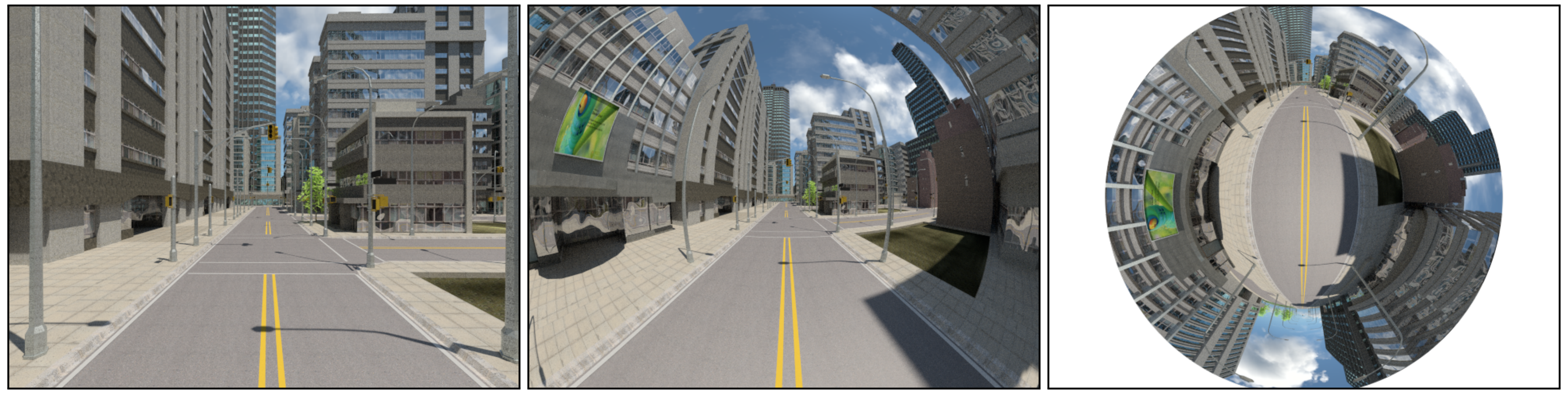
- **Why learn Visual Ego-motion / Odometry?**
  - **Varied camera optics:** Pinhole, Fisheye, Catadioptric



Varied Camera Optics
(a) Pinhole (b) Fisheye (c) Catadioptric

MOTIVATION

- **Why learn Visual Ego-motion / Odometry?**
  - **Varied camera optics:** Pinhole, Fisheye, Catadioptric
  - **Motion constraints:** Unconstrained VO, Constrained VO



Varied Camera Optics
(a) Pinhole (b) Fisheye (c) Catadioptric

Variants

- 2-D to 2-D
- 3-D to 3-D
- 3-D to 2-D (PnP)

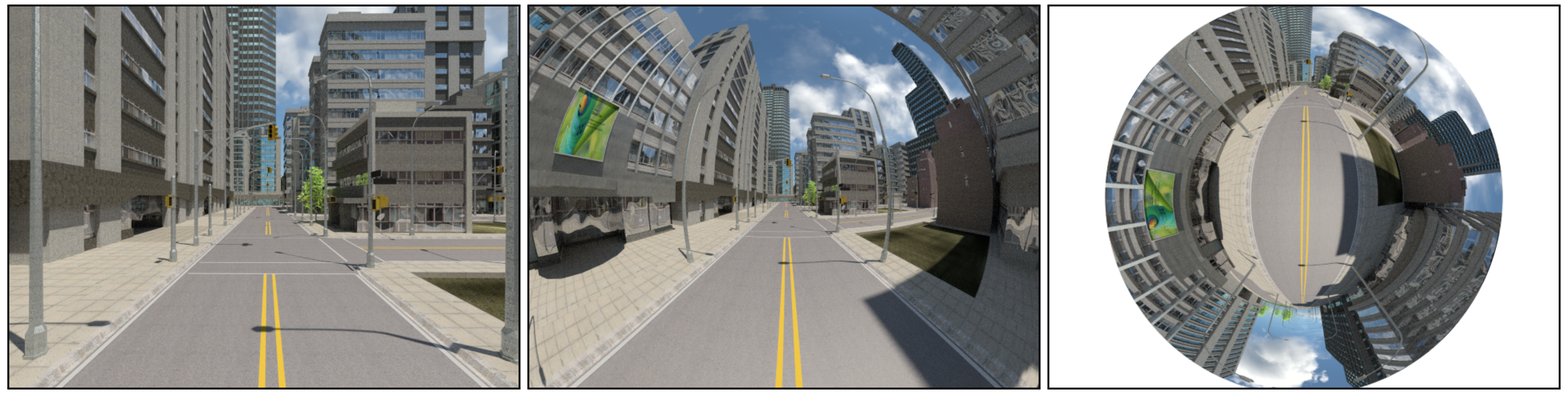
2-D to 2-D Variants

- 5-point
- 8-point
- 1-point, 2-point
- Stereo, RGB-D

[Scaramuzza et. al 2011]

MOTIVATION

- **Why learn Visual Ego-motion / Odometry?**
  - **Varied camera optics:** Pinhole, Fisheye, Catadioptric
  - **Motion constraints:** Unconstrained VO, Constrained VO
  - **Tedious calibration / monitoring:** Intrinsics, Extrinsics



Varied Camera Optics
(a) Pinhole (b) Fisheye (c) Catadioptric

GROWING SENSOR CONFIGURATION

MIT DGC Vehicle (2007)

Uber ATG Vehicle (2017)

Variants

- 2-D to 2-D
- 3-D to 3-D
- 3-D to 2-D (PnP)

2-D to 2-D Variants

- 5-point
- 8-point
- 1-point, 2-point
- Stereo, RGB-D

[Scaramuzza et. al 2011]

SELF-SUPERVISED
VISUAL EGO-MOTION LEARNING

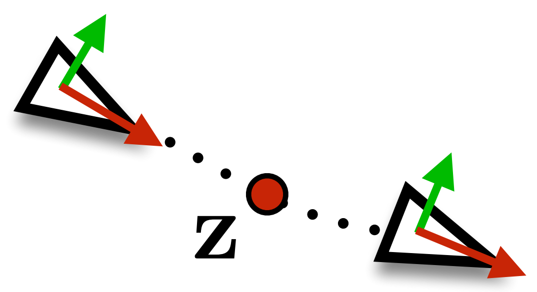
SELF-SUPERVISED
VISUAL EGO-MOTION LEARNING

- **“Ground-truth” Trajectory Generation**
  - Generate target variables for self-supervision

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- **"Ground-truth" Trajectory Generation**
  - Generate target variables for self-supervision

$$f(\mathcal{I}_{t-1}, \mathcal{I}_t) \mapsto$$



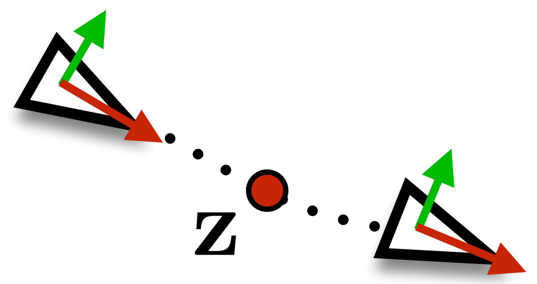
Subsequent Images

Odometry
(Relative motion)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

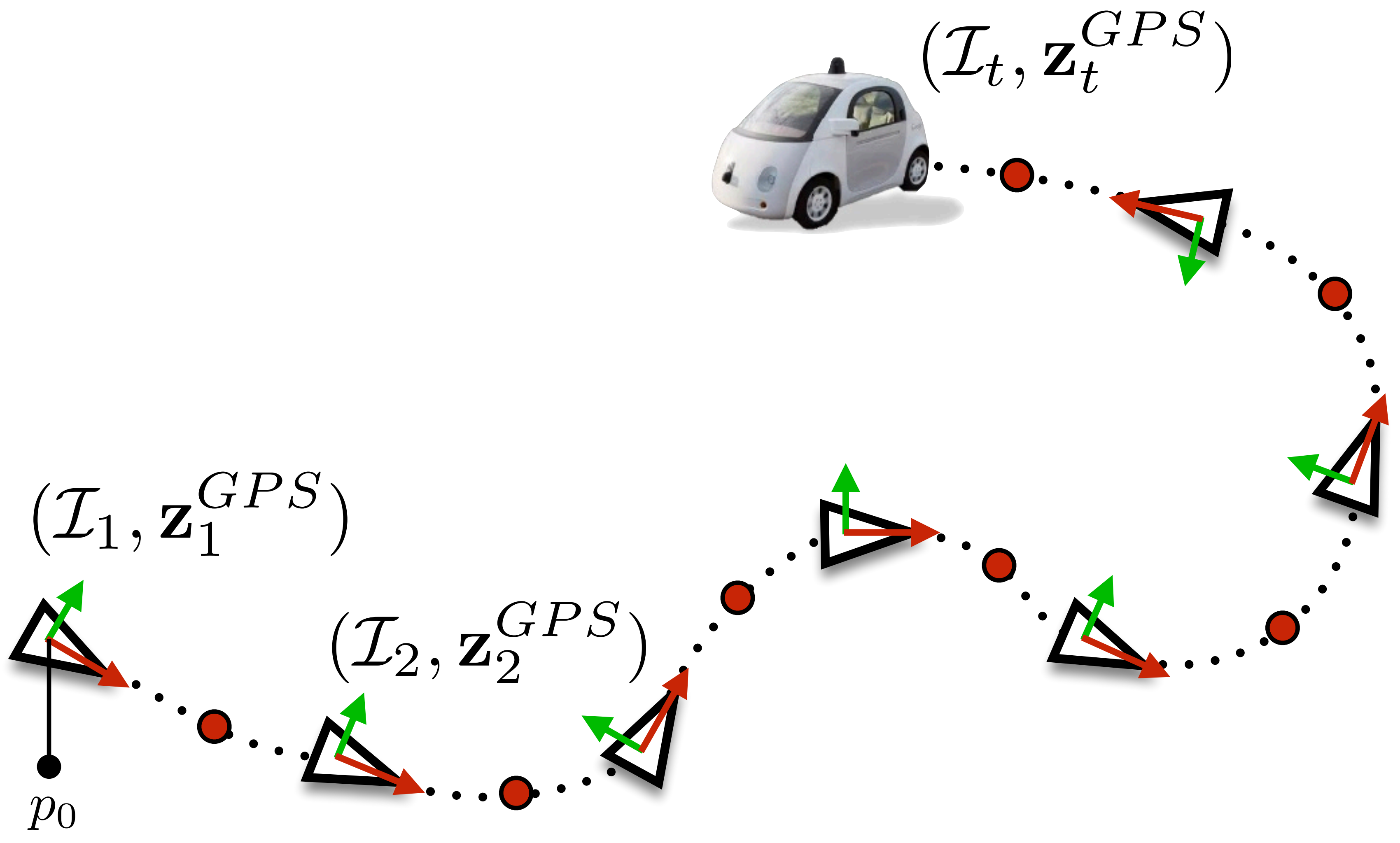
- **"Ground-truth" Trajectory Generation**
  - Generate target variables for self-supervision

$$f(\mathcal{I}_{t-1}, \mathcal{I}_t) \mapsto$$



Subsequent Images

Odometry
(Relative motion)

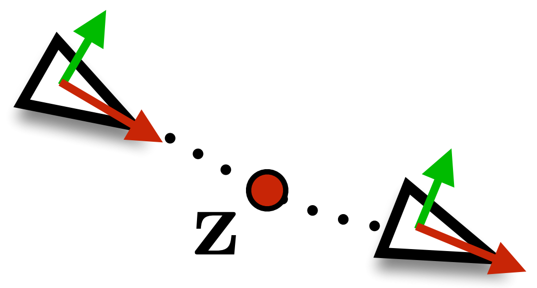


  - Natural synchronization of Images/GPS/INS/Wheel Odometry to first solve a GPS-aided localization problem

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

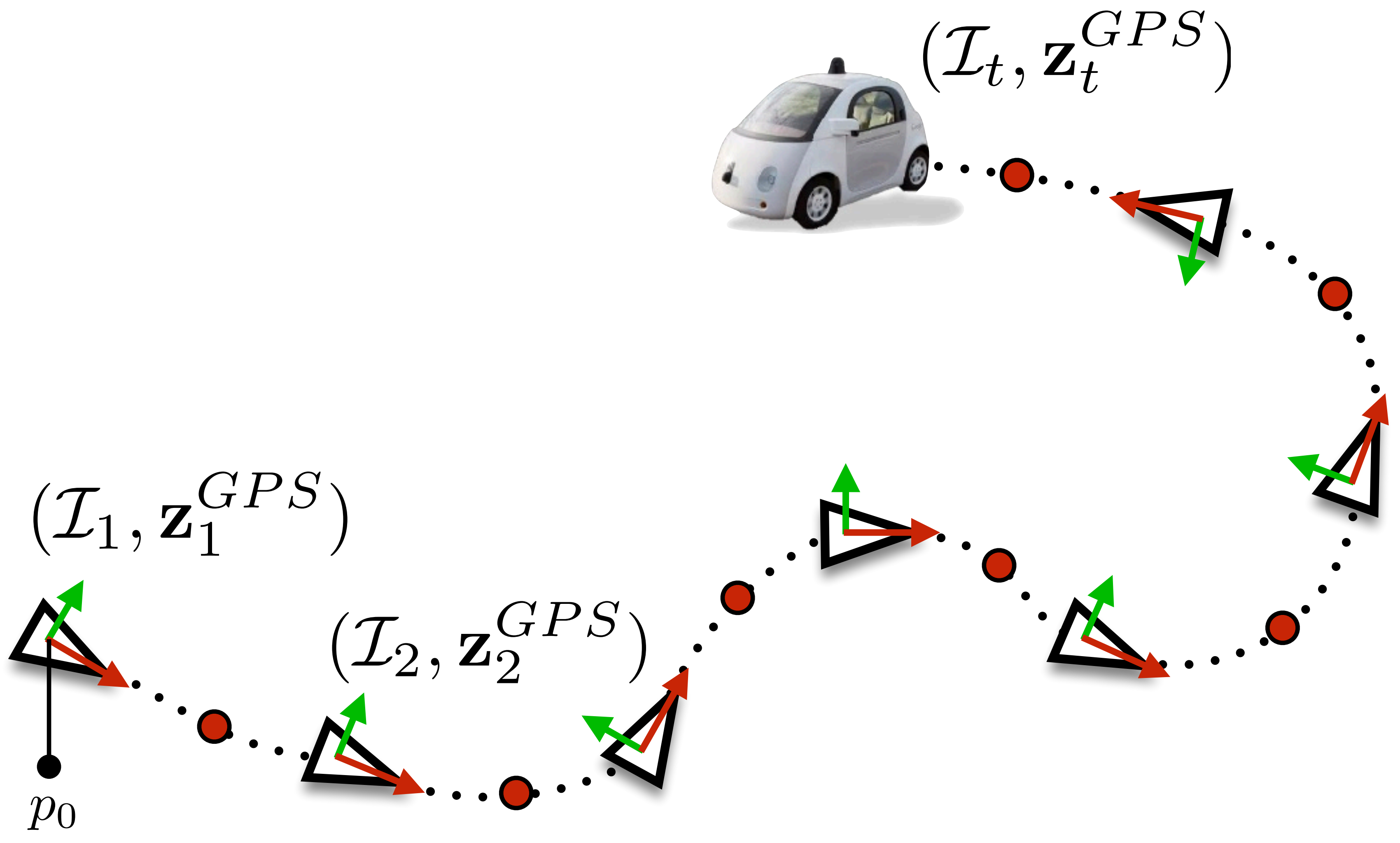
- **"Ground-truth" Trajectory Generation**
  - Generate target variables for self-supervision

$$f(\mathcal{I}_{t-1}, \mathcal{I}_t) \quad \mapsto$$

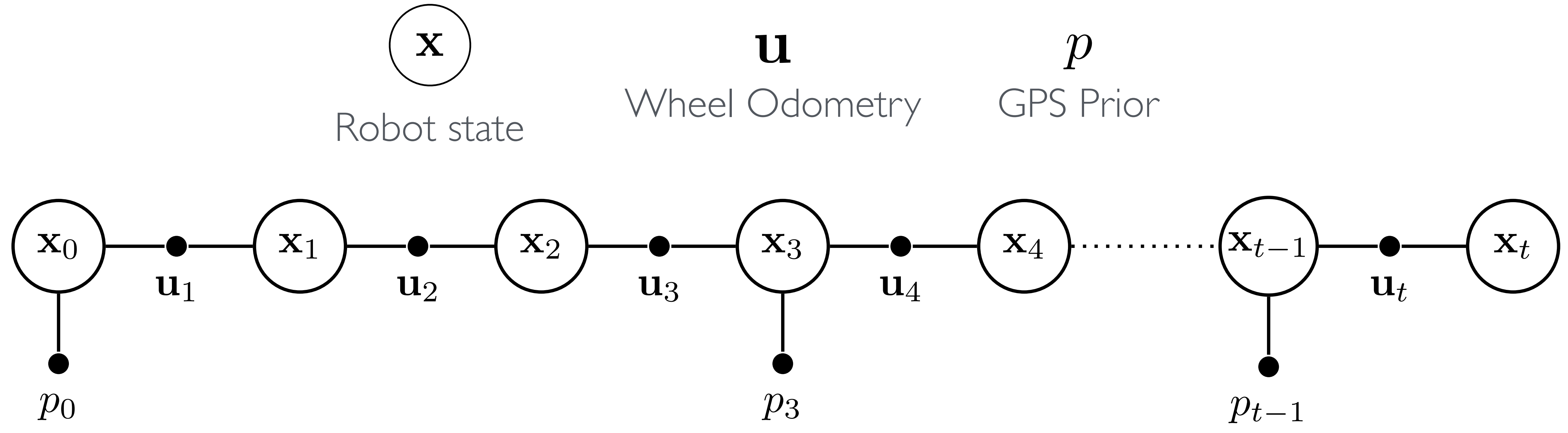


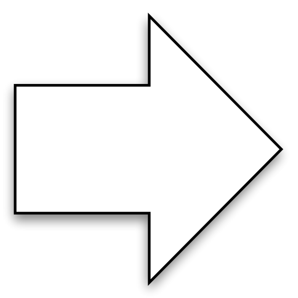
Subsequent Images

Odometry (Relative motion)

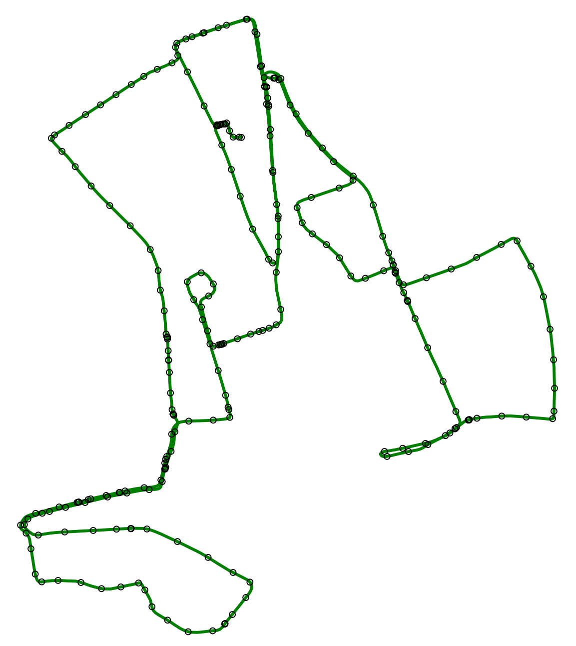


  - Natural synchronization of Images/GPS/INS/Wheel Odometry to first solve a GPS-aided localization problem





Fused Ego-motion Trajectory

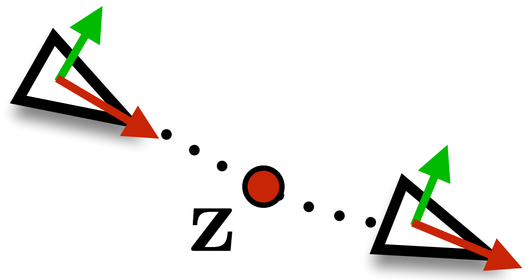


Long-term, drift-free, accurate robot trajectory

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

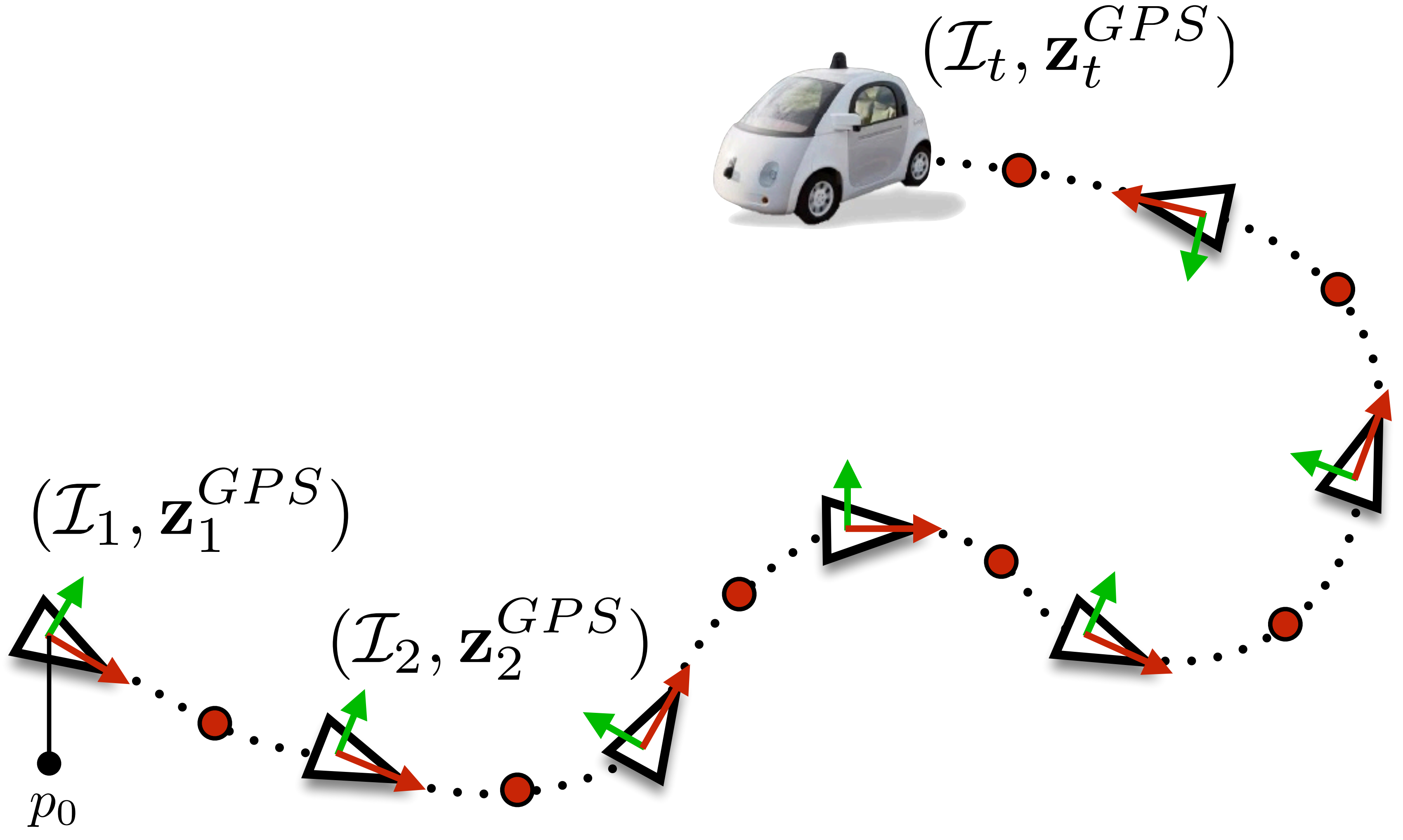
- **"Ground-truth" Trajectory Generation**
  - Generate target variables for self-supervision

$$f(\mathcal{I}_{t-1}, \mathcal{I}_t) \mapsto$$



Subsequent Images

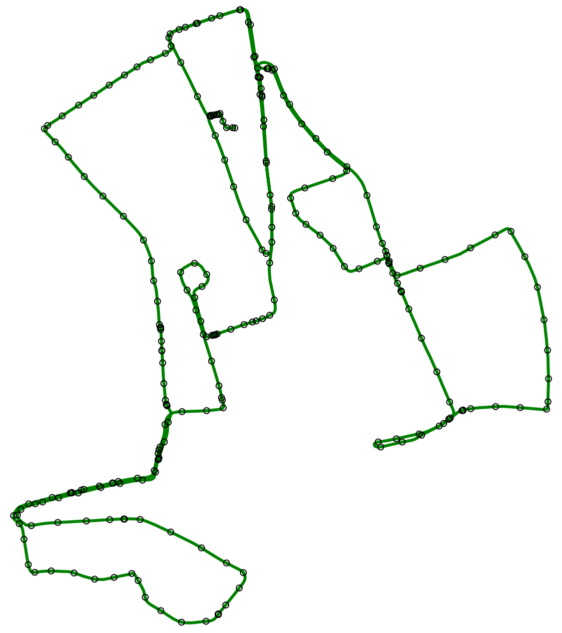
Odometry
(Relative motion)



  - Natural synchronization of Images/GPS/INS/Wheel Odometry to first solve a GPS-aided localization problem

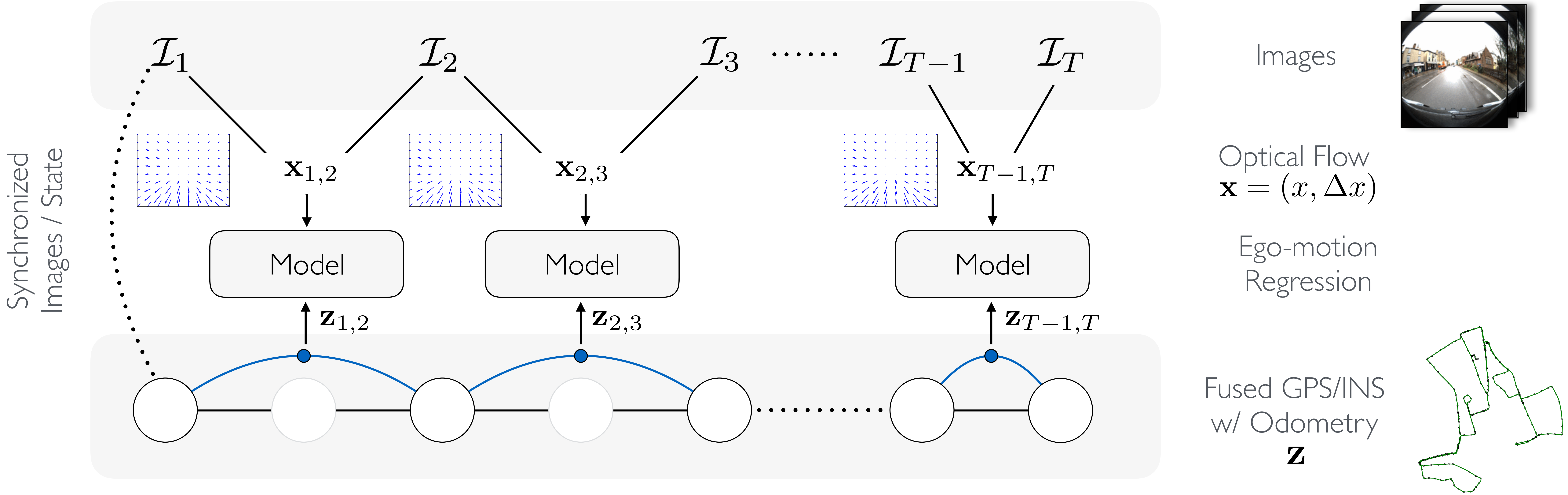
$$\mathbf{X}^* = \arg\max_{\mathbf{X}} p(\mathbf{X} \mid \mathbf{U}, \mathbf{Z_g})$$
$$= \arg\min_{\mathbf{X}} \left\{ \underbrace{\sum_{i=1}^{M} \| f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i \|^2_{\sum_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{j=1}^{G} \| h_g(\mathbf{x}_j) - \mathbf{z}_j \|^2_{\sum_g}}_{\text{GPS Measurement Priors}} \right\}$$

Fused Ego-motion Trajectory



Long-term, drift-free, accurate robot trajectory

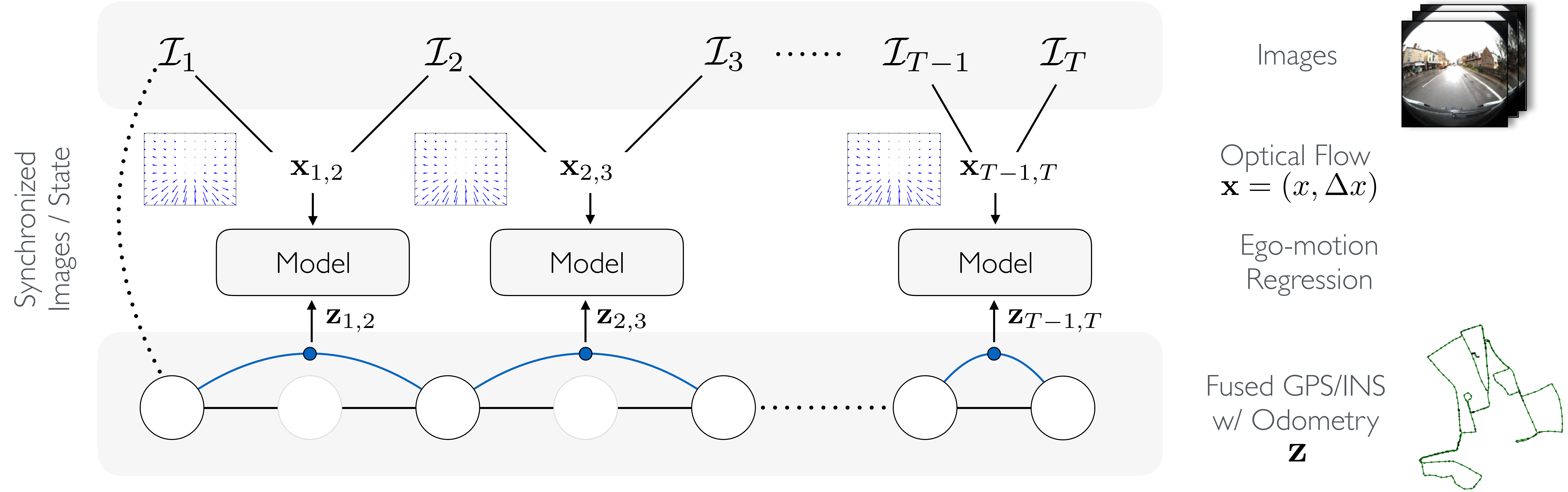
SELF-SUPERVISED VISUAL EGO-MOTION LEARNING



$$f : \mathbf{x} \mapsto \mathbf{z}$$

EGO-MOTION REGRESSION

SELF-SUPERVISED
VISUAL EGO-MOTION LEARNING



$$f : \mathbf{x} \mapsto \mathbf{z}$$

EGO-MOTION REGRESSION

Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Contributions

Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

SELF-SUPERVISED
VISUAL EGO-MOTION LEARNING

- Contributions
  - Ego-motion as a learned density estimation problem

Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Contributions
  - Ego-motion as a learned density estimation problem
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)

Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Contributions
  - Ego-motion as a learned density estimation problem
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)
  - Introspective model-based reasoning

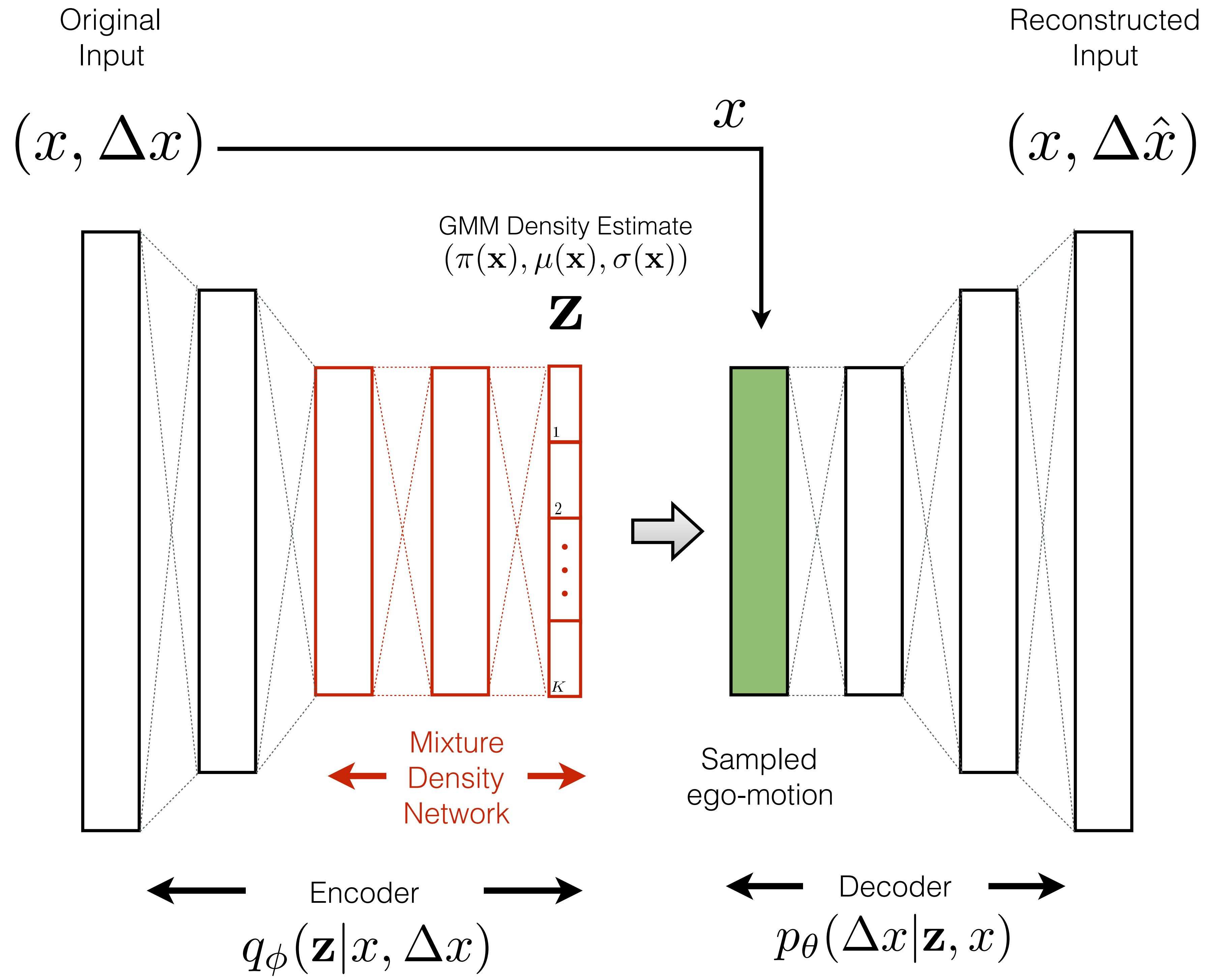
Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Contributions
  - Ego-motion as a learned density estimation problem
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)
  - Introspective model-based reasoning

Mixture Density Network (MDN) in a Conditional VAE (C-VAE)

Towards Visual Ego-motion Learning in Robots (IROS '17)



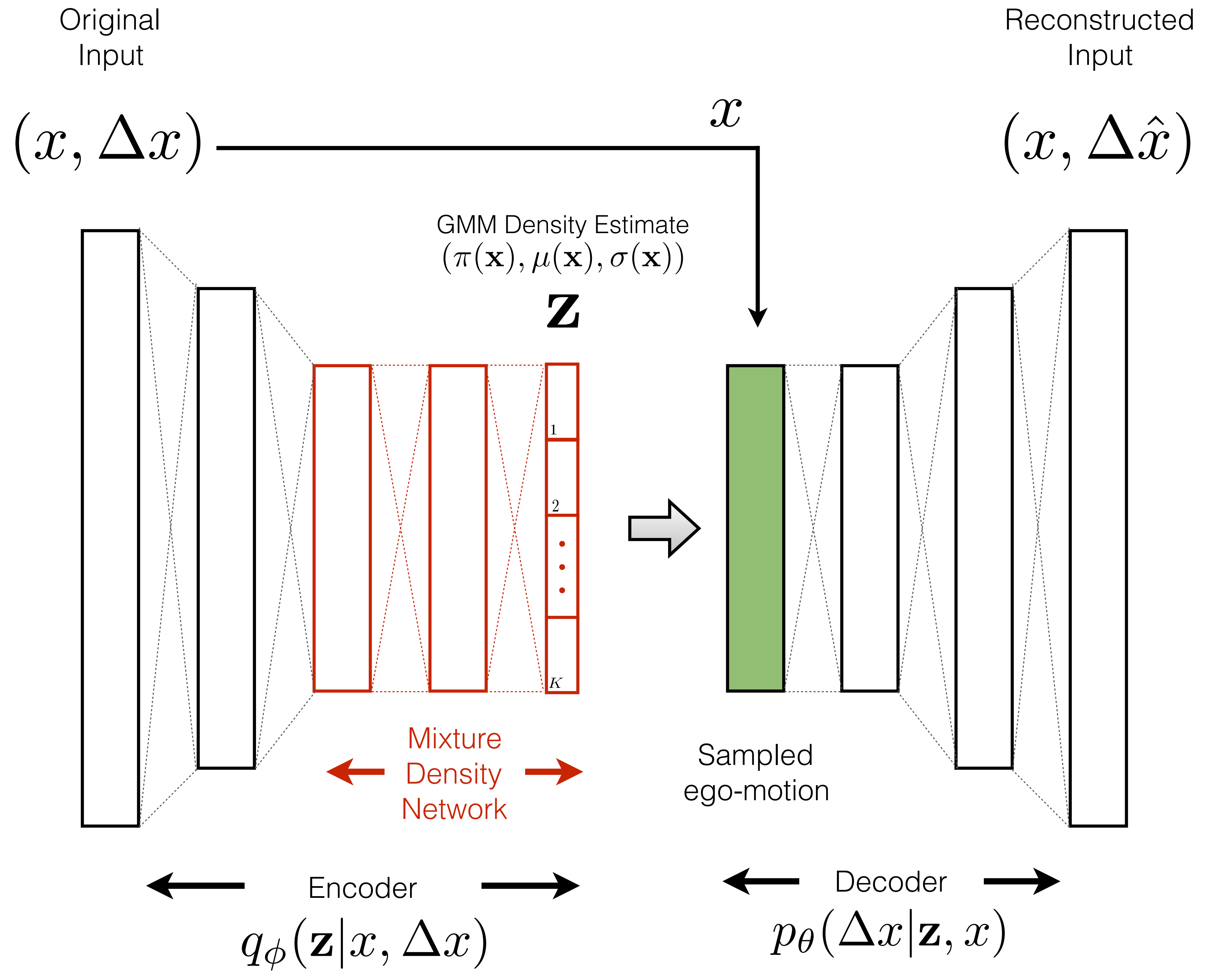
Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Contributions
  - Ego-motion as a learned density estimation problem
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)
  - Introspective model-based reasoning

Mixture Density Network (MDN) in a Conditional VAE (C-VAE)

Towards Visual Ego-motion Learning in Robots (IROS '17)



Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

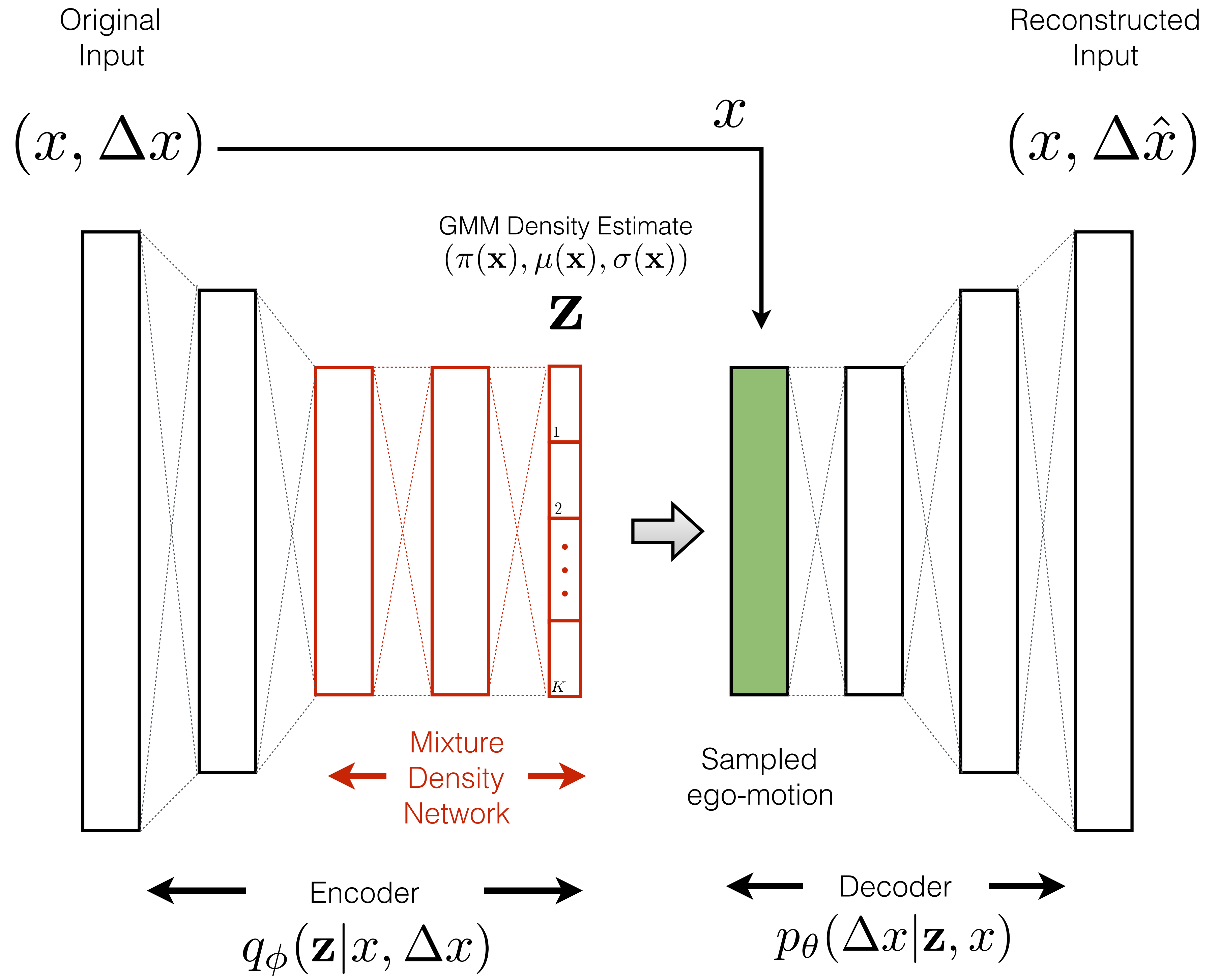
SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Contributions
  - Ego-motion as a learned density estimation problem
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)
  - Introspective model-based reasoning

$\mathbf{z}$	Ego-motion density estimate
$\mathbf{x} = (x, \Delta x)$	Input feature location, and optical flow
$p_\theta(\Delta x \vert \mathbf{z}, x)$	Decoder estimating scene flow given input feature location and sampled ego-motion
$q_\phi(\mathbf{z} \vert x, \Delta x)$	Encoder estimating ego-motion pdf given input feature location and flow

Mixture Density Network (MDN) in a Conditional VAE (C-VAE)

Towards Visual Ego-motion Learning in Robots (IROS '17)



Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Ego-motion Density Estimation
  - Mixture Density Network (MDN): Neural Network whose outputs are parameters of a Gaussian Mixture Model (GMM)

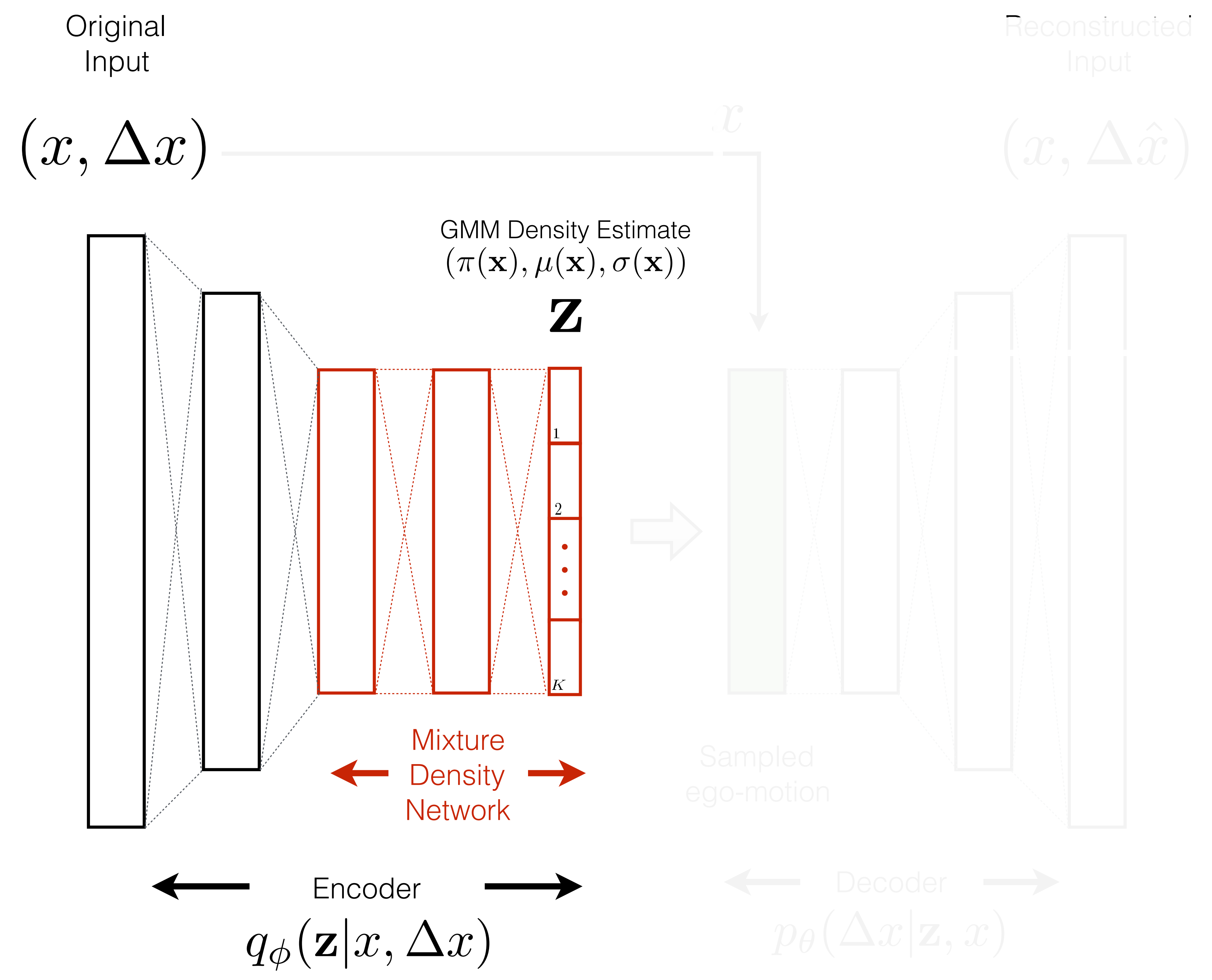
$$f^{vo} : \mathbf{x} \mapsto \underbrace{\Big(\mu(\mathbf{x}_{t-1,t}), \sigma(\mathbf{x}_{t-1,t}), \pi(\mathbf{x}_{t-1,t})\Big)}_{\mathbf{z}}$$

Optical Flow (under $\mathbf{x}$)

Ego-motion density estimate (under $\mathbf{z}$)

Mixture Density Network (MDN) in a Conditional VAE (C-VAE)

Towards Visual Ego-motion Learning in Robots (IROS '17)



SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Ego-motion Density Estimation
  - Mixture Density Network (MDN): Neural Network whose outputs are parameters of a Gaussian Mixture Model (GMM)

$$f^{vo} : \mathbf{x} \mapsto \underbrace{\Big(\mu(\mathbf{x}_{t-1,t}), \sigma(\mathbf{x}_{t-1,t}), \pi(\mathbf{x}_{t-1,t})\Big)}_{\mathbf{z}}$$

Optical Flow ($\mathbf{x}$)

Ego-motion density estimate ($\mathbf{z}$)

GMM

$$p(\mathbf{z}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_k(\mathbf{x}_i)\mathcal{N}(\mathbf{z} \mid \mu_k(\mathbf{x}_i), \sigma_k^2(\mathbf{x}_i))$$

(Ego-motion density estimate given optical flow)

$$\mathcal{L}_{MDN} = -\sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k(\mathbf{x}_n)\mathcal{N}(\mathbf{z} \mid \mu_k(\mathbf{x}_n), \sigma_k^2(\mathbf{x}_n)) \right\}$$

(Minimize neg. log-likelihood under the GMM model)

Outputs

$$\pi_k(\boldsymbol{x}) = \frac{\exp(a_k^\pi)}{\sum_{l=1}^{K} \exp(a_l^\pi)}$$

$$\sigma_k(\boldsymbol{x}) = \exp(a_k^\sigma), \quad \mu_k(\boldsymbol{x}) = a_k^\mu$$

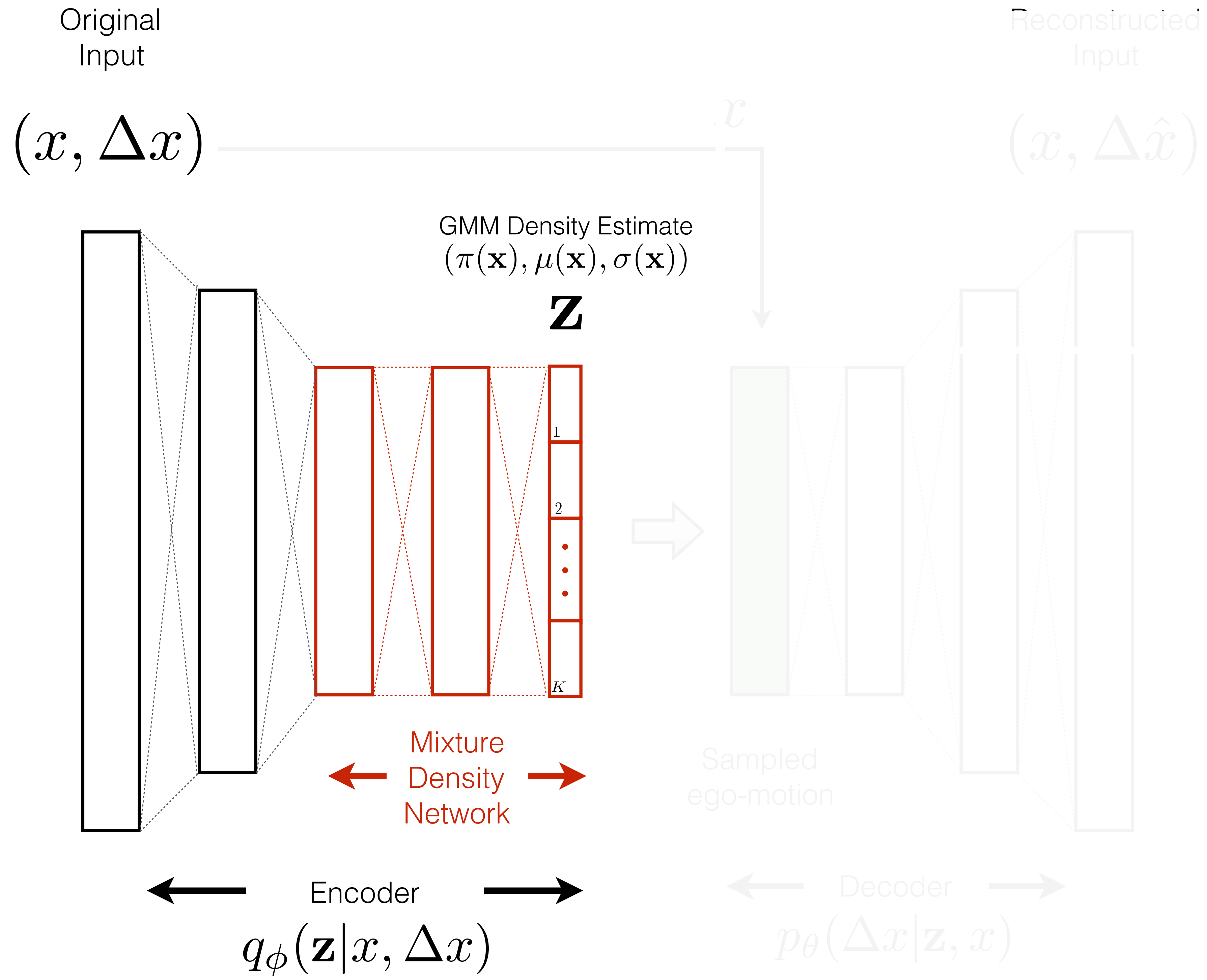
Constraints (via activations)

$$\textstyle\sum_K \pi_k(\boldsymbol{x}) = 1$$

$$0 \leq \pi_k(\boldsymbol{x}) \leq 1$$

Mixture Density Network (MDN) in a Conditional VAE (C-VAE)

Towards Visual Ego-motion Learning in Robots (IROS '17)



Sparse Optical Flow ⇨ Ego-motion Density ⇨ Predicted Feature Tracks

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Density Estimation with flow introspection
  - Mixture Density Network (MDN): Neural Network whose outputs are parameters of a Gaussian Mixture Model (GMM)

$$f^{vo} : \underbrace{\mathbf{x}}_{\text{Optical Flow}} \mapsto \underbrace{\Big(\mu(\mathbf{x}_{t-1,t}), \sigma(\mathbf{x}_{t-1,t}), \pi(\mathbf{x}_{t-1,t})\Big)}_{\mathbf{z}\ \text{Ego-motion density estimate}}$$

  - Conditional-VAE (C-VAE) to reconstruct flow vectors given ego-motion

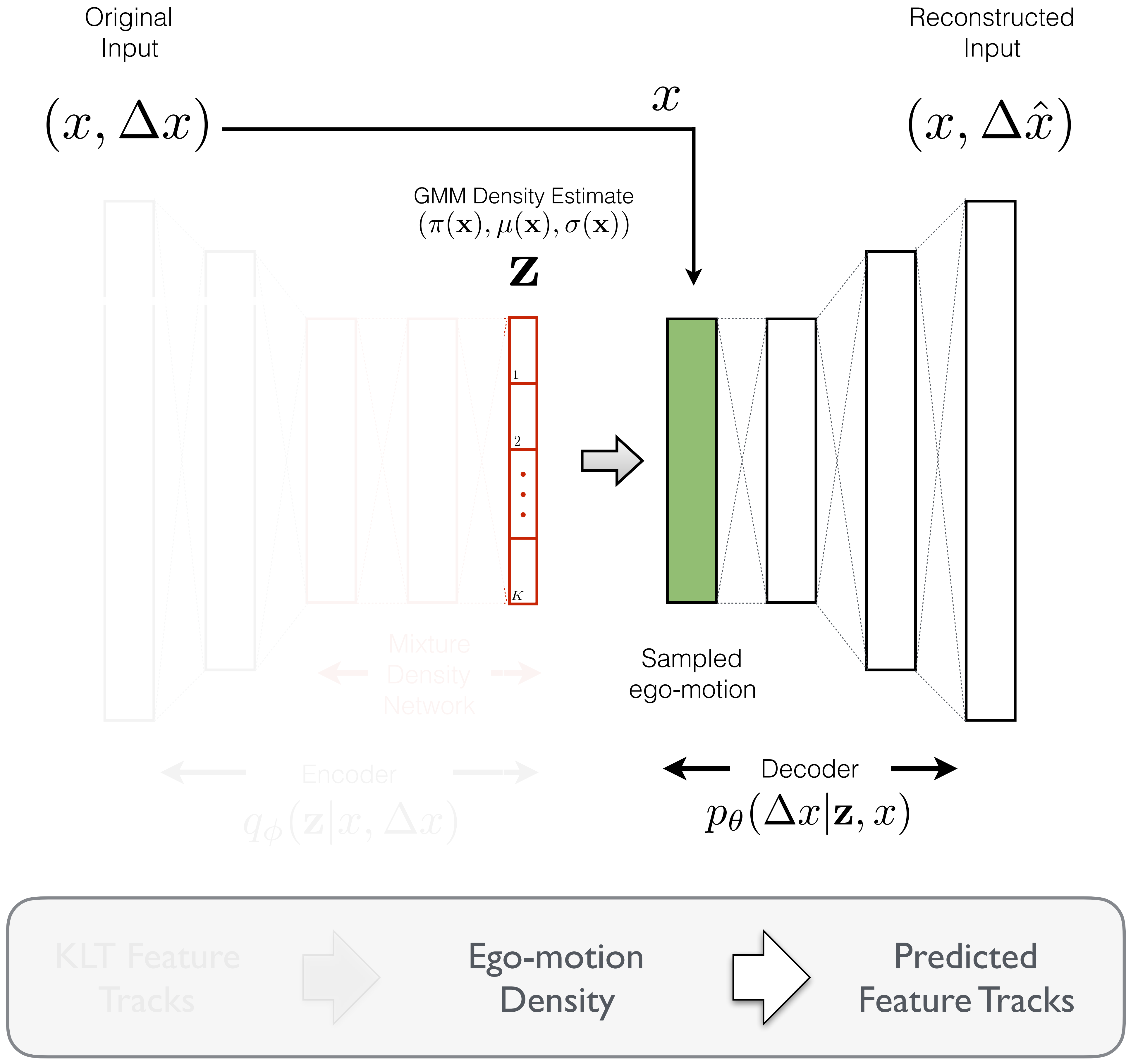
C-VAE

$$\mathcal{L}_{\text{CVAE}} = \underbrace{\mathbb{E}\big[\log p_\theta(\Delta x|\mathbf{z}, x)\big]}_{\text{Reconstruction Error}} - \underbrace{D_{KL}\big[q_\phi(\mathbf{z}|x, \Delta x)||p_\theta(\mathbf{z}|x)\big]}_{\text{Variational Regularization}}$$

(Variational Lower Bound Objective)

Mixture Density Network (MDN) in a Conditional VAE (C-VAE)

Towards Visual Ego-motion Learning in Robots (IROS '17)



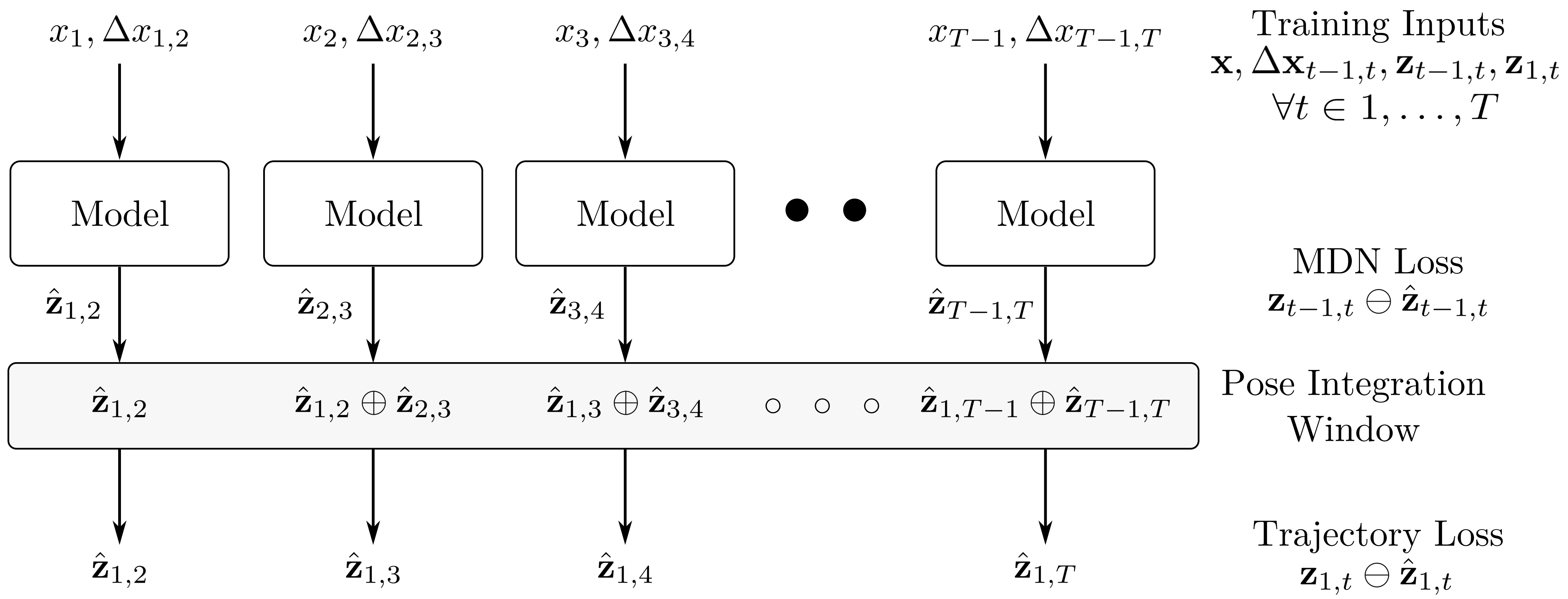
SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Multi-Objective Minimization

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

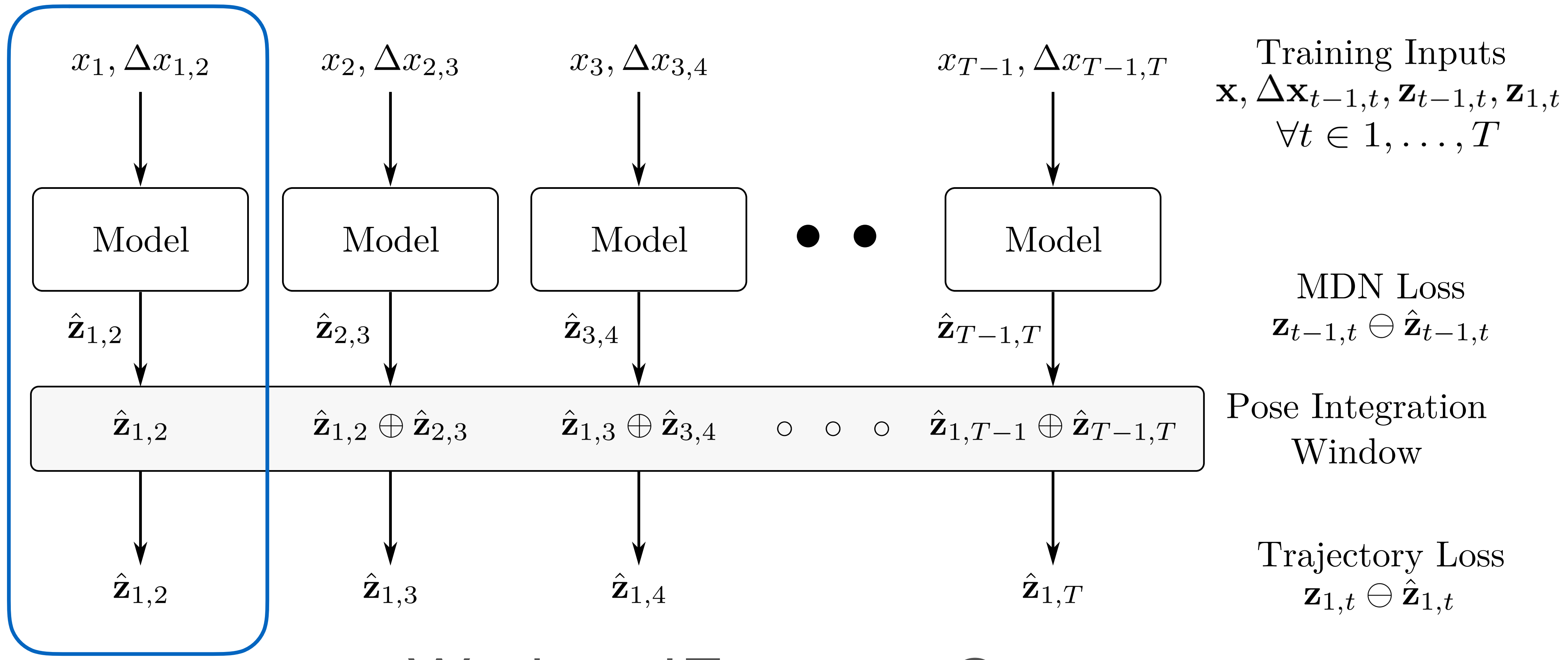
- Multi-Objective Minimization
  - Two-stage optimization



Windowed Trajectory Optimization

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Multi-Objective Minimization
  - Two-stage optimization
  - **Local** MDN loss minimizes short-term ego-motion trajectories, but prone to bias



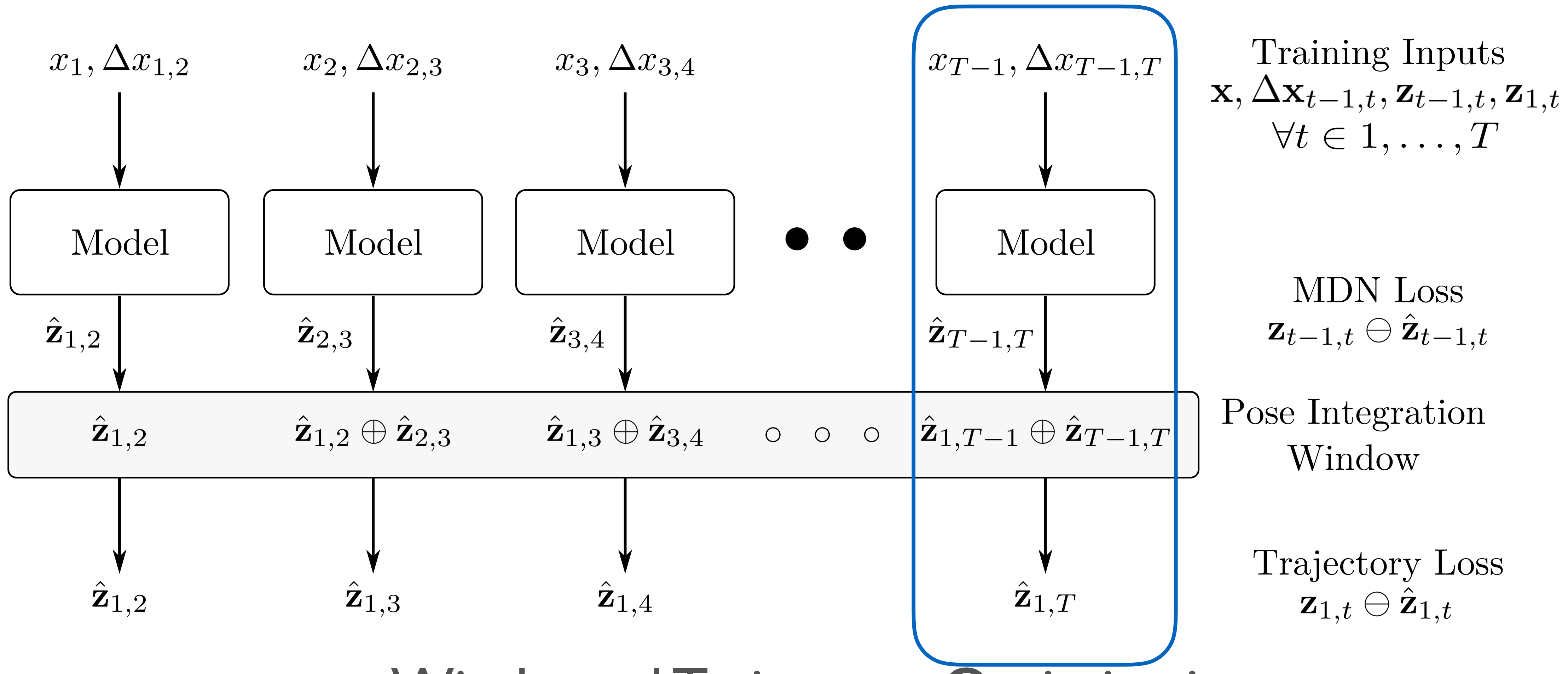
Windowed Trajectory Optimization

Stage 1

$$\mathcal{L}_{\text{ENC}} = \boxed{\underbrace{\sum_t \mathcal{L}^t_{MDN}\left(f^{vo}(\mathbf{x}), \mathbf{z}_{t-1,t}\right)}_{\text{MDN Loss}}}$$

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Multi-Objective Minimization
  - Two-stage optimization
  - **Local** MDN loss minimizes short-term ego-motion trajectories, but prone to bias



Windowed Trajectory Optimization

Stage I

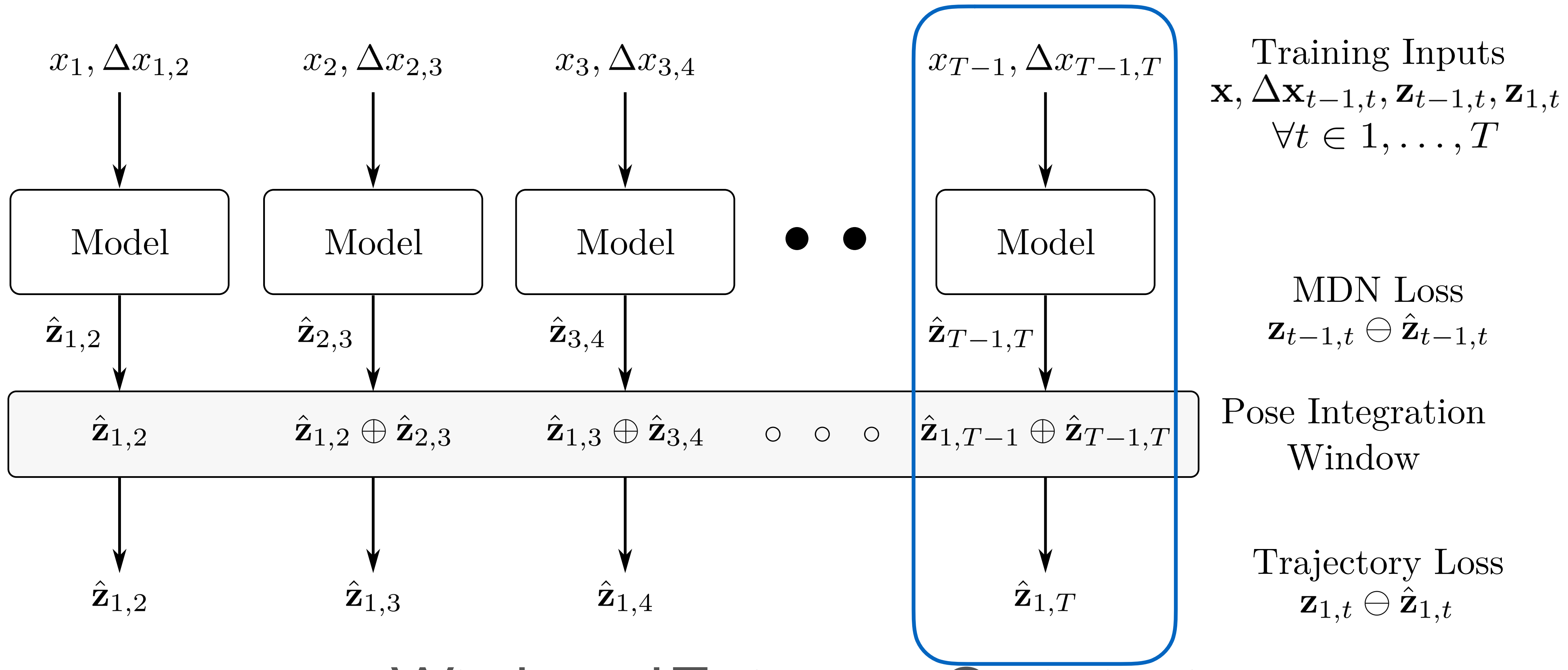
$$\mathcal{L}_{\text{ENC}} = \boxed{\underbrace{\sum_t \mathcal{L}^t_{MDN}\left(f^{vo}(\mathbf{x}), \mathbf{z}_{t-1,t}\right)}_{\text{MDN Loss}}}$$

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

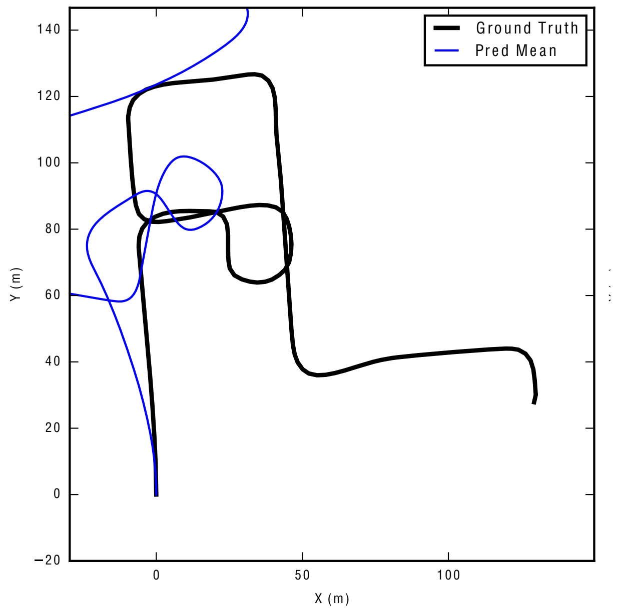
- Multi-Objective Minimization
  - Two-stage optimization
  - **Local** MDN loss minimizes short-term ego-motion trajectories, but prone to bias

Stage 1

$$\mathcal{L}_{\text{ENC}} = \boxed{\underbrace{\sum_t \mathcal{L}^t_{MDN}\left(f^{vo}(\mathbf{x}), \mathbf{z}_{t-1,t}\right)}_{\text{MDN Loss}}}$$



Windowed Trajectory Optimization



Stage 1
(Final)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

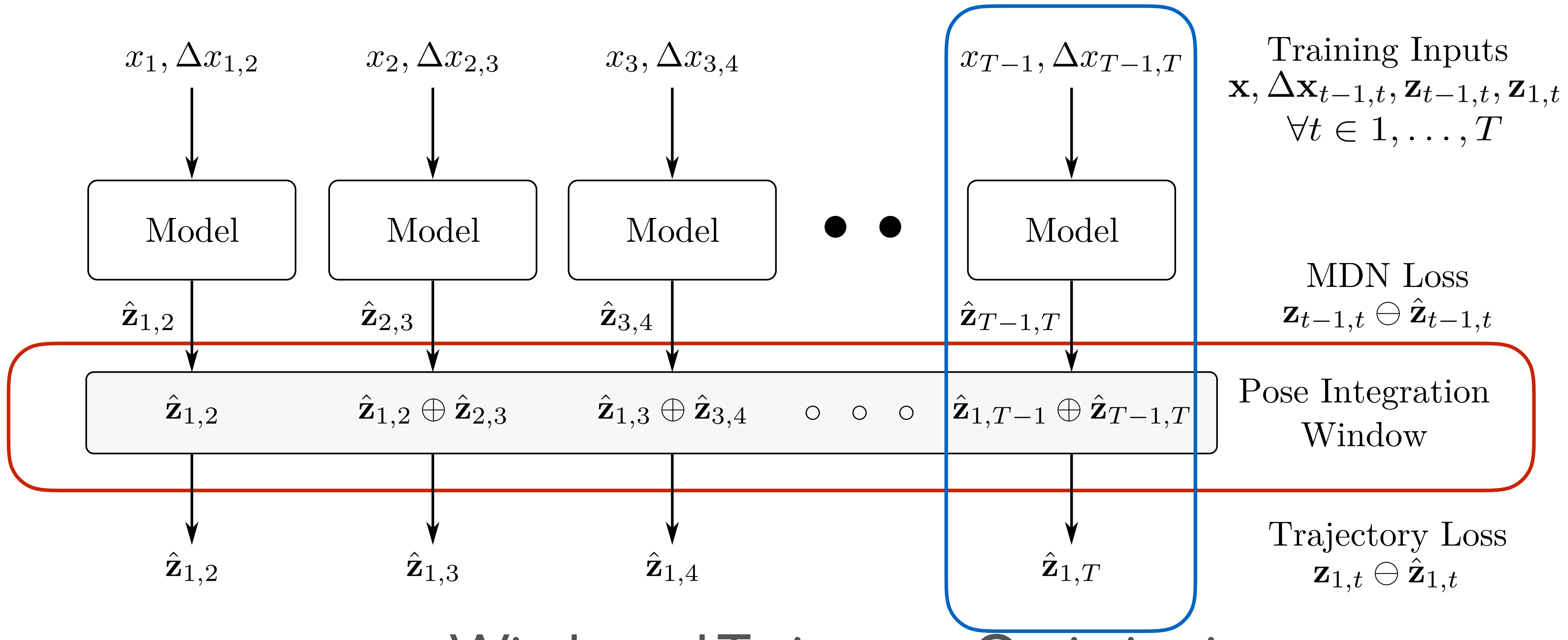
- Multi-Objective Minimization
  - Two-stage optimization
  - **Local** MDN loss minimizes short-term ego-motion trajectories, but prone to bias
  - **Global** Trajectory loss minimizes long-term ego-motion prediction bias

Stage 1

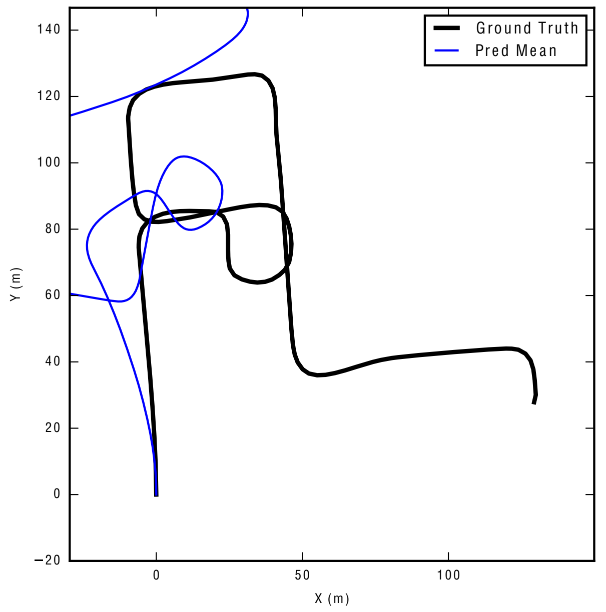
$$\mathcal{L}_{\text{ENC}} = \underbrace{\sum_t \mathcal{L}^t_{MDN}\Big(f^{vo}(\mathbf{x}), \mathbf{z}_{t-1,t}\Big)}_{\text{MDN Loss}}$$

Stage 2

$$\mathcal{L}_{\text{ENC}} = \underbrace{\sum_t \mathcal{L}^t_{MDN}\Big(f^{vo}(\mathbf{x}), \mathbf{z}_{t-1,t}\Big)}_{\text{MDN Loss}} + \underbrace{\sum_t \mathcal{L}^t_{TRAJ}(\mathbf{z}_{1,t} \ominus \hat{\mathbf{z}}_{1,t})}_{\text{Overall Trajectory Loss}}$$



Windowed Trajectory Optimization



Stage 1
(Final)

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

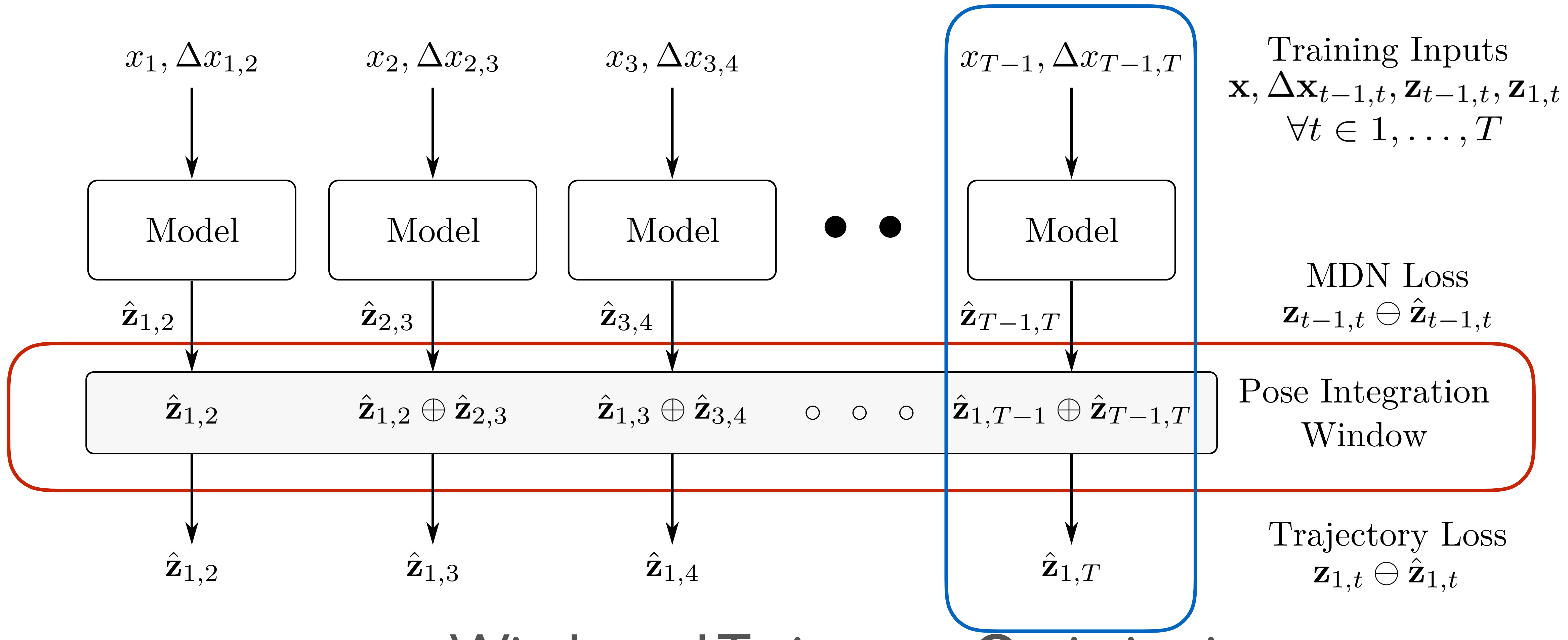
- Multi-Objective Minimization
  - Two-stage optimization
  - **Local** MDN loss minimizes short-term ego-motion trajectories, but prone to bias
  - **Global** Trajectory loss minimizes long-term ego-motion prediction bias

Stage 1

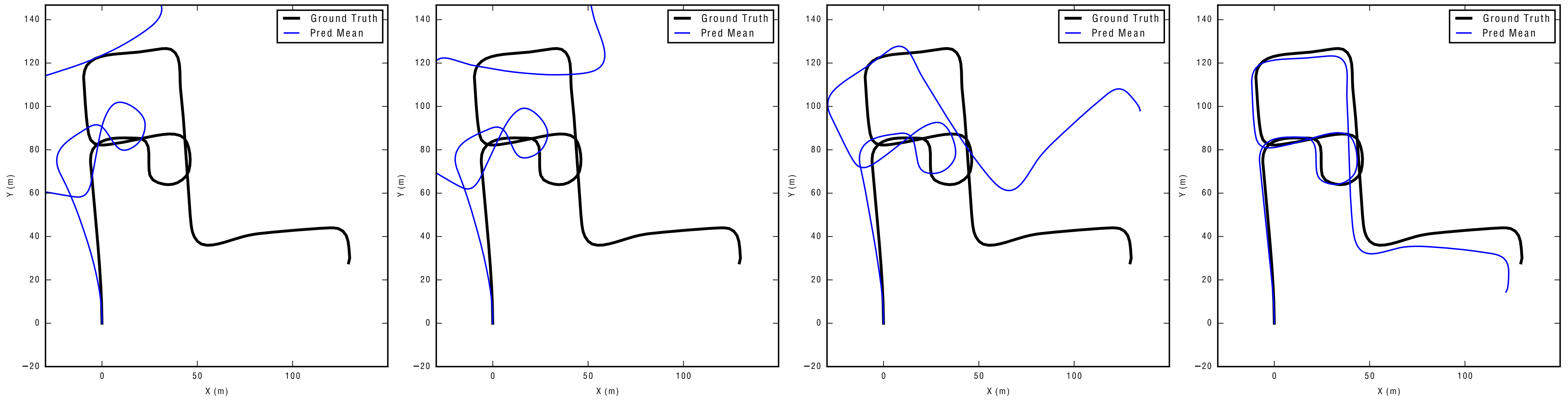
$$\mathcal{L}_{\text{ENC}} = \underbrace{\sum_t \mathcal{L}^t_{MDN}\left(f^{vo}(\mathbf{x}), \mathbf{z}_{t-1,t}\right)}_{\text{MDN Loss}}$$

Stage 2

$$\mathcal{L}_{\text{ENC}} = \underbrace{\sum_t \mathcal{L}^t_{MDN}\left(f^{vo}(\mathbf{x}), \mathbf{z}_{t-1,t}\right)}_{\text{MDN Loss}} + \underbrace{\sum_t \mathcal{L}^t_{TRAJ}(\mathbf{z}_{1,t} \ominus \hat{\mathbf{z}}_{1,t})}_{\text{Overall Trajectory Loss}}$$



Windowed Trajectory Optimization

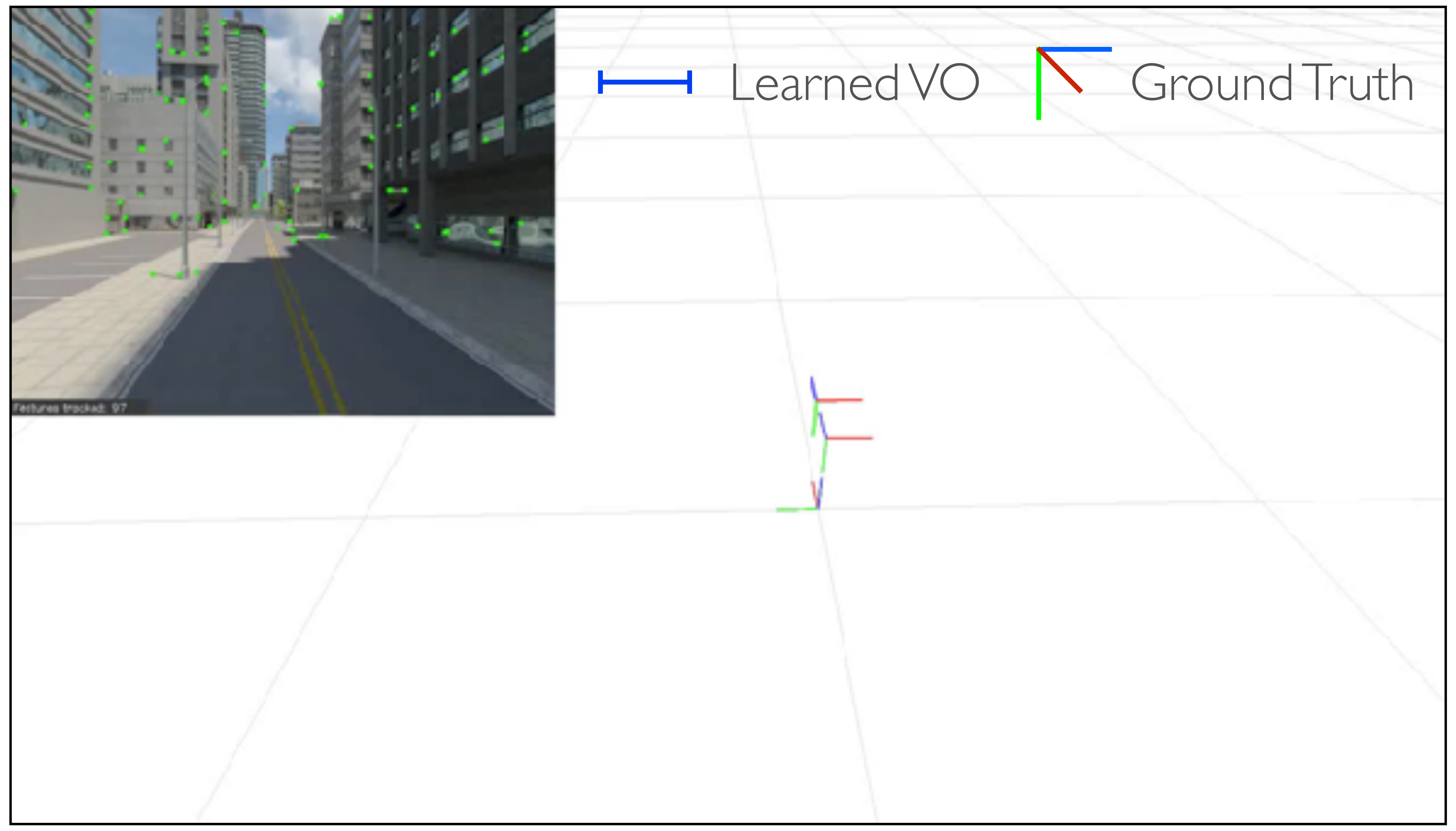


Stage 1 (Final) | **Stage 2** (Epoch 4) | **Stage 2** (Epoch 8) | **Stage 2** (Epoch 18)

Two-stage Optimization

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

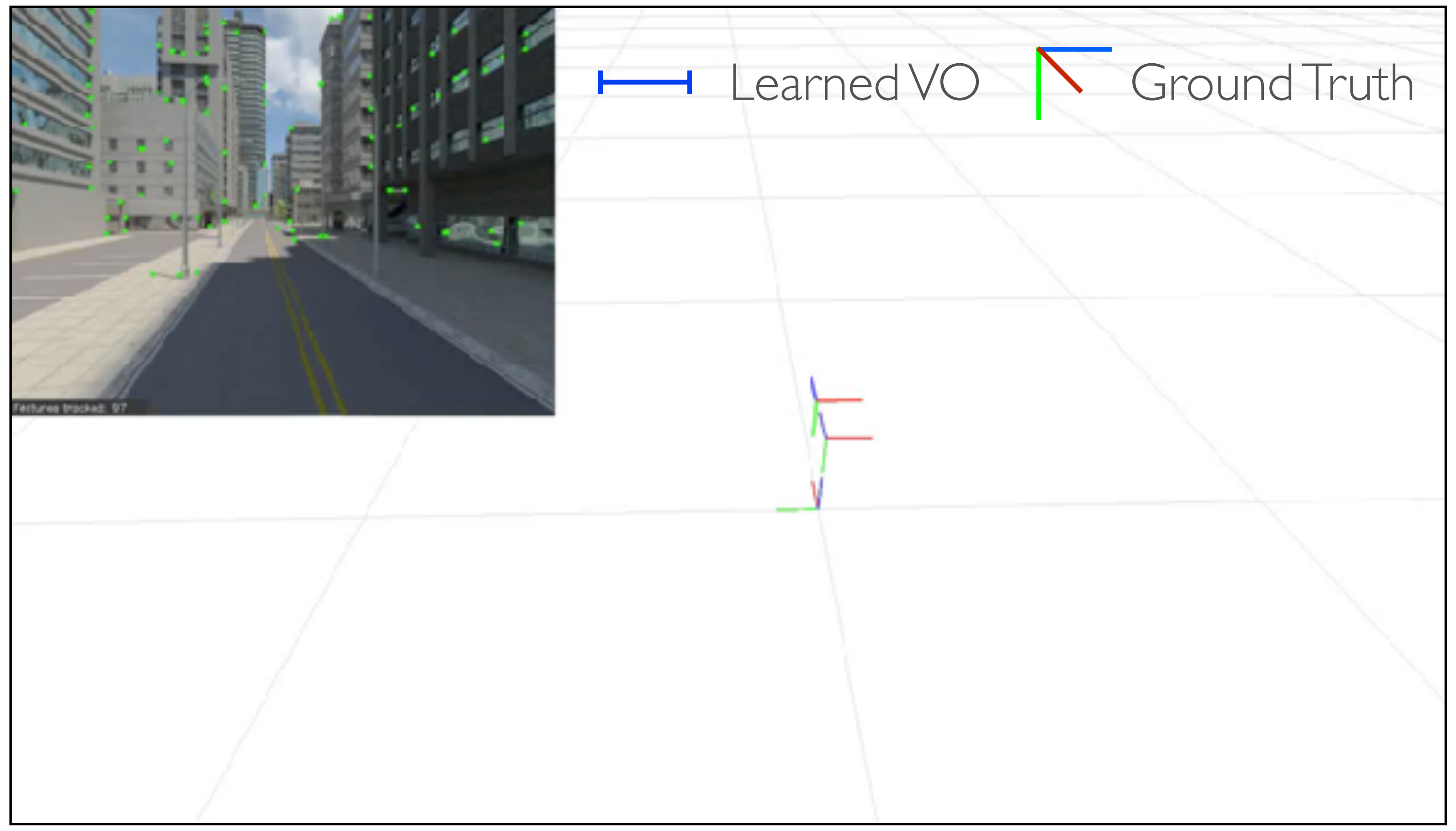
- Learning to recover ego-motion from feature tracks
  - Robust and adaptive (Tunable architectural capacity)
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)
  - Powerful model based reasoning (Scene flow introspection)



Learned VO

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

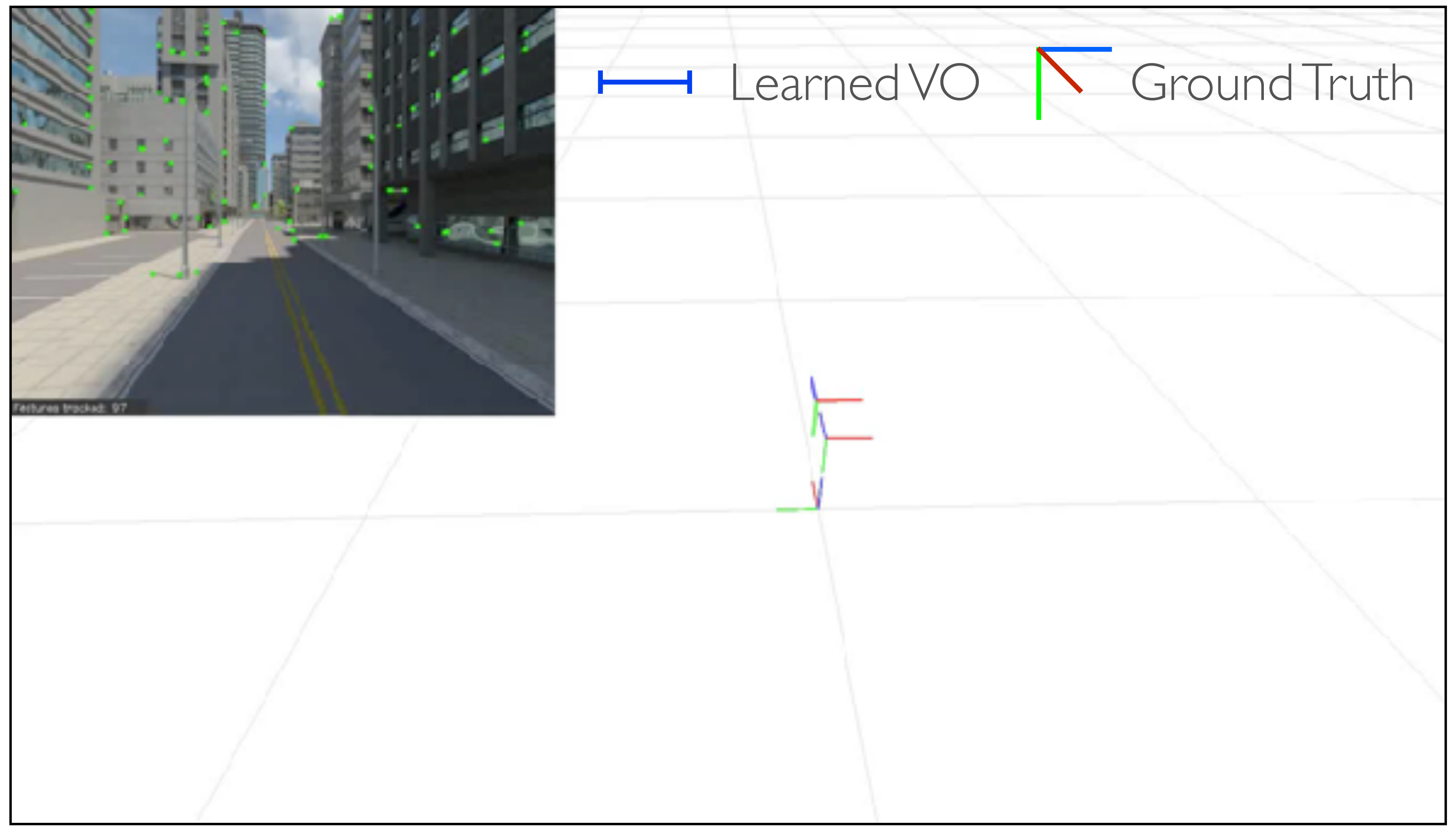
- Learning to recover ego-motion from feature tracks
  - Robust and adaptive (Tunable architectural capacity)
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)
  - Powerful model based reasoning (Scene flow introspection)



Learned VO

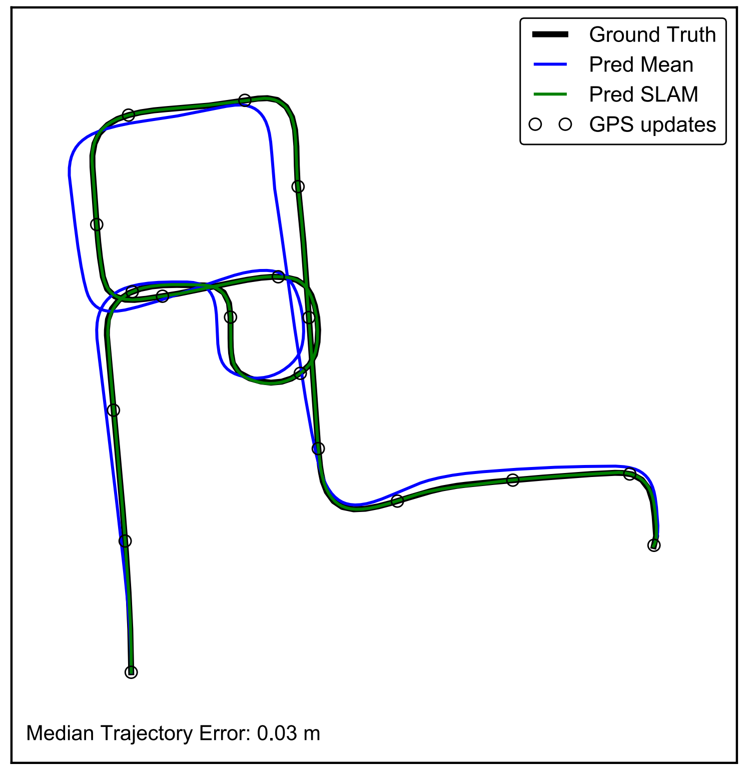
SELF-SUPERVISED
VISUAL EGO-MOTION LEARNING

- Learning to recover ego-motion from feature tracks
  - Robust and adaptive (Tunable architectural capacity)
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)
  - Powerful model based reasoning (Scene flow introspection)

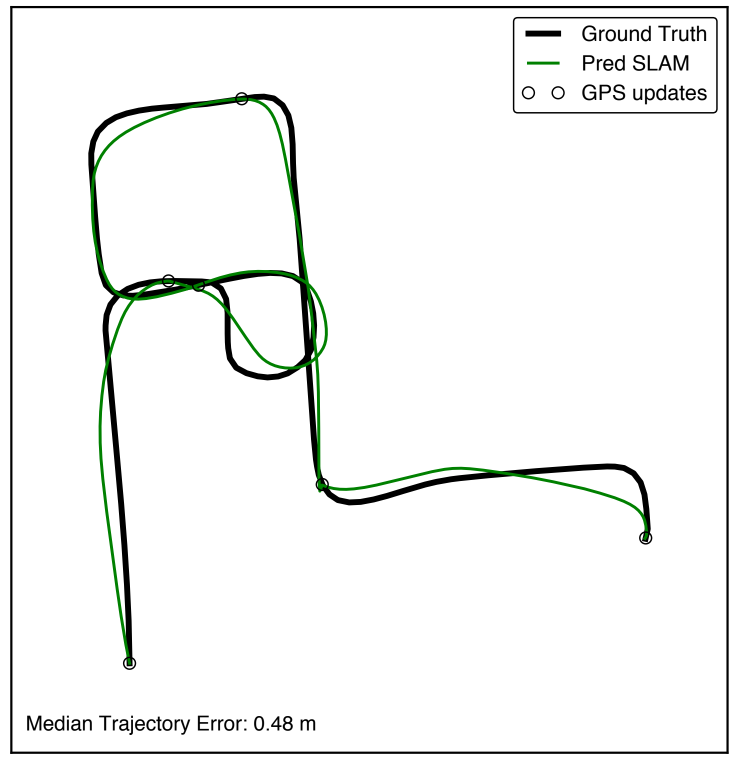


Learned VO

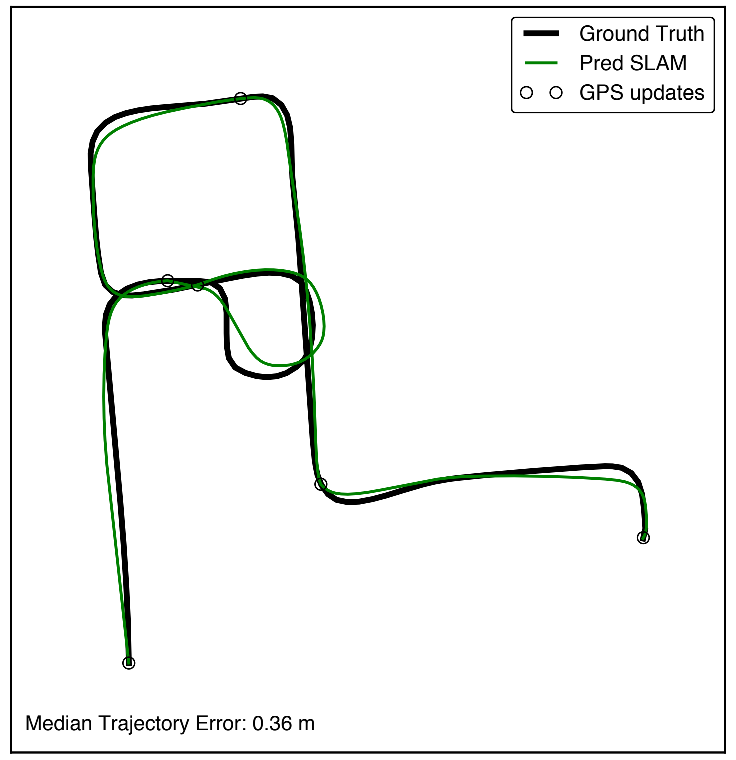
Trajectory Estimation and Optimization
(Learned VO + intermittent GPS updates)



Pinhole



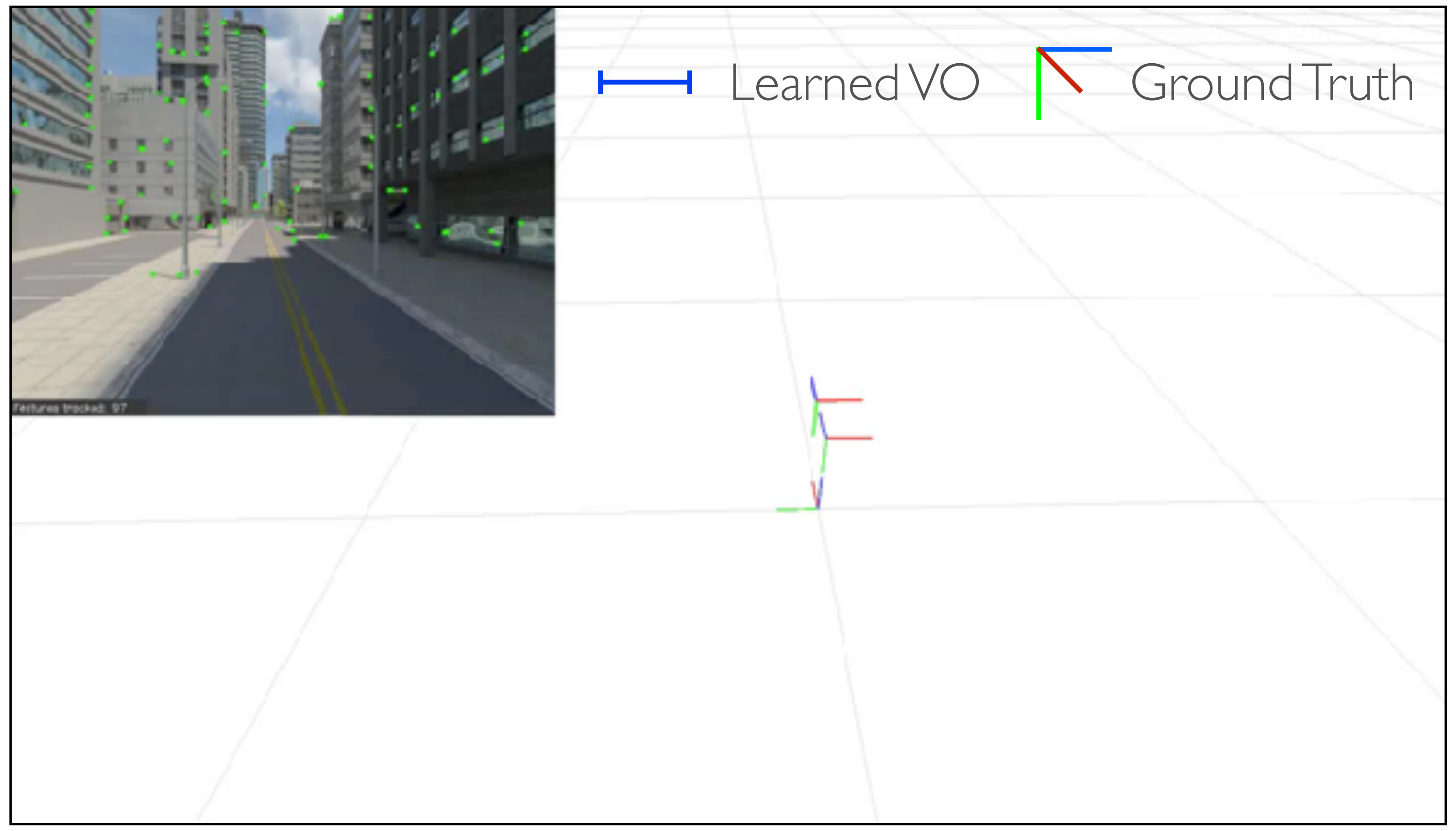
Fisheye



Catadioptric

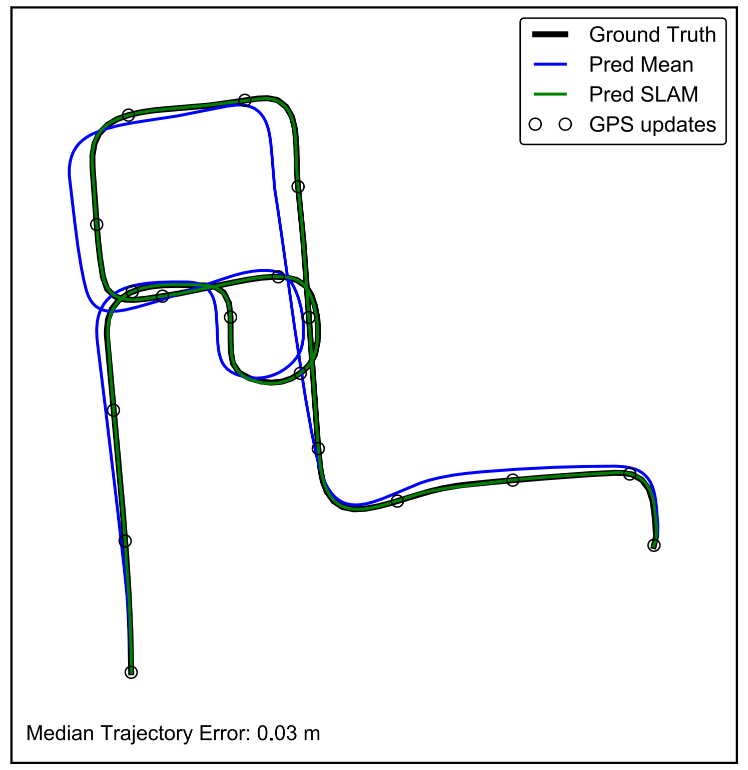
SELF-SUPERVISED VISUAL EGO-MOTION LEARNING

- Learning to recover ego-motion from feature tracks
  - Robust and adaptive (Tunable architectural capacity)
  - Generic camera optics (Pinhole, Fisheye, Catadioptric)
  - Powerful model based reasoning (Scene flow introspection)

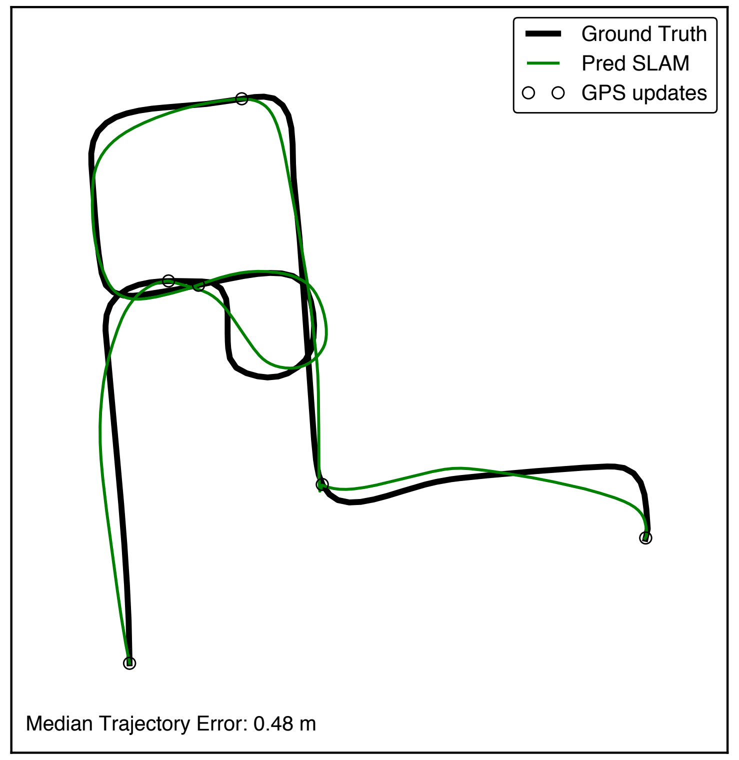


Learned VO

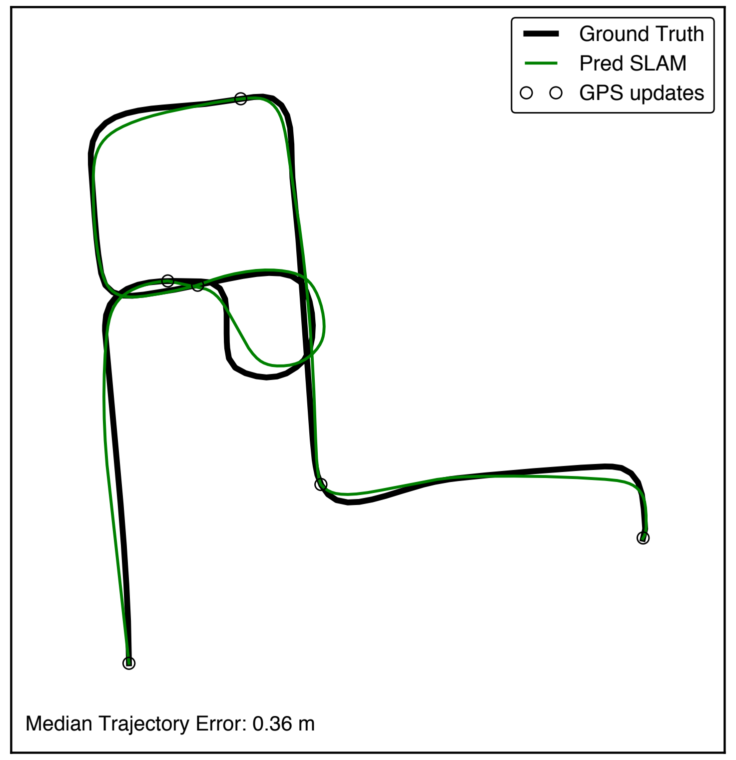
Trajectory Estimation and Optimization
(Learned VO + intermittent GPS updates)



Pinhole



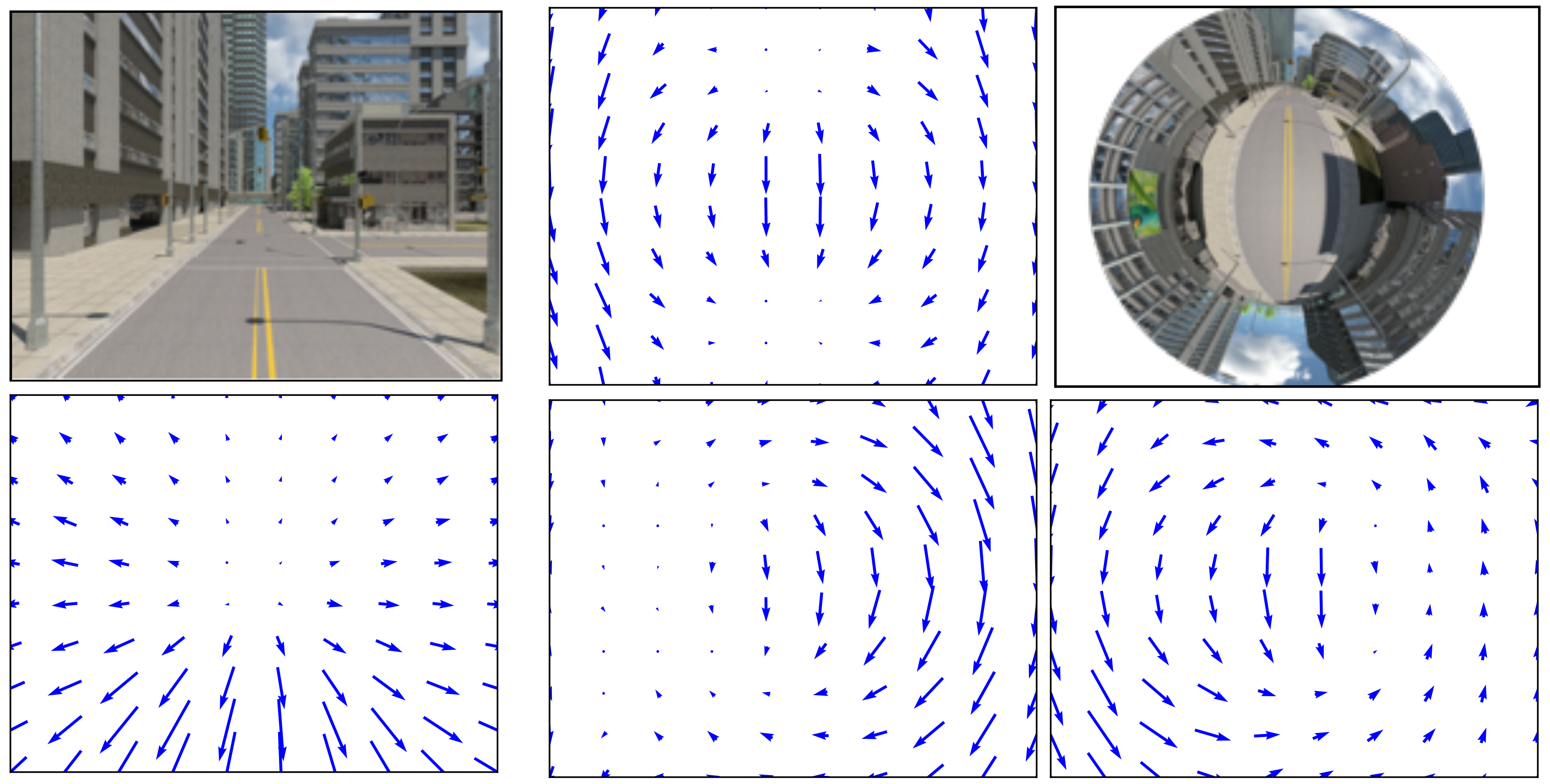
Fisheye



Catadioptric

Generative Modeling
(Scene flow introspection)

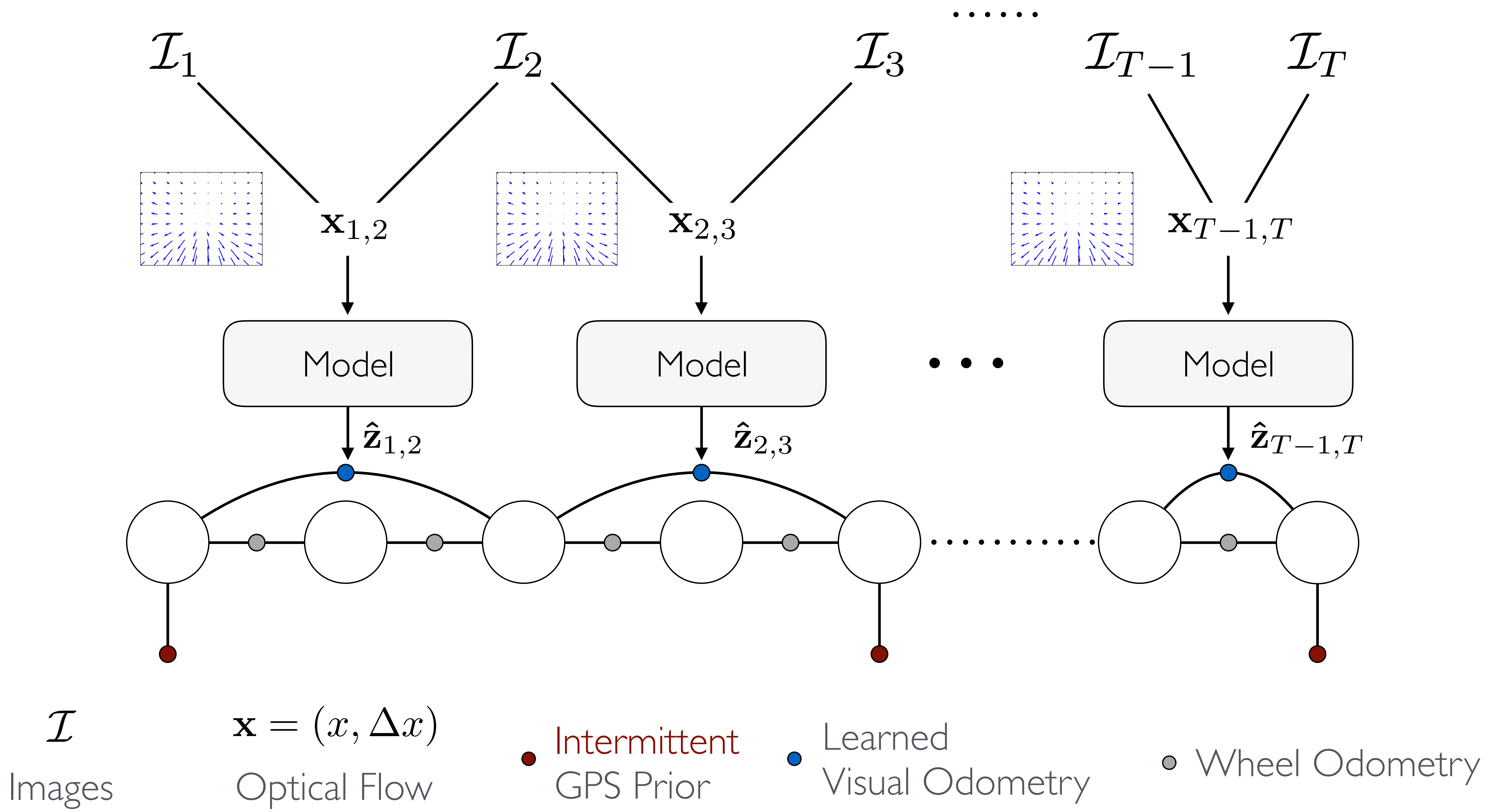
Forward



Forward | Left | Right

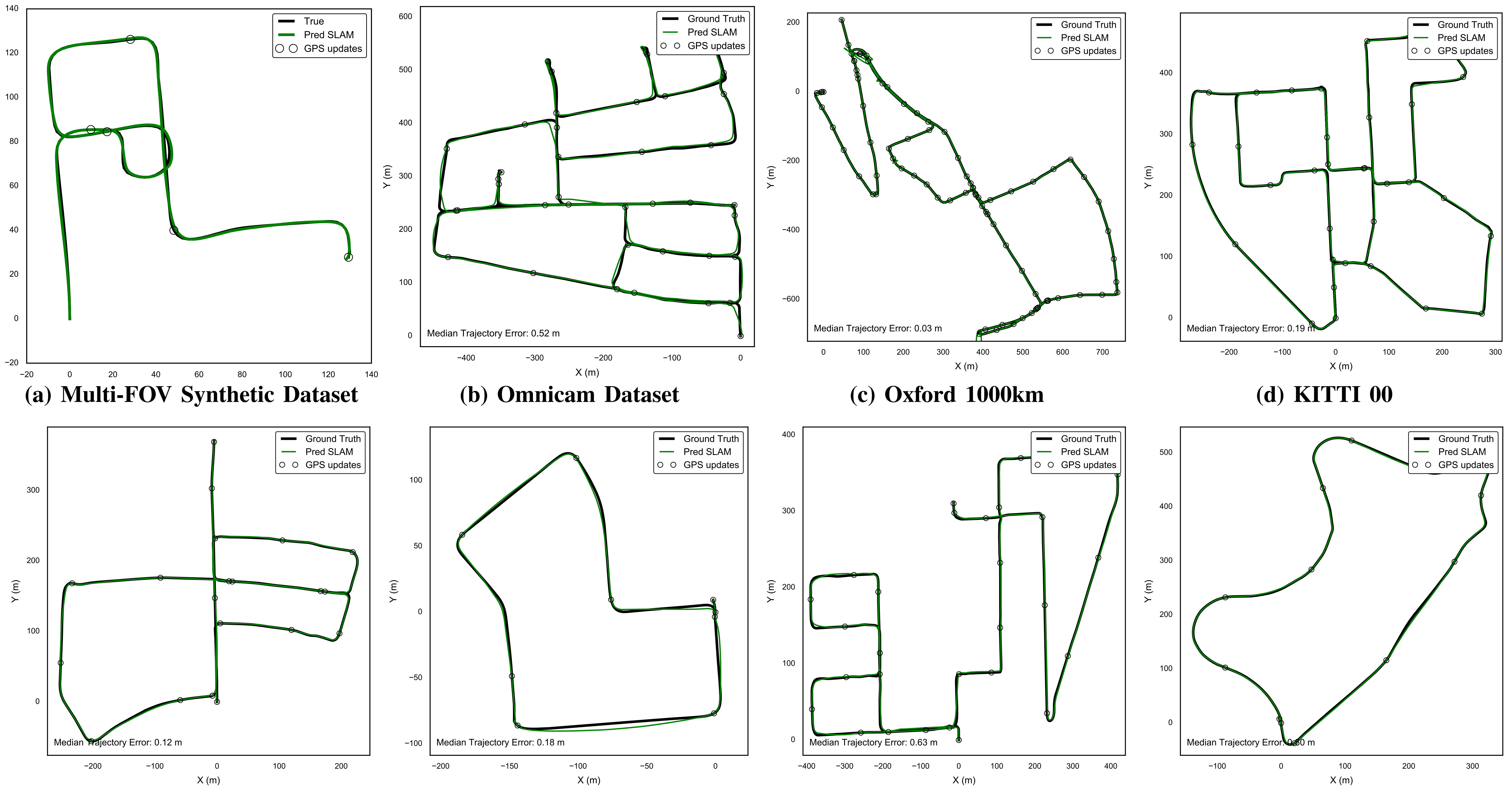
Pinhole | **Catadioptric**

SELF-SUPERVISED VISUAL EGO-MOTION LEARNING



DEPLOYMENT

VISUAL EGO-MOTION PERFORMANCE



(a) Multi-FOV Synthetic Dataset (b) Omnicam Dataset (c) Oxford 1000km (d) KITTI 00

(e) KITTI 05 (f) KITTI 07 (g) KITTI 08 (h) KITTI 09

Sensor fusion with learned ego-motion on various datasets

Fusing our learned visual ego-motion with intermittent GPS updates

(Datasets: Multi-FOV Synthetic Dataset, Oxford 1000km, KITTI)

VISUAL EGO-MOTION PERFORMANCE

Dataset	Camera	Median Trajectory Error
KITTI	Pinhole	0.02 - 0.63 m
Multi-FOV	Pinhole	0.18 m
Multi-FOV	Fisheye	0.48 m
Multi-FOV	Catadioptric	0.36 m
Oxford	Pinhole	0.03 m
KITTI-Omni	Catadioptric	0.52 m

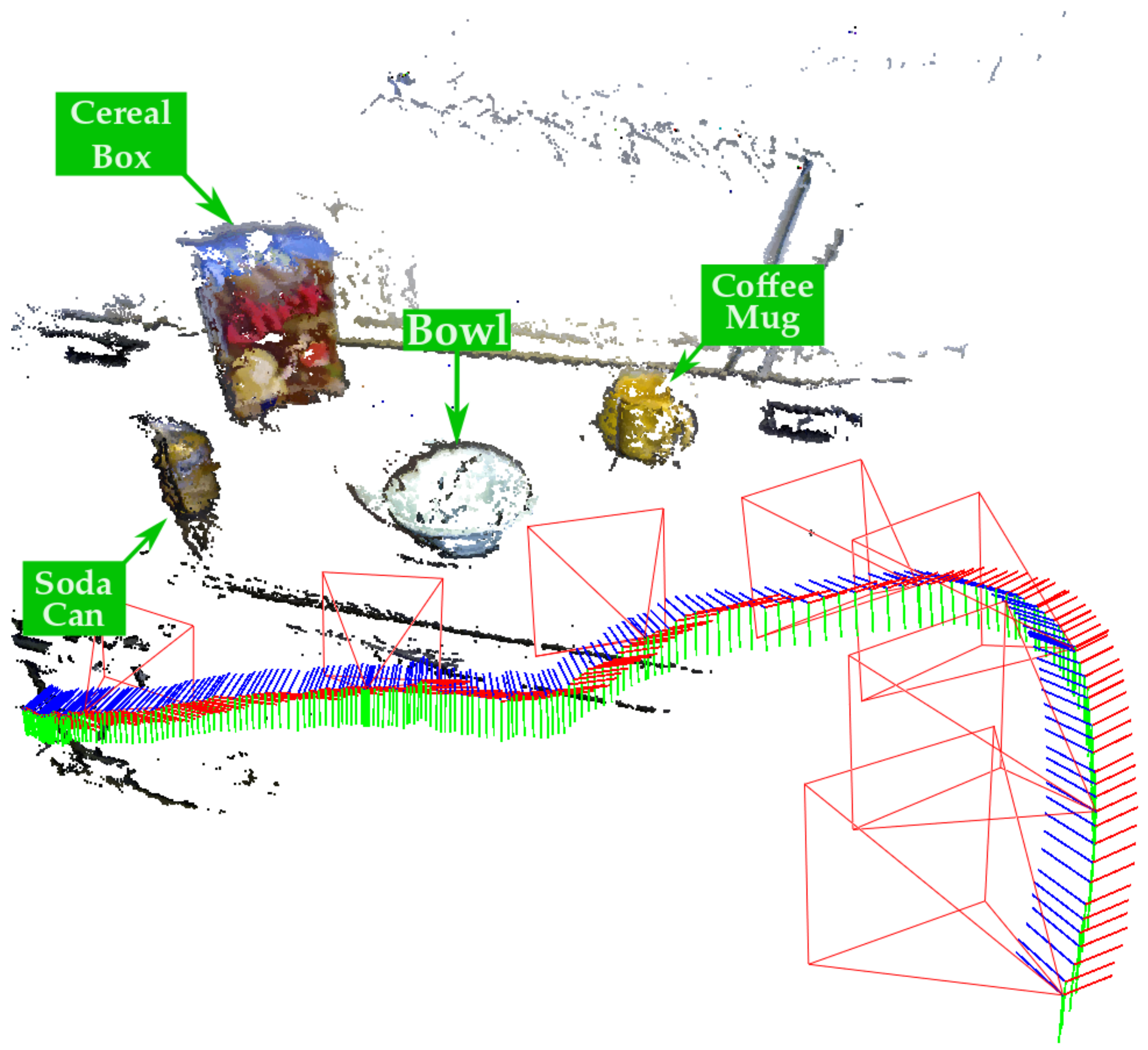
Trajectory Prediction Performance

Fusing our learned visual ego-motion with intermittent GPS updates

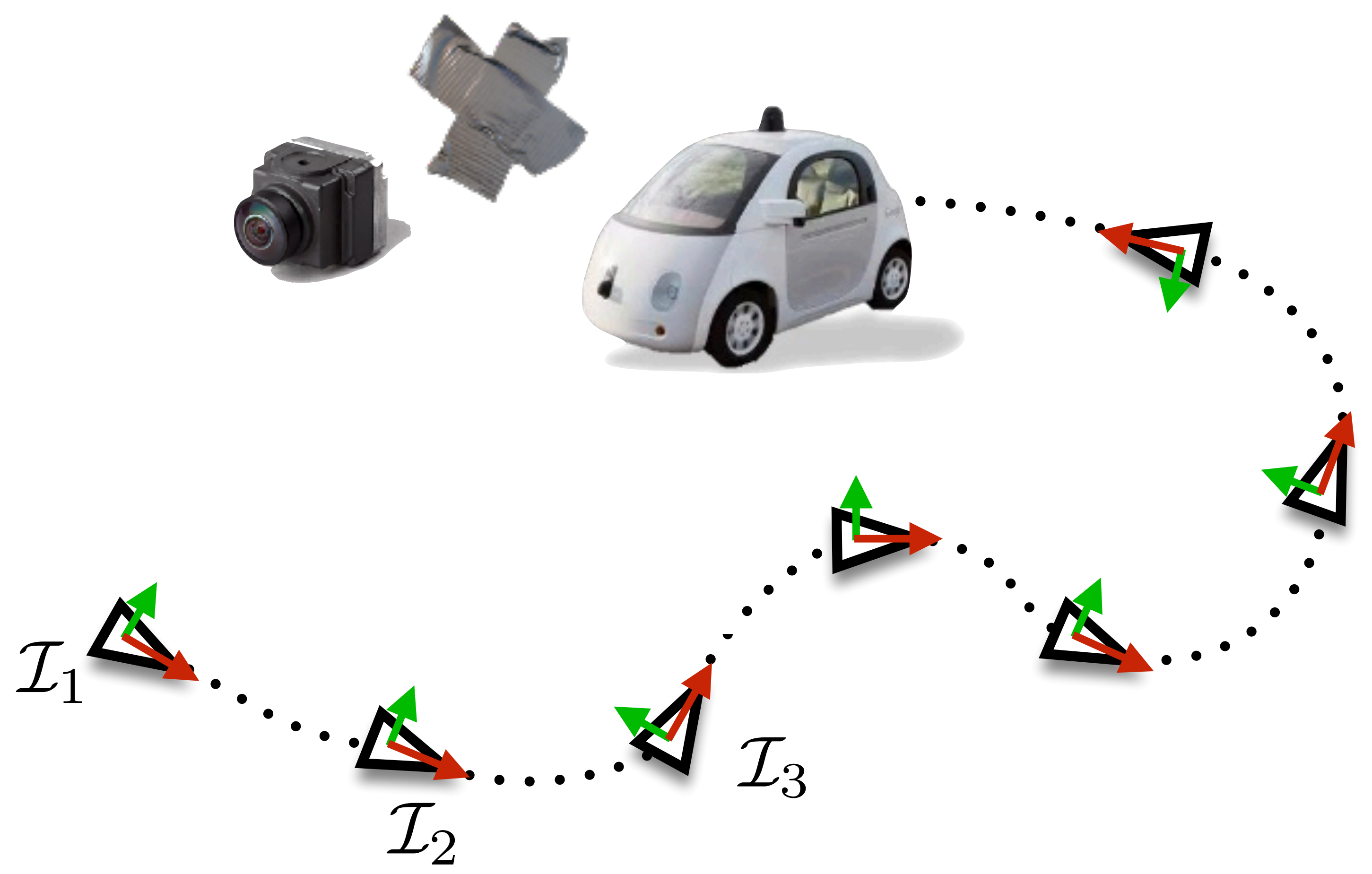
(Datasets: Multi-FOV Synthetic Dataset, Oxford 1000km, KITTI)

SLAM AS A SUPERVISORY SIGNAL

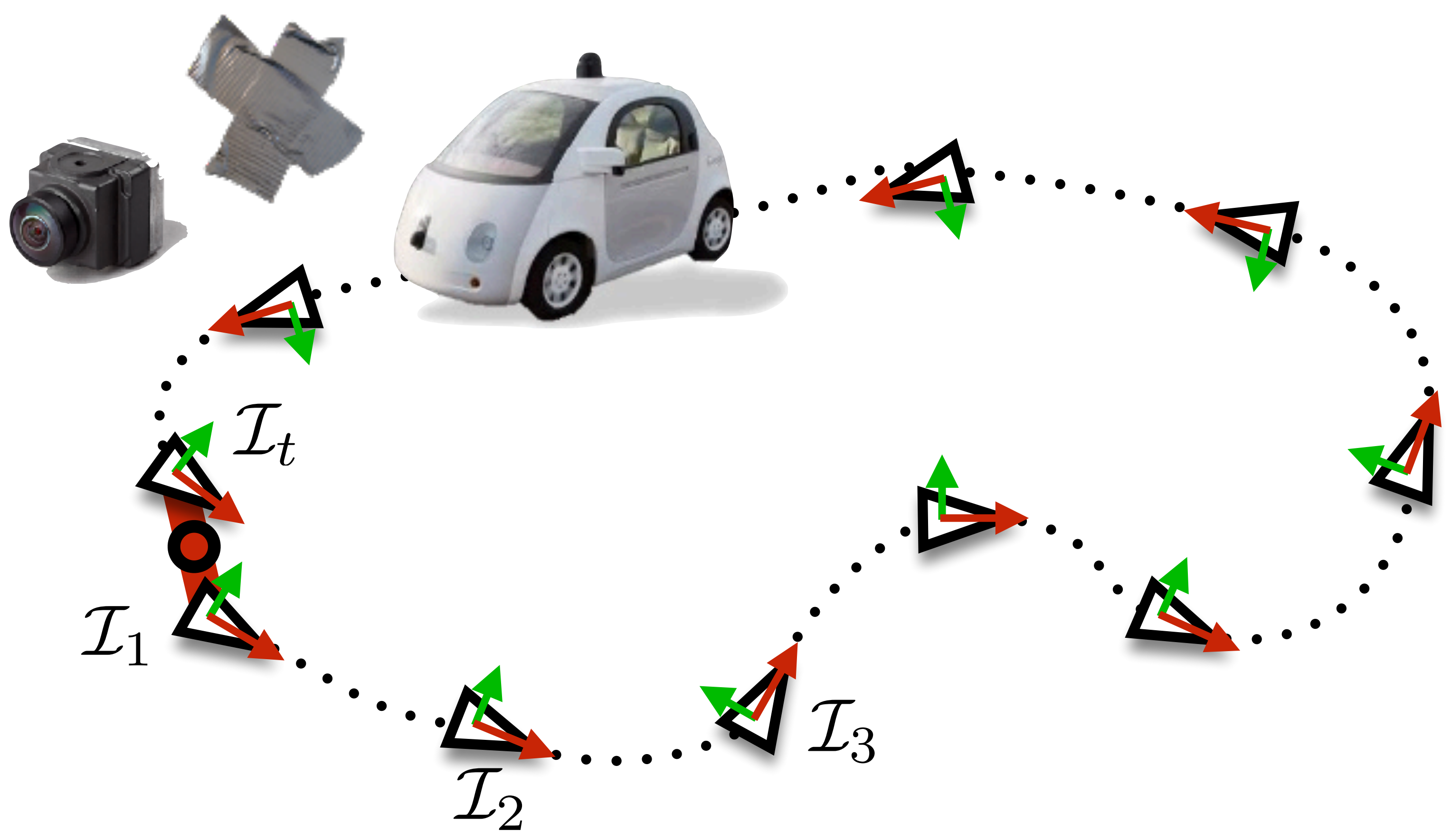
Monocular SLAM-Supported Object Recognition



Self-Supervised Visual Ego-motion Learning



Self-Supervised Visual Place Recognition Learning



SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

Correspondence Engine
(Geometric data association)

Self-Supervision
(SLAM-aided supervision)

Knowledge Transfer
(Bootstrapping)

SLAM AS A SUPERVISORY SIGNAL

Monocular SLAM-Supported Object Recognition

Correspondence Engine

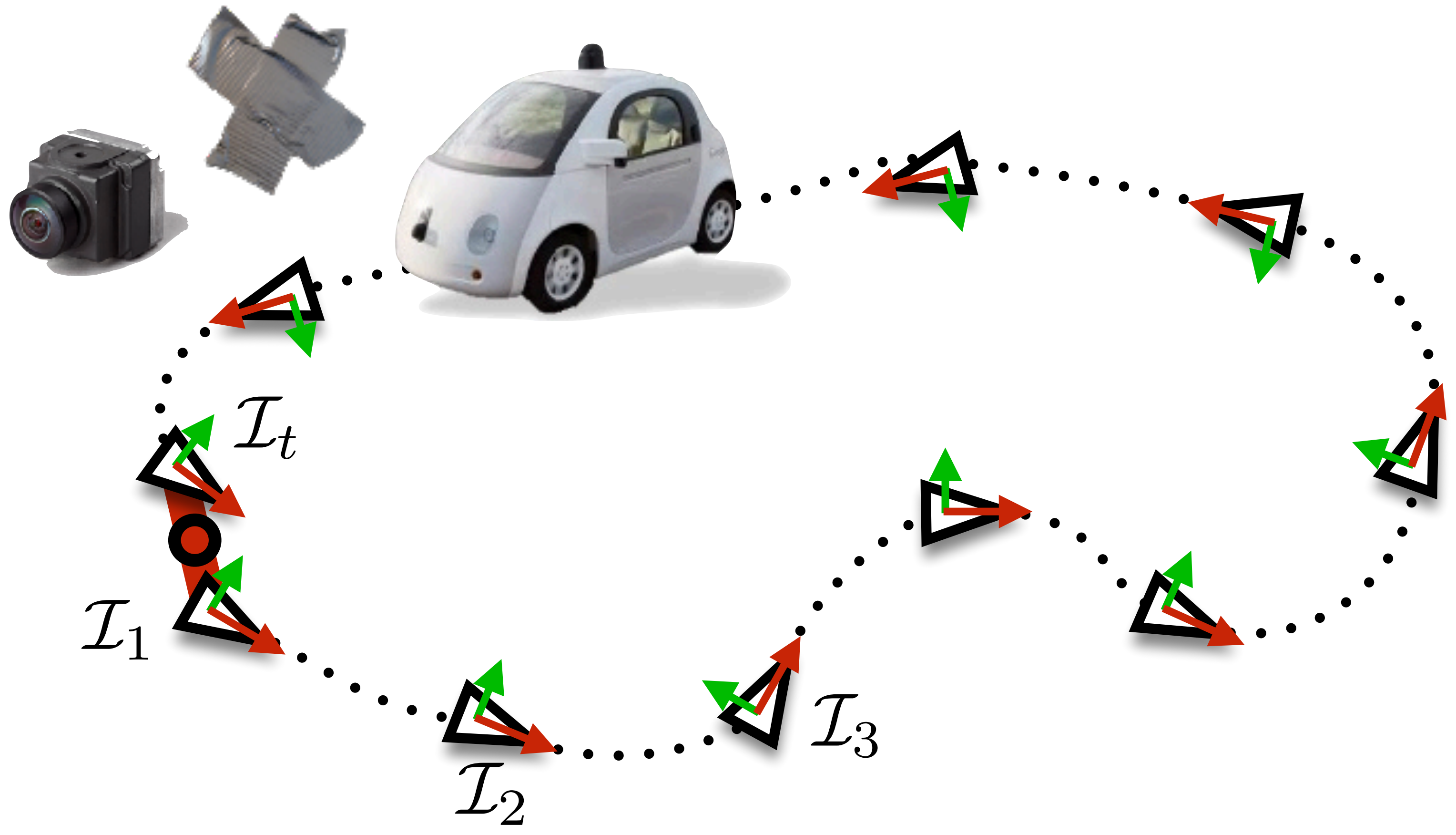
(Geometric data association)

Self-Supervised Visual Ego-motion Learning

Self-Supervision

(SLAM-aided supervision)

Self-Supervised Visual Place Recognition Learning



Knowledge Transfer

(Bootstrapping)

SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

VISION-BASED LOOP-CLOSURE DETECTION

- **Visual Place-Recognition / Loop-Closure Detection**
  - Identifying previously visited places to reduce the odometry drift

VISION-BASED LOOP-CLOSURE DETECTION

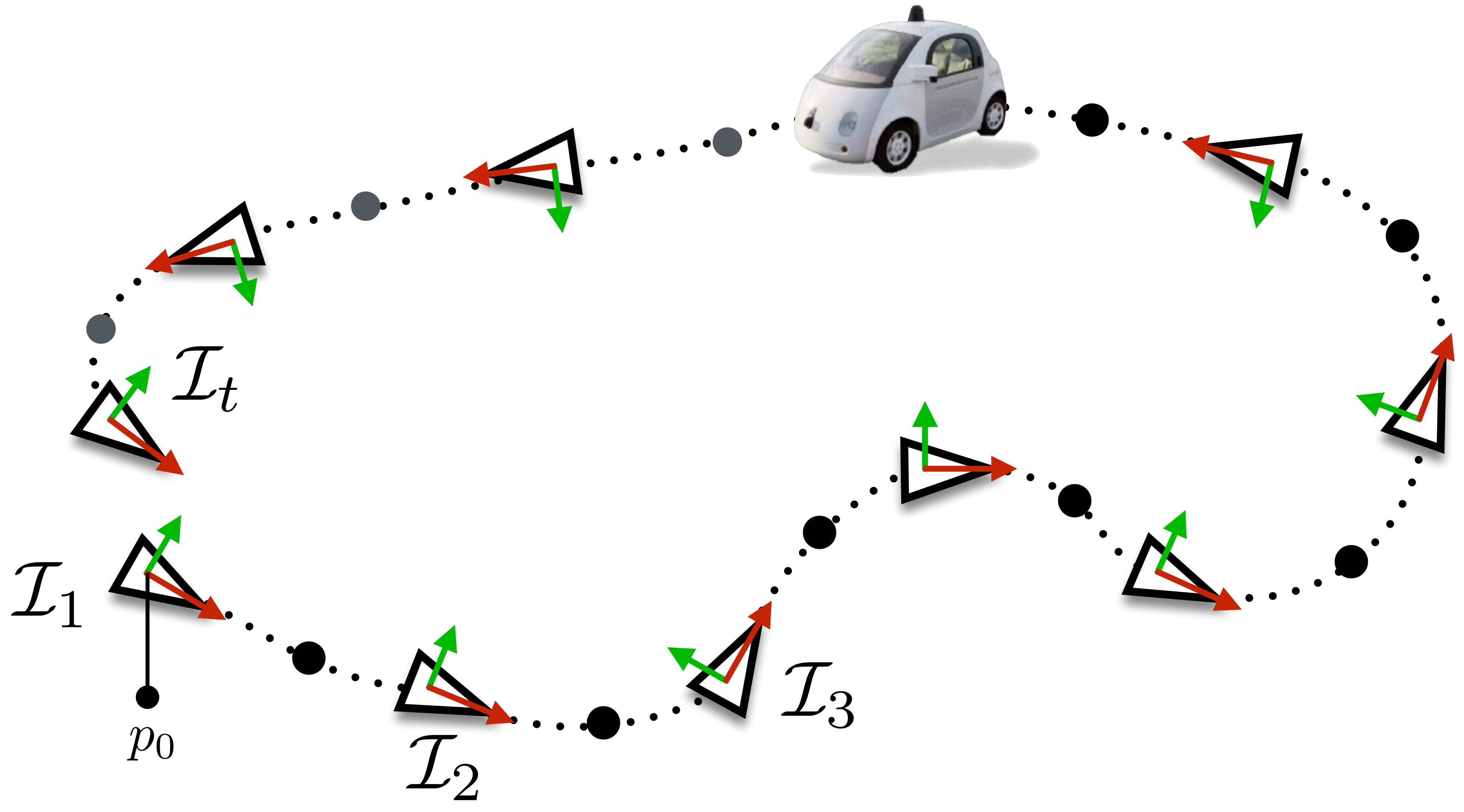
- **Visual Place-Recognition / Loop-Closure Detection**
  - Identifying previously visited places to reduce the odometry drift

VISION-BASED LOOP-CLOSURE DETECTION

- **Visual Place-Recognition / Loop-Closure Detection**
  - Identifying previously visited places to reduce the odometry drift

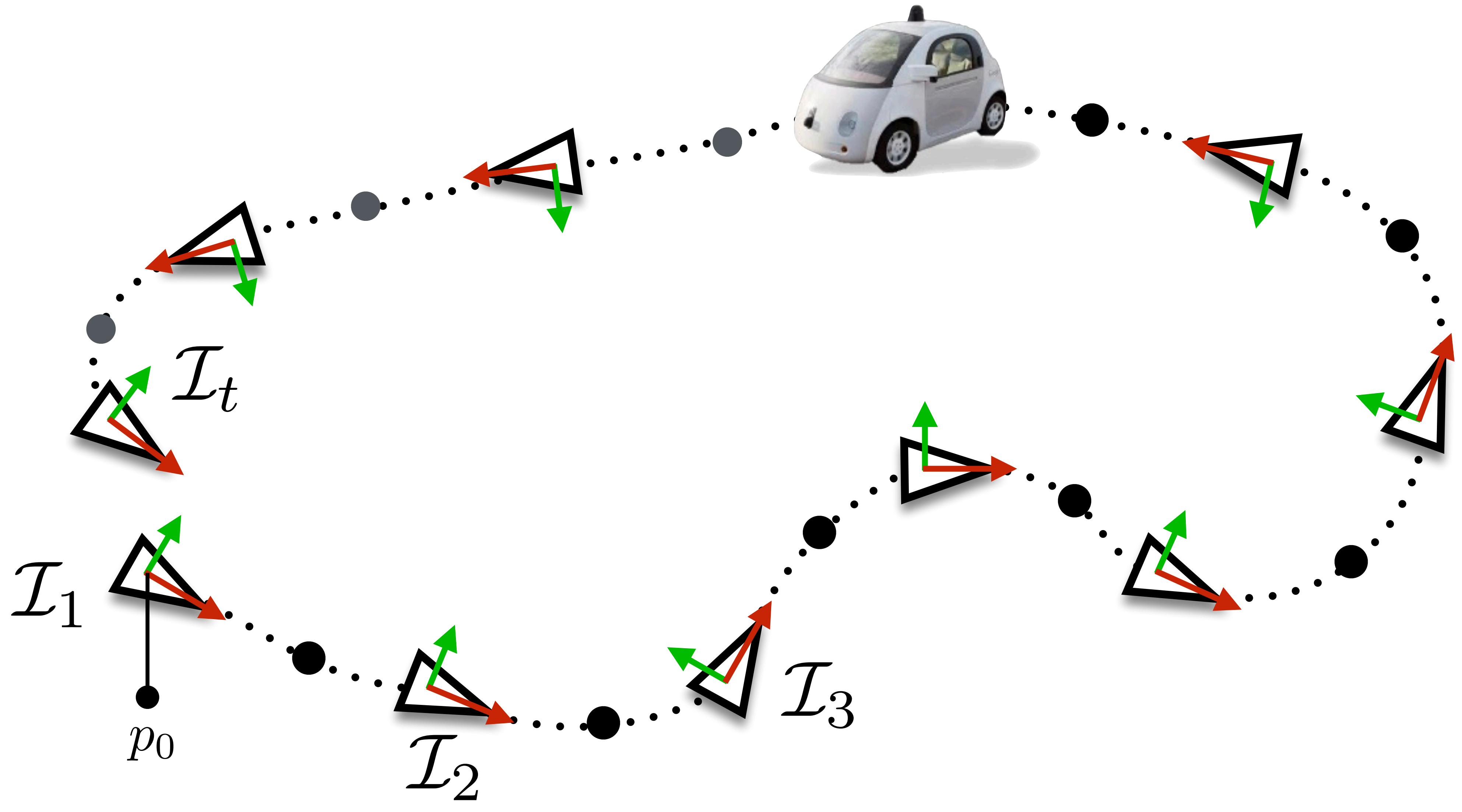
VISION-BASED LOOP-CLOSURE DETECTION

- **Visual Place-Recognition / Loop-Closure Detection**
  - Identifying previously visited places to reduce the odometry drift



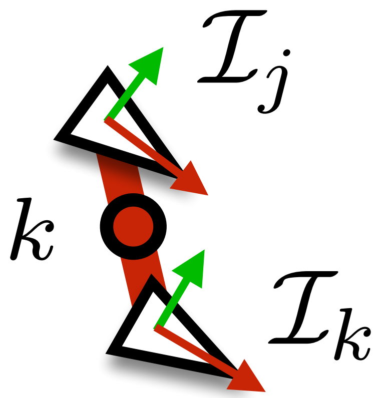
VISION-BASED LOOP-CLOSURE DETECTION

- **Visual Place-Recognition / Loop-Closure Detection**
  - Identifying previously visited places to reduce the odometry drift



DETERMINE f such that

$$f(\mathcal{I}_j) \simeq f(\mathcal{I}_k) \quad \mapsto \quad c_{j,k}$$

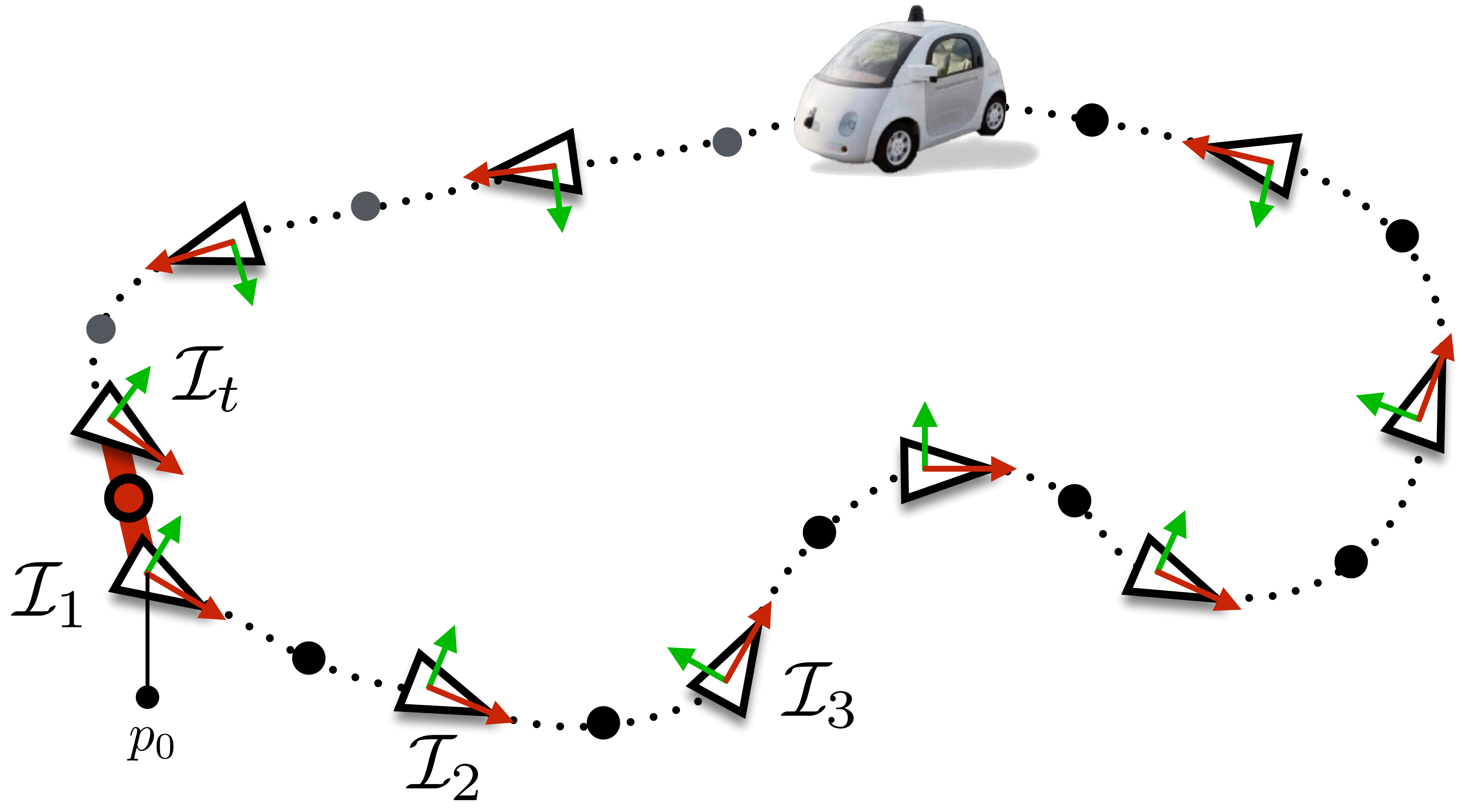


Temporally-distant Images

Loop-closure Constraint

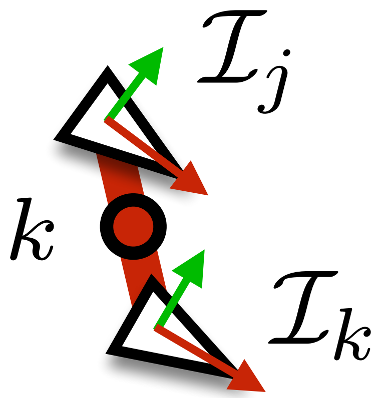
VISION-BASED LOOP-CLOSURE DETECTION

- **Visual Place-Recognition / Loop-Closure Detection**
  - Identifying previously visited places to reduce the odometry drift



DETERMINE f such that

$$f(\mathcal{I}_j) \simeq f(\mathcal{I}_k) \mapsto c_{j,k}$$



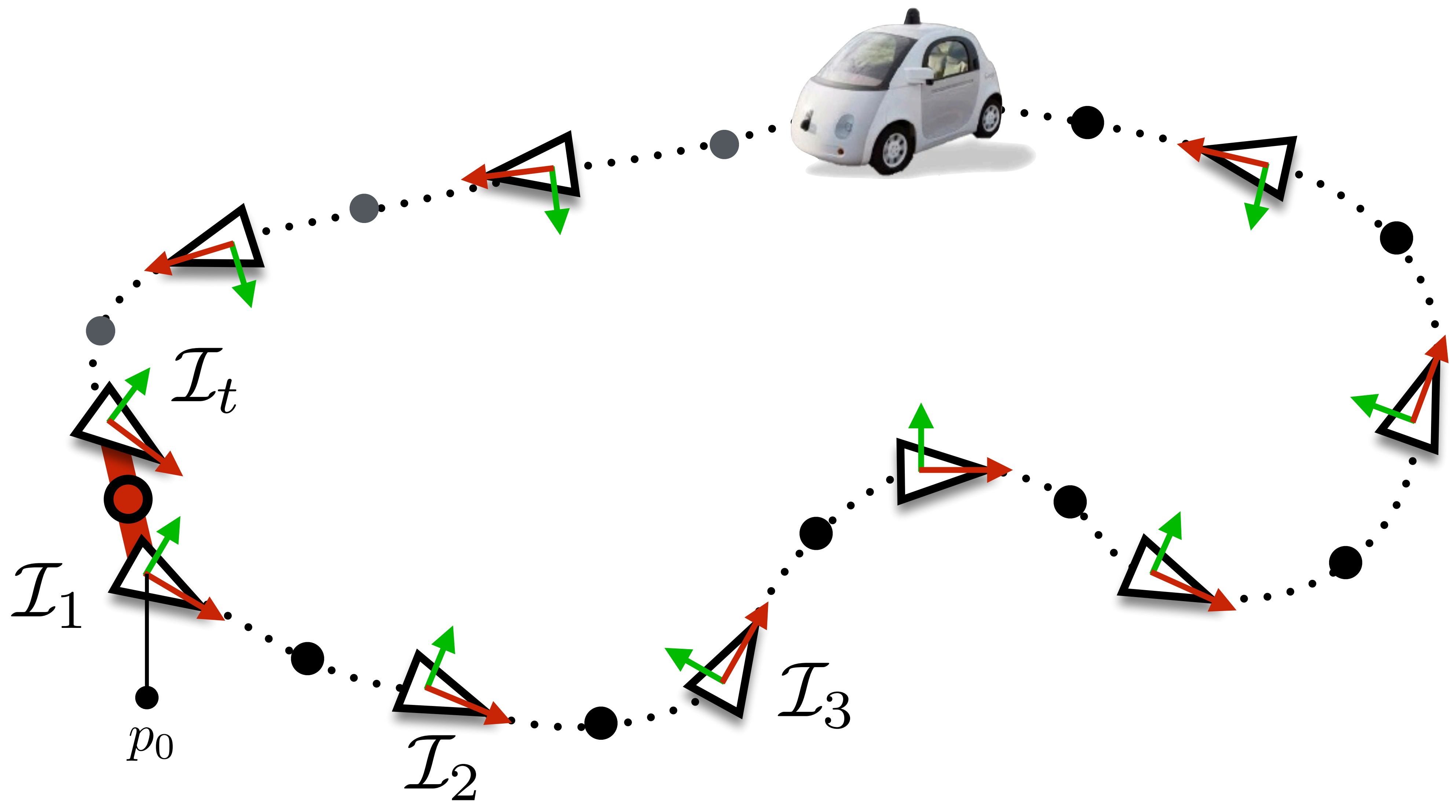
Temporally-distant Images

Loop-closure Constraint

VISION-BASED LOOP-CLOSURE DETECTION

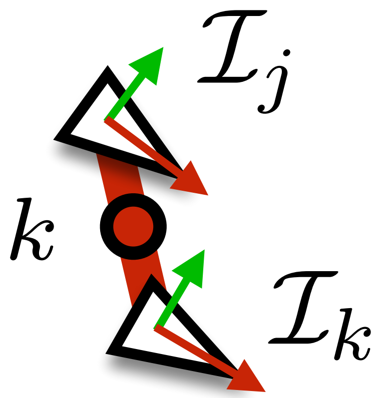
Visual Place-Recognition / Loop-Closure Detection

- Identifying previously visited places to reduce the odometry drift



DETERMINE f such that

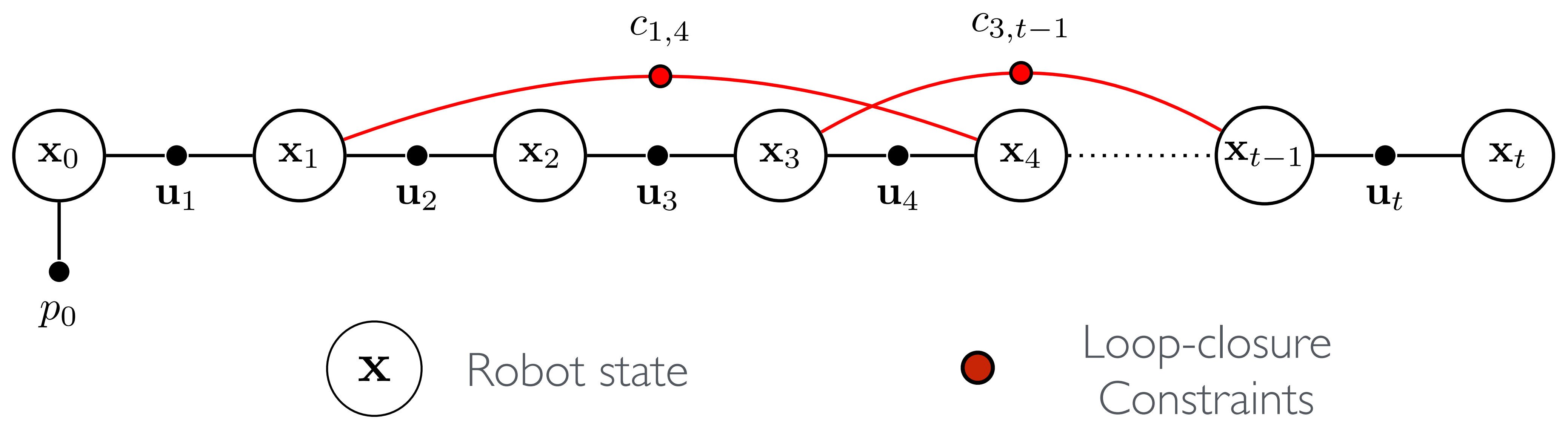
$$f(\mathcal{I}_j) \simeq f(\mathcal{I}_k) \mapsto c_{j,k}$$



Temporally-distant Images

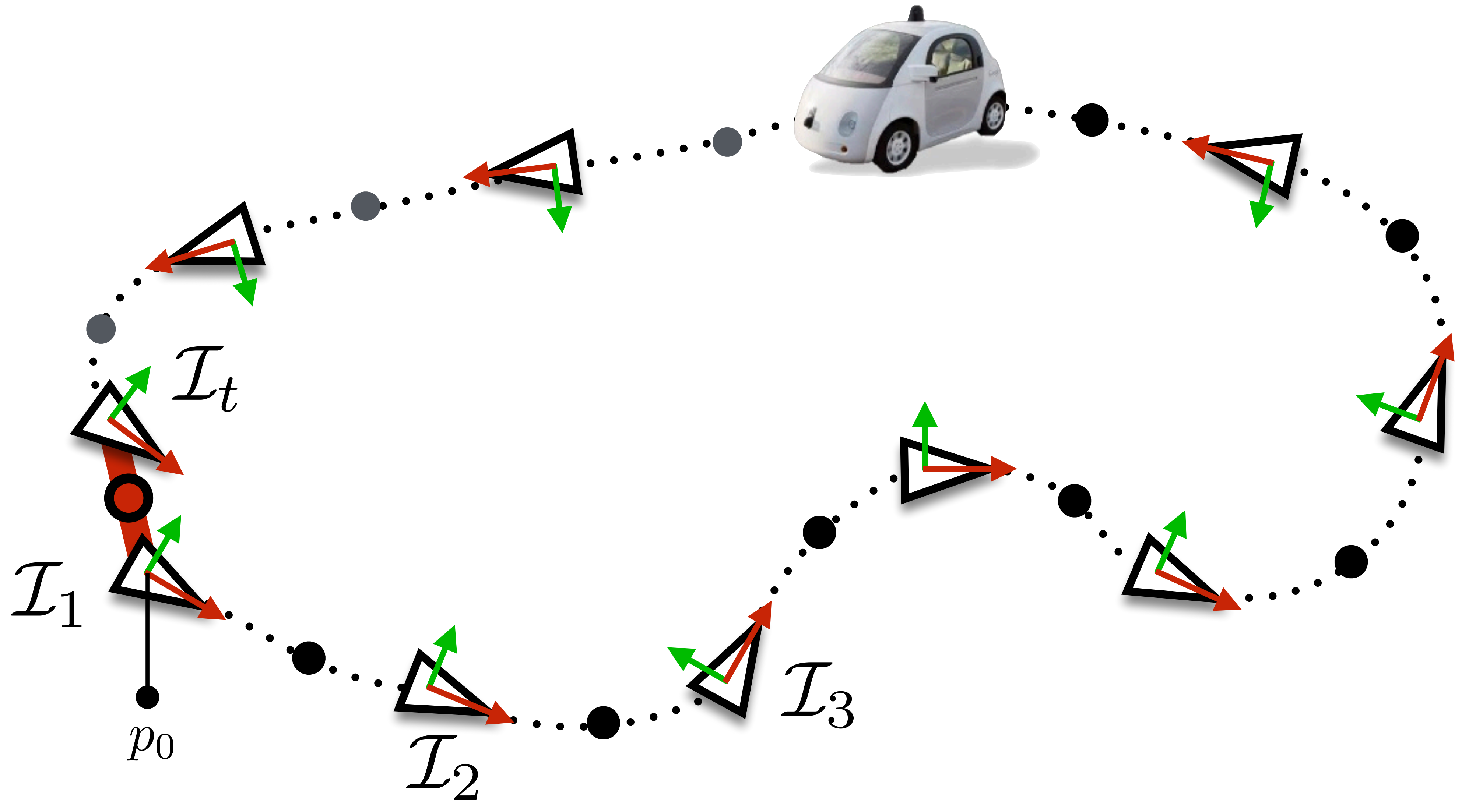
Loop-closure Constraint

Factor Graph for Vision-based Pose-Graph SLAM



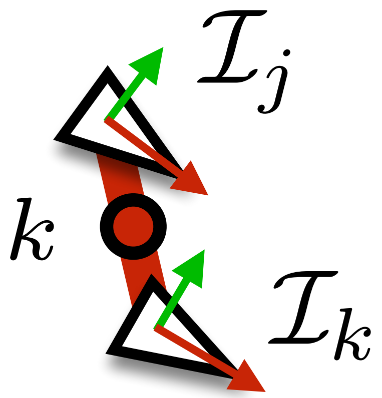
VISION-BASED LOOP-CLOSURE DETECTION

- **Visual Place-Recognition / Loop-Closure Detection**
  - Identifying previously visited places to reduce the odometry drift



DETERMINE f such that

$$f(\mathcal{I}_j) \simeq f(\mathcal{I}_k) \mapsto c_{j,k}$$



Temporally-distant Images

Loop-closure Constraint

Factor Graph for Vision-based Pose-Graph SLAM

$$\begin{aligned}
\mathbf{X}^* &= \arg\max_{\mathbf{X}} p(\mathbf{X} \mid \mathbf{U}, \mathbf{Z_c}) \\
&= \arg\min_{\mathbf{X}} \left\{ \underbrace{\sum_{i=1}^{M} \| f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i \|^2_{\Sigma_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{(j,k)\in\mathcal{C}} \| h_c(\mathbf{x}_j, \mathbf{x}_k) - \mathbf{z}_{jk} \|^2_{\Sigma_c}}_{\text{Loop-Closure Constraint Factors}} \right\}
\end{aligned}$$

MOTIVATION

- **Visual Place Recognition / Loop-closure Detection**
  as a front-end measurement for Vision-based SLAM

MOTIVATION

- **Visual Place Recognition / Loop-closure Detection** as a front-end measurement for Vision-based SLAM
  - Histogram-based: BoVW *[Sivic 2003, Levin 2004, Nister 2006]*
  - Whole-Image: GIST / Binarized Images *[Sunderhauf 2011]*
  - FABMAP (BoW + Chow-Liu Approx) *[Cummins 2008]*
  - Temporal: SeqSLAM, CAT-SLAM *[Milford 2012, Maddern 2012]*
  - Density-based: Placeless place-recognition *[Lynen 2014]*

MOTIVATION

- **Visual Place Recognition / Loop-closure Detection**
  as a front-end measurement for Vision-based SLAM
  - Histogram-based: BoVW *[Sivic 2003, Levin 2004, Nister 2006]*
  - Whole-Image: GIST / Binarized Images *[Sunderhauf 2011]*
  - FABMAP (BoW + Chow-Liu Approx) *[Cummins 2008]*
  - Temporal: SeqSLAM, CAT-SLAM *[Milford 2012, Maddern 2012]*
  - Density-based: Placeless place-recognition *[Lynen 2014]*

Hand-engineered descriptions and metrics for matching

SIFT, SURF, ORB, BRIEF, GIST
BOW, VLAD, Fisher Vectors
L1, L2, Cosine, Hamming Distance

MOTIVATION

- **Visual Place Recognition / Loop-closure Detection**
  as a front-end measurement for Vision-based SLAM
  - Histogram-based: BoVW *[Sivic 2003, Levin 2004, Nister 2006]*
  - Whole-Image: GIST / Binarized Images *[Sunderhauf 2011]*
  - FABMAP (BoW + Chow-Liu Approx) *[Cummins 2008]*
  - Temporal: SeqSLAM, CAT-SLAM *[Milford 2012, Maddern 2012]*
  - Density-based: Placeless place-recognition *[Lynen 2014]*

Hand-engineered descriptions and metrics for matching

SIFT, SURF, ORB, BRIEF, GIST
BOW, VLAD, Fisher Vectors
L1, L2, Cosine, Hamming Distance

- **Convolutional Neural Networks**
  - Places205: Scene Recognition *[Zhou 2014, Zhou 2015]*
  - NetVLAD *[Arandjelovic 2017]*
  - Place Recognition with ConvNet Landmarks *[Sunderhauf 2015]*
  - CNN-based Place Recognition *[Chen 2017]*

MOTIVATION

- **Visual Place Recognition / Loop-closure Detection**
  as a front-end measurement for Vision-based SLAM
  - Histogram-based: BoVW *[Sivic 2003, Levin 2004, Nister 2006]*
  - Whole-Image: GIST / Binarized Images *[Sunderhauf 2011]*
  - FABMAP (BoW + Chow-Liu Approx) *[Cummins 2008]*
  - Temporal: SeqSLAM, CAT-SLAM *[Milford 2012, Maddern 2012]*
  - Density-based: Placeless place-recognition *[Lynen 2014]*

Hand-engineered descriptions and metrics for matching

SIFT, SURF, ORB, BRIEF, GIST
BOW, VLAD, Fisher Vectors
L1, L2, Cosine, Hamming Distance

- **Convolutional Neural Networks**
  - Places205: Scene Recognition *[Zhou 2014, Zhou 2015]*
  - NetVLAD *[Arandjelovic 2017]*
  - Place Recognition with ConvNet Landmarks *[Sunderhauf 2015]*
  - CNN-based Place Recognition *[Chen 2017]*

Supervising scene recognition is tedious / expensive

Require large amounts of training data

Rich feature capacity

Scalable

Pre-trained recognition models

MOTIVATION

- **Visual Place Recognition / Loop-closure Detection**
  as a front-end measurement for Vision-based SLAM
  - Histogram-based: BoVW *[Sivic 2003, Levin 2004, Nister 2006]*
  - Whole-Image: GIST / Binarized Images *[Sunderhauf 2011]*
  - FABMAP (BoW + Chow-Liu Approx) *[Cummins 2008]*
  - Temporal: SeqSLAM, CAT-SLAM *[Milford 2012, Maddern 2012]*
  - Density-based: Placeless place-recognition *[Lynen 2014]*

Hand-engineered descriptions and metrics for matching

SIFT, SURF, ORB, BRIEF, GIST

BOW, VLAD, Fisher Vectors

L1, L2, Cosine, Hamming Distance

Learn a new metric for matching

- **Convolutional Neural Networks**
  - Places205: Scene Recognition *[Zhou 2014, Zhou 2015]*
  - NetVLAD *[Arandjelovic 2017]*
  - Place Recognition with ConvNet Landmarks *[Sunderhauf 2015]*
  - CNN-based Place Recognition *[Chen 2017]*

Supervising scene recognition is tedious / expensive

Require large amounts of training data

Rich feature capacity | Scalable | Pre-trained recognition models

MOTIVATION

- **Visual Place Recognition / Loop-closure Detection**
  as a front-end measurement for Vision-based SLAM
  - Histogram-based: BoVW *[Sivic 2003, Levin 2004, Nister 2006]*
  - Whole-Image: GIST / Binarized Images *[Sunderhauf 2011]*
  - FABMAP (BoW + Chow-Liu Approx) *[Cummins 2008]*
  - Temporal: SeqSLAM, CAT-SLAM *[Milford 2012, Maddern 2012]*
  - Density-based: Placeless place-recognition *[Lynen 2014]*

Learn a new metric for matching

- **Convolutional Neural Networks**
  - Places205: Scene Recognition *[Zhou 2014, Zhou 2015]*
  - NetVLAD *[Arandjelovic 2017]*
  - Place Recognition with ConvNet Landmarks *[Sunderhauf 2015]*
  - CNN-based Place Recognition *[Chen 2017]*

SLAM-aware Self-Supervision
in Mobile Robots

Rich feature capacity

Scalable

Pre-trained recognition models

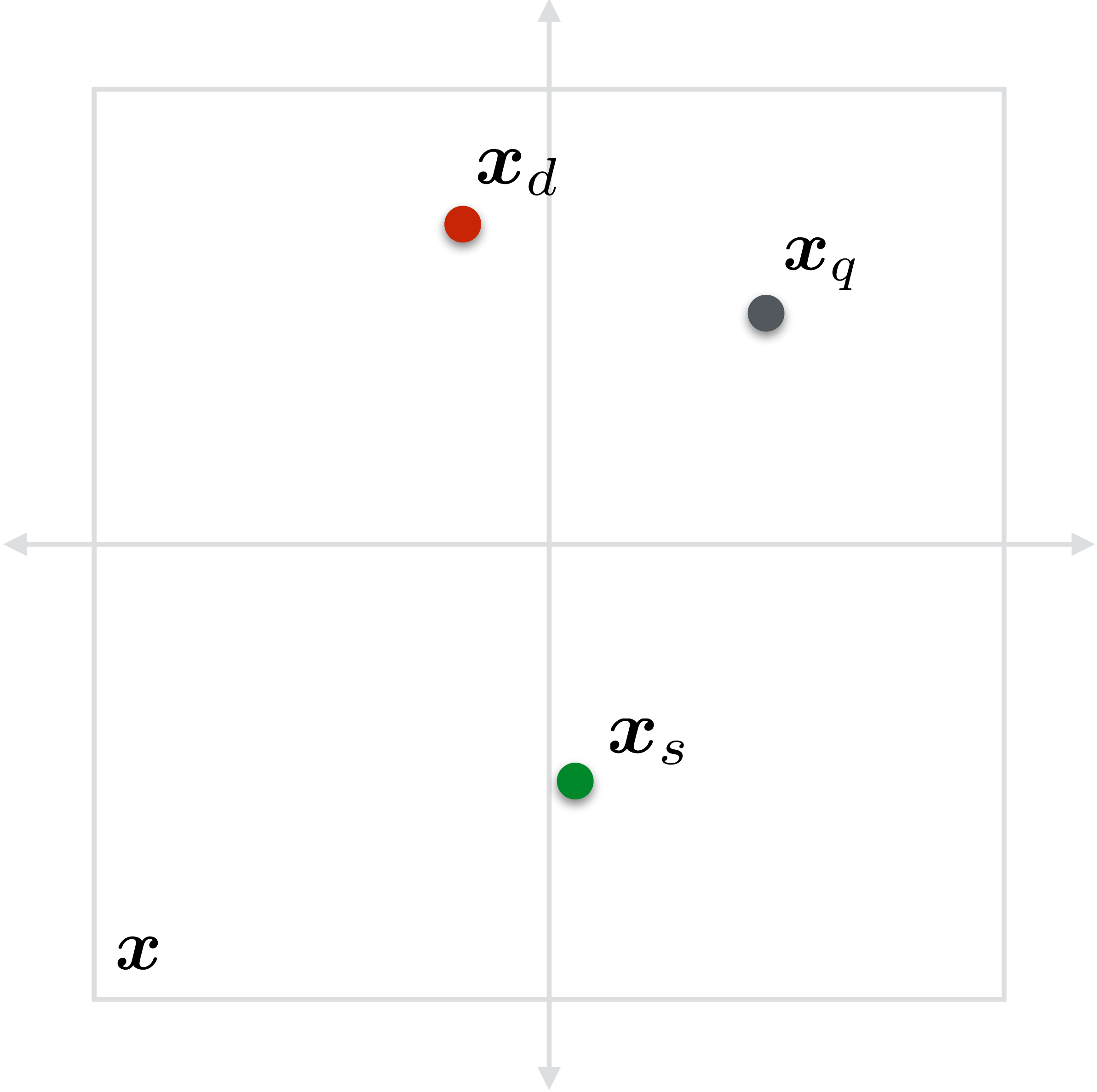
METRIC LEARNING

Learn a new metric for matching

METRIC LEARNING

Learn a new metric for matching

Arbitrarily-defined Distance Measure
(Meaningless)



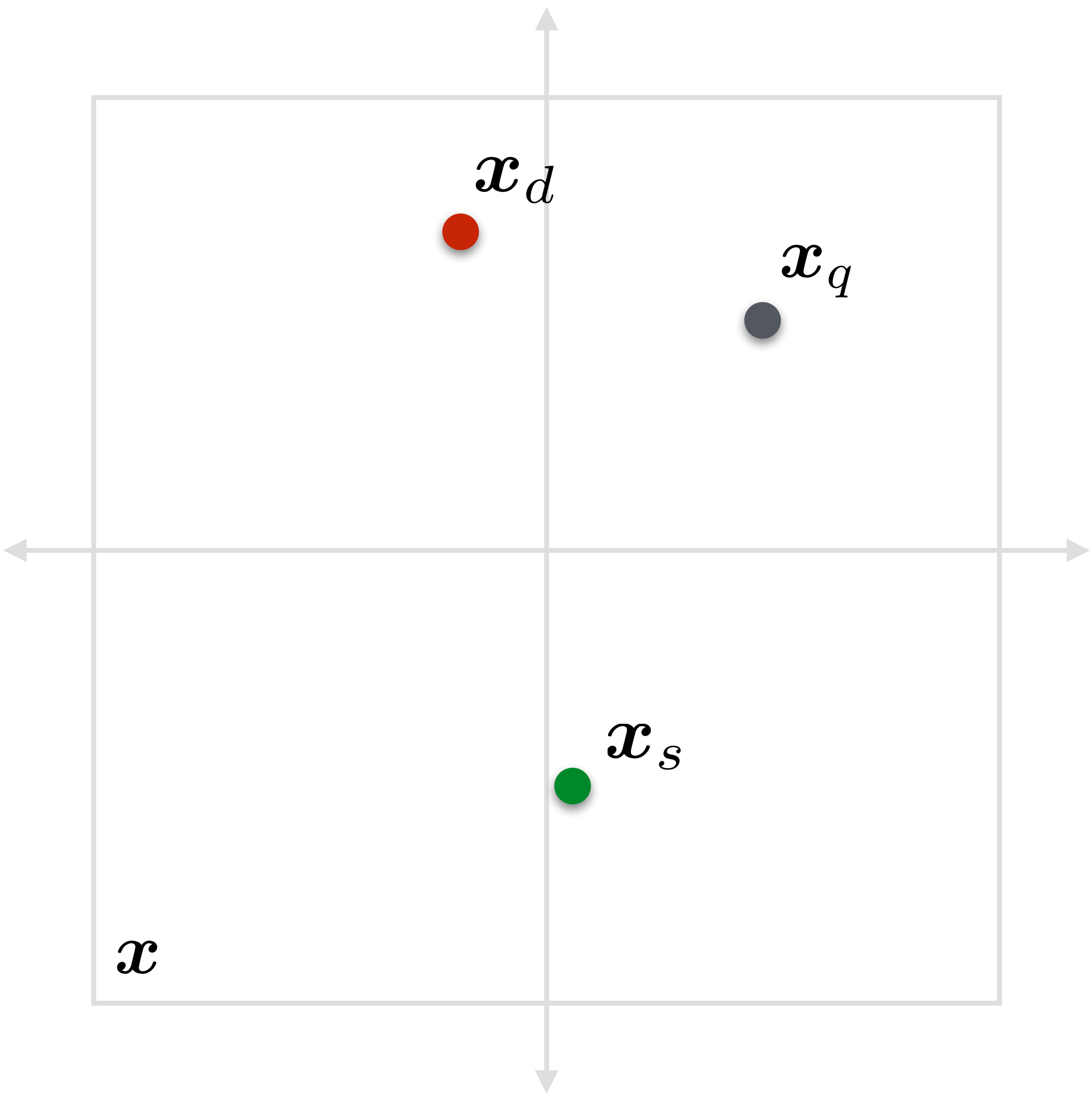
$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

$$\mathcal{X}_S := \{(\boldsymbol{x}_q, \boldsymbol{x}_s) \mid \boldsymbol{x}_q \text{ and } \boldsymbol{x}_s \text{ are in the } \textit{same} \text{ class}\}$$
$$\mathcal{X}_D := \{(\boldsymbol{x}_q, \boldsymbol{x}_d) \mid \boldsymbol{x}_q \text{ and } \boldsymbol{x}_d \text{ are in } \textit{different} \text{ classes}\}$$

METRIC LEARNING

Learn a new metric for matching

Arbitrarily-defined Distance Measure
(Meaningless)

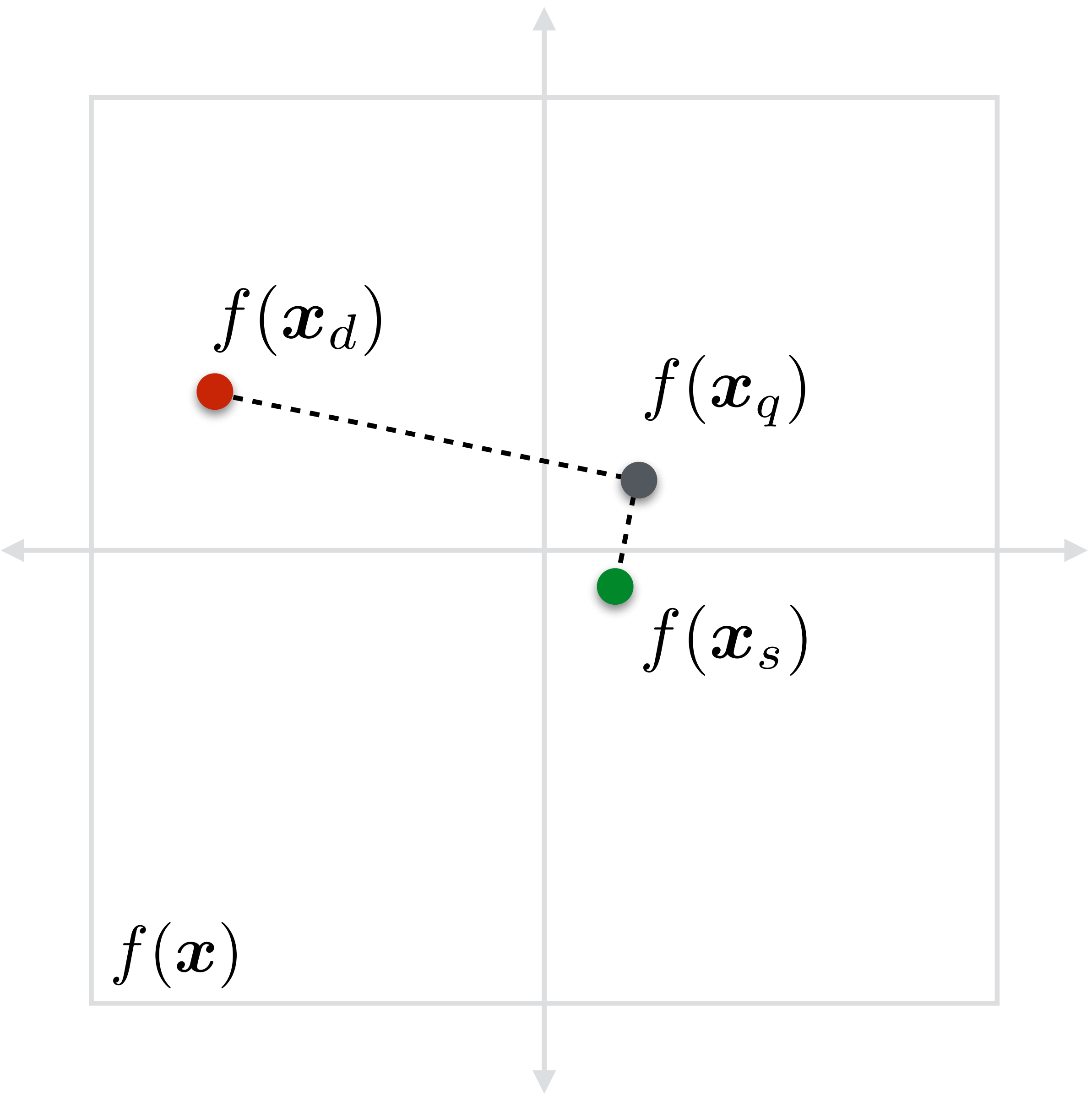


$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

Determine

$$f(\boldsymbol{x}; \theta)$$

"Semantic" Distance Measure
(Task appropriate)



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f(\boldsymbol{x}_i; \theta) - f(\boldsymbol{x}_j; \theta)\|_2$$

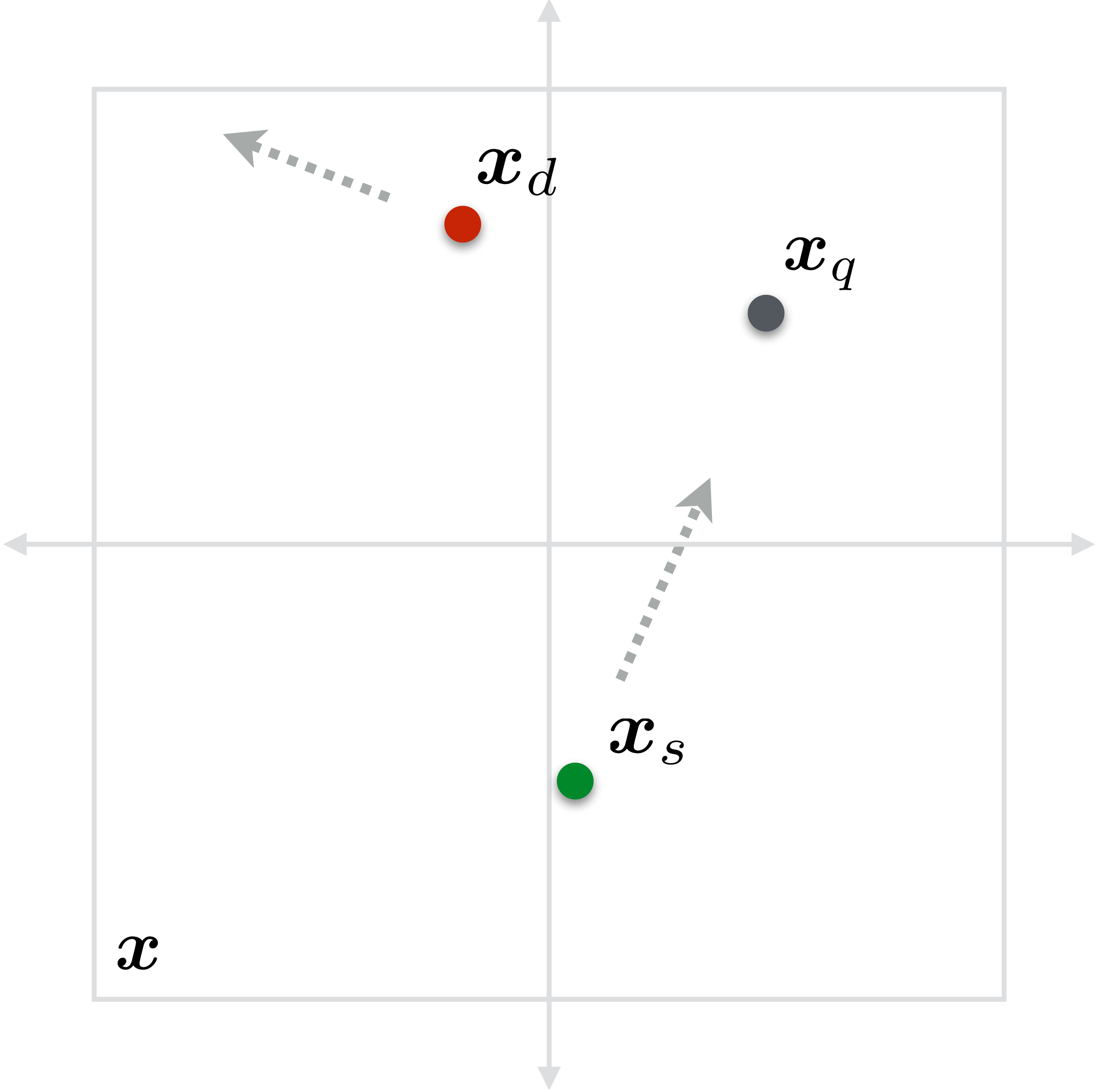
$$\mathcal{X}_S := \{(\boldsymbol{x}_q, \boldsymbol{x}_s) \mid \boldsymbol{x}_q \text{ and } \boldsymbol{x}_s \text{ are in the } \textit{same} \text{ class}\}$$

$$\mathcal{X}_D := \{(\boldsymbol{x}_q, \boldsymbol{x}_d) \mid \boldsymbol{x}_q \text{ and } \boldsymbol{x}_d \text{ are in } \textit{different} \text{ classes}\}$$

METRIC LEARNING

Learn a new metric for matching

Arbitrarily-defined Distance Measure
(Meaningless)



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

Determine

$$f(\boldsymbol{x}; \theta)$$

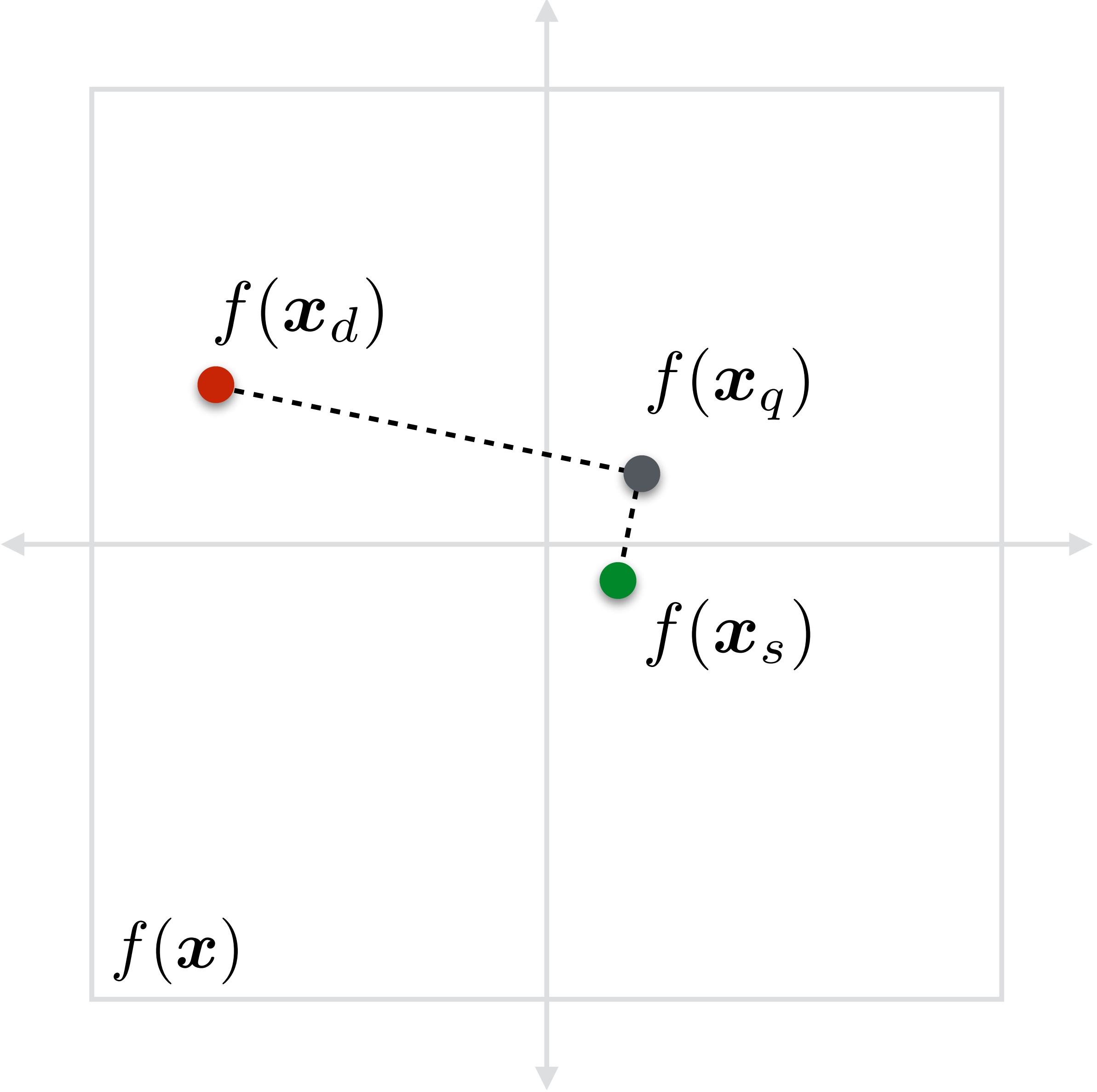
such that we minimize

$$\mathcal{L}(\theta) = \underbrace{\sum_{(\boldsymbol{x}_q, \boldsymbol{x}_s) \in \mathcal{X}_S} \ell_p(\boldsymbol{x}_q, \boldsymbol{x}_s)}_{\text{Penalize similar examples that are far away}} + \underbrace{\sum_{(\boldsymbol{x}_q, \boldsymbol{x}_d) \in \mathcal{X}_D} \ell_n(\boldsymbol{x}_q, \boldsymbol{x}_d)}_{\text{Penalize dissimilar examples that are nearby}}$$

$$\mathcal{X}_S := \{(\boldsymbol{x}_q, \boldsymbol{x}_s) \mid \boldsymbol{x}_q \text{ and } \boldsymbol{x}_s \text{ are in the } \textit{same} \text{ class}\}$$

$$\mathcal{X}_D := \{(\boldsymbol{x}_q, \boldsymbol{x}_d) \mid \boldsymbol{x}_q \text{ and } \boldsymbol{x}_d \text{ are in } \textit{different} \text{ classes}\}$$

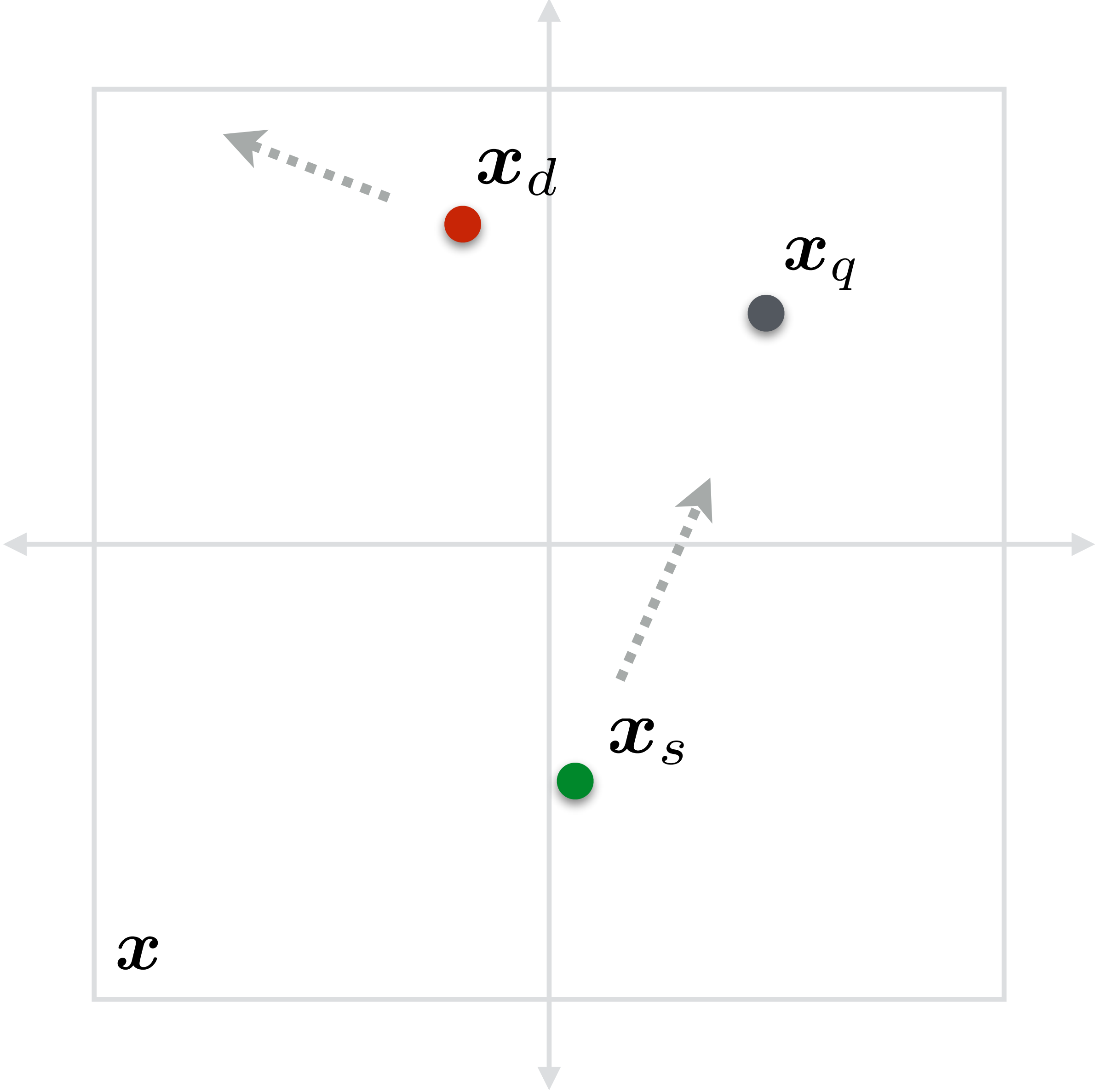
"Semantic" Distance Measure
(Task appropriate)



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f(\boldsymbol{x}_i; \theta) - f(\boldsymbol{x}_j; \theta)\|_2$$

METRIC LEARNING

Arbitrarily-defined Distance Measure
(Meaningless)



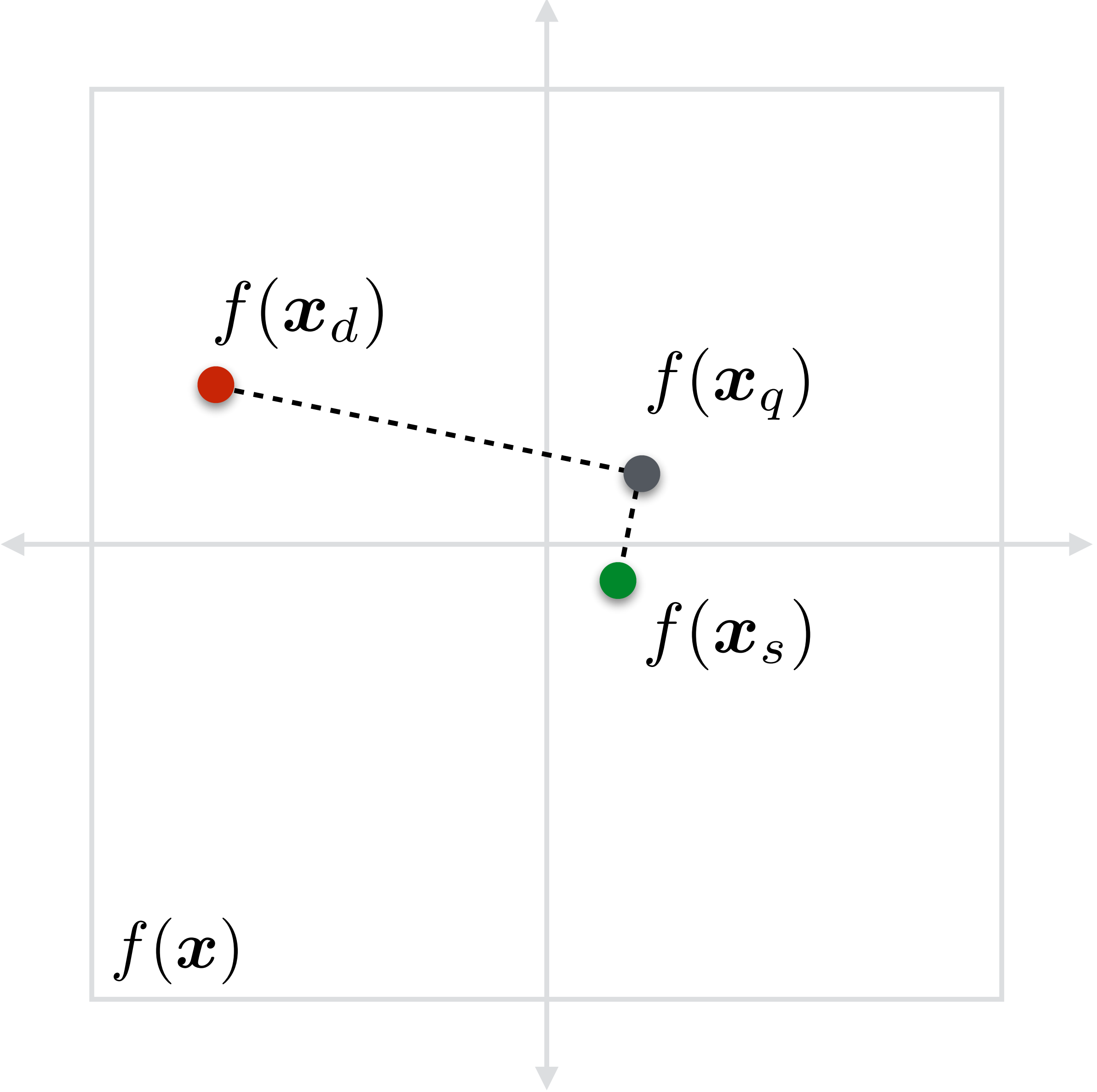
$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

Determine

$$f(\boldsymbol{x}; \theta)$$

such that we minimize

"Semantic" Distance Measure
(Task appropriate)

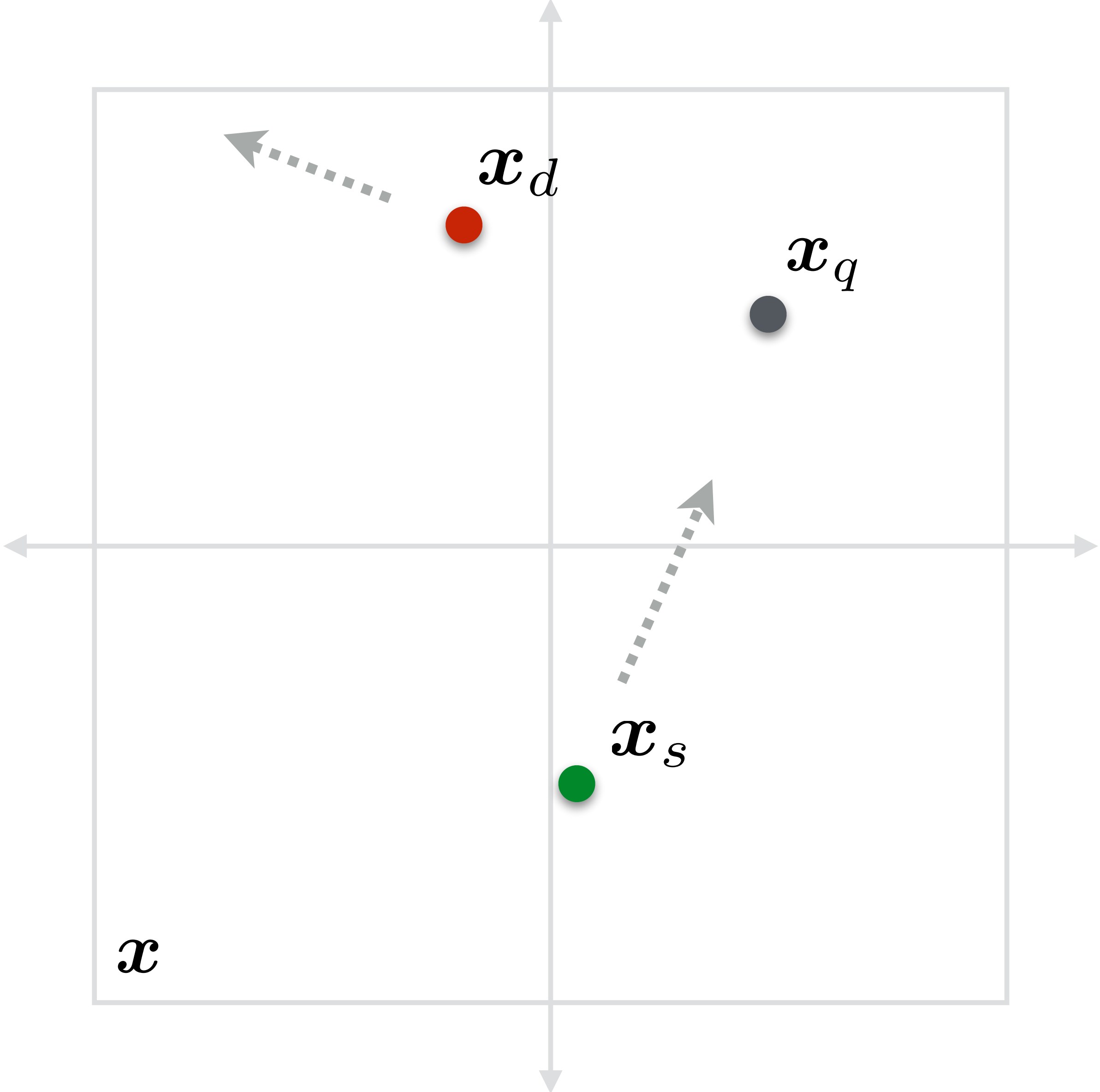


$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f(\boldsymbol{x}_i; \theta) - f(\boldsymbol{x}_j; \theta)\|_2$$

METRIC LEARNING

via Contrastive Loss
(Chopra et al. 2005)

Arbitrarily-defined Distance Measure
(Meaningless)



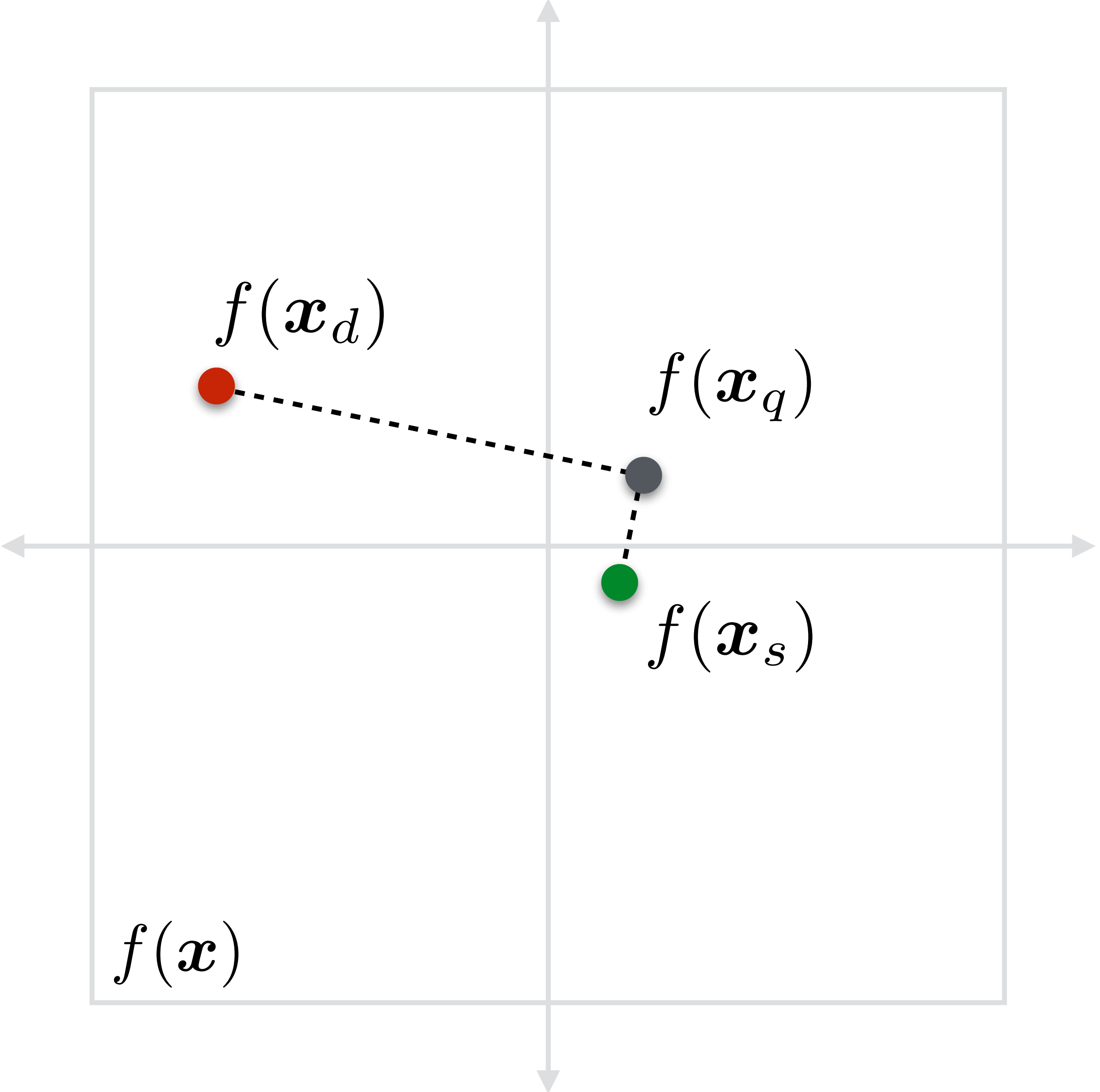
$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

Determine

$$f(\boldsymbol{x}; \theta)$$

such that we minimize

"Semantic" Distance Measure
(Task appropriate)



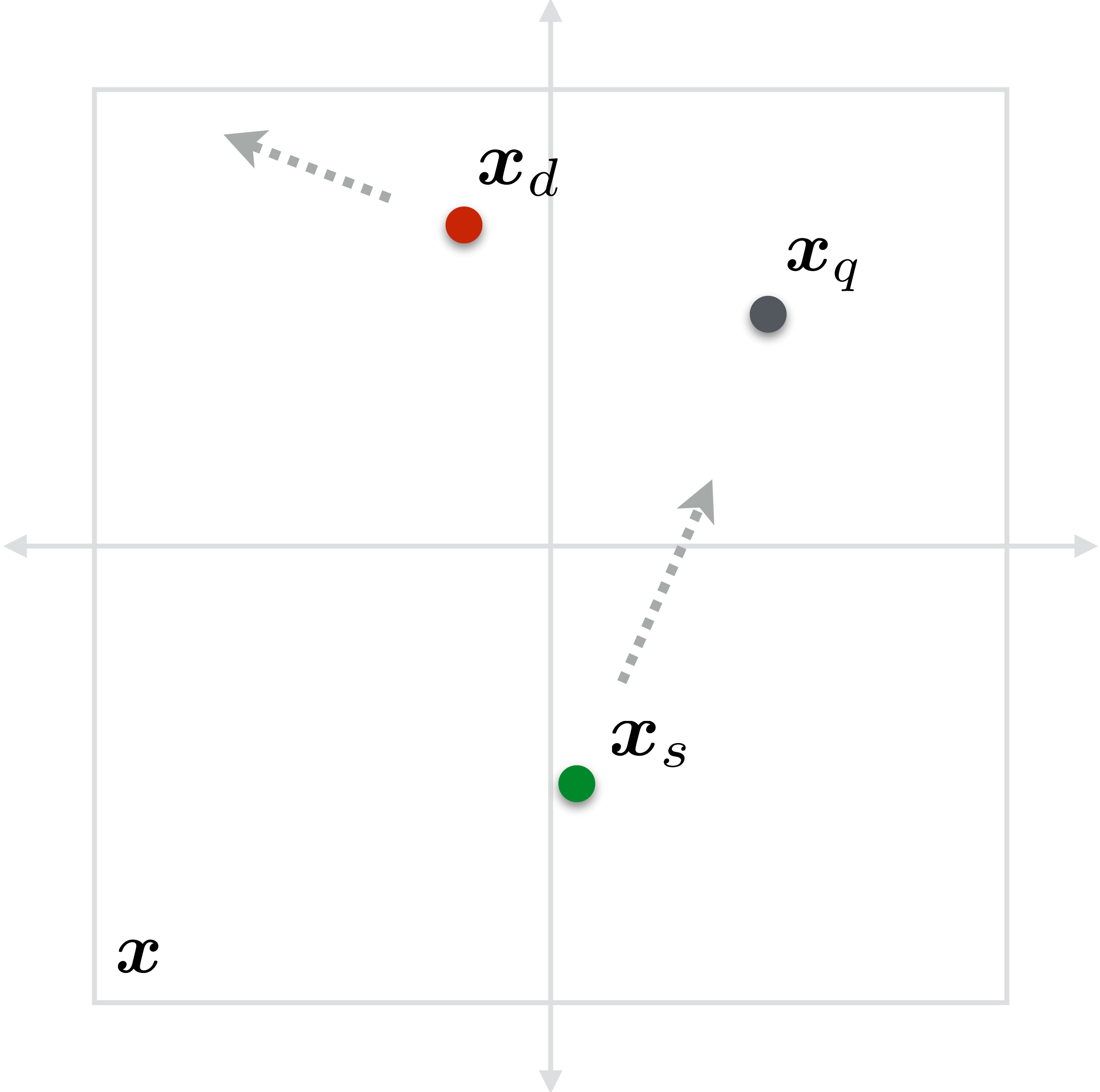
$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f(\boldsymbol{x}_i; \theta) - f(\boldsymbol{x}_j; \theta)\|_2$$

Learning a similarity metric discriminatively, with application to face verification
Chopra et al. 2005

METRIC LEARNING

via Contrastive Loss
(Chopra et al. 2005)

Arbitrarily-defined Distance Measure
(Meaningless)



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

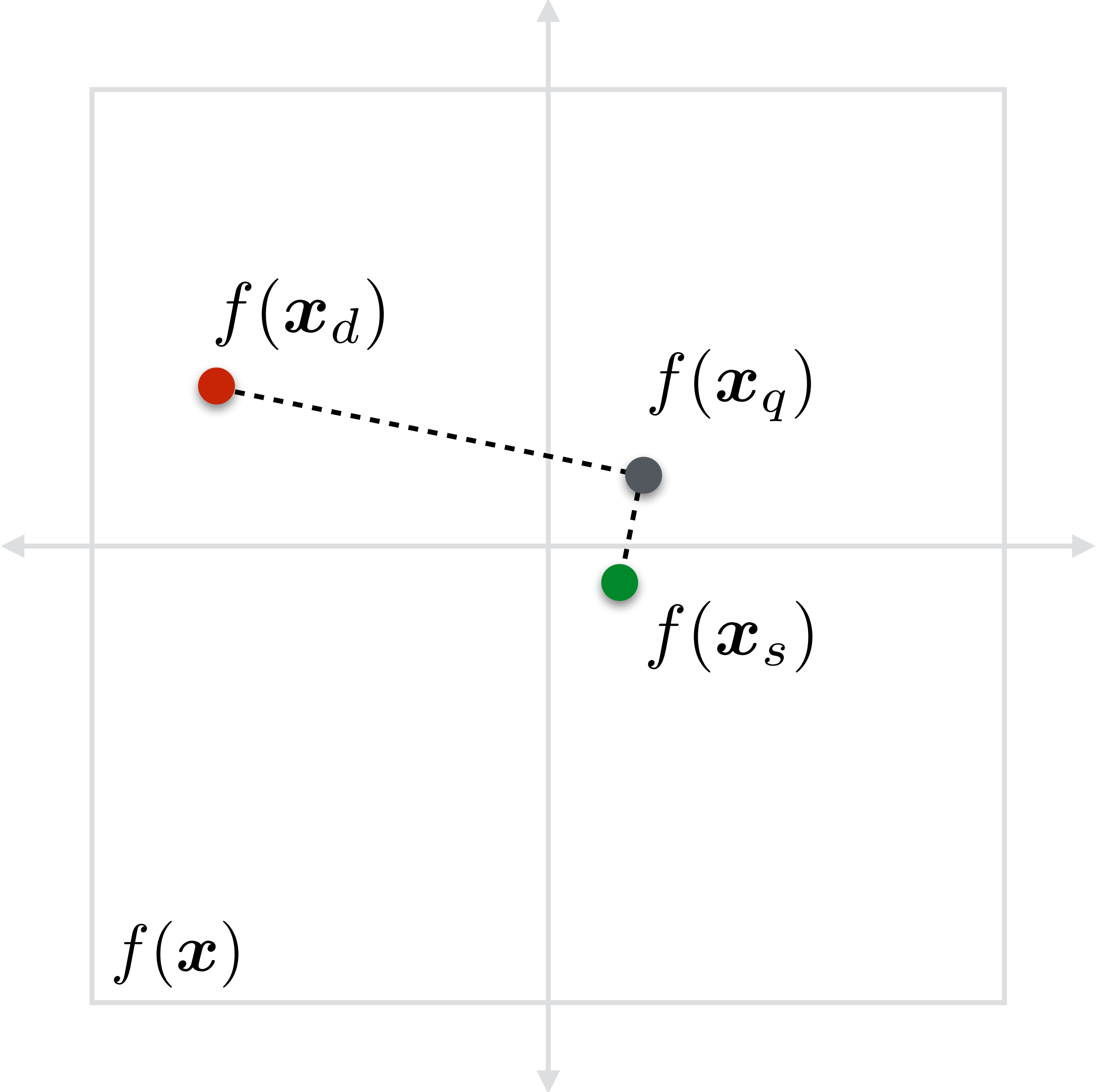
Determine

$$f(\boldsymbol{x}; \theta)$$

such that we minimize

$$\mathcal{L}(\theta) = \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{X}} y D(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + (1 - y) \Big[\alpha - D(\boldsymbol{x}_i, \boldsymbol{x}_j)\Big]_+^2$$

"Semantic" Distance Measure
(Task appropriate)



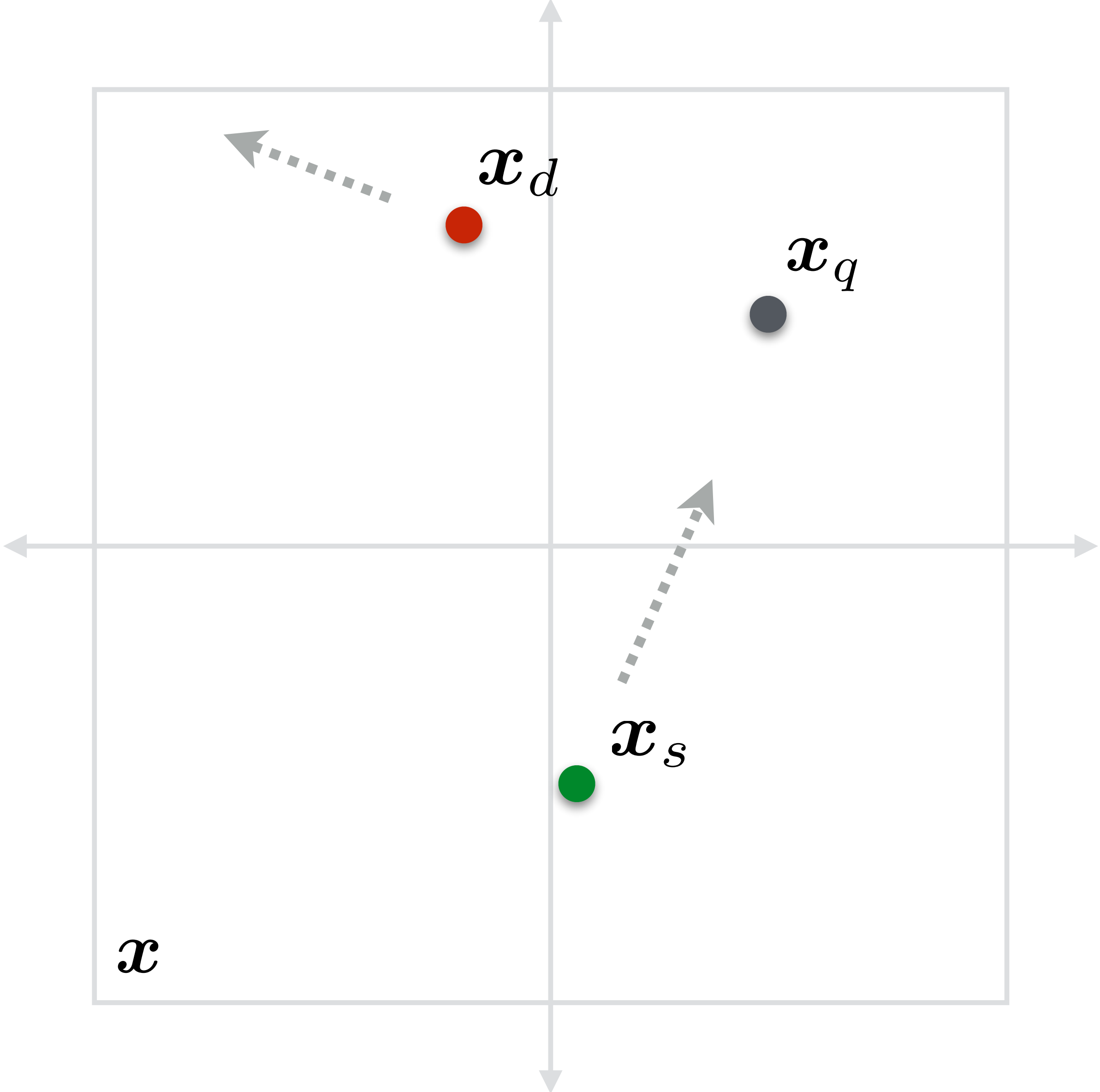
$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f(\boldsymbol{x}_i; \theta) - f(\boldsymbol{x}_j; \theta)\|_2$$

Learning a similarity metric discriminatively, with application to face verification
Chopra et al. 2005

METRIC LEARNING

via Contrastive Loss
(Chopra et al. 2005)

Arbitrarily-defined Distance Measure
(Meaningless)



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

Determine

$$f(\boldsymbol{x}; \theta)$$

such that we minimize

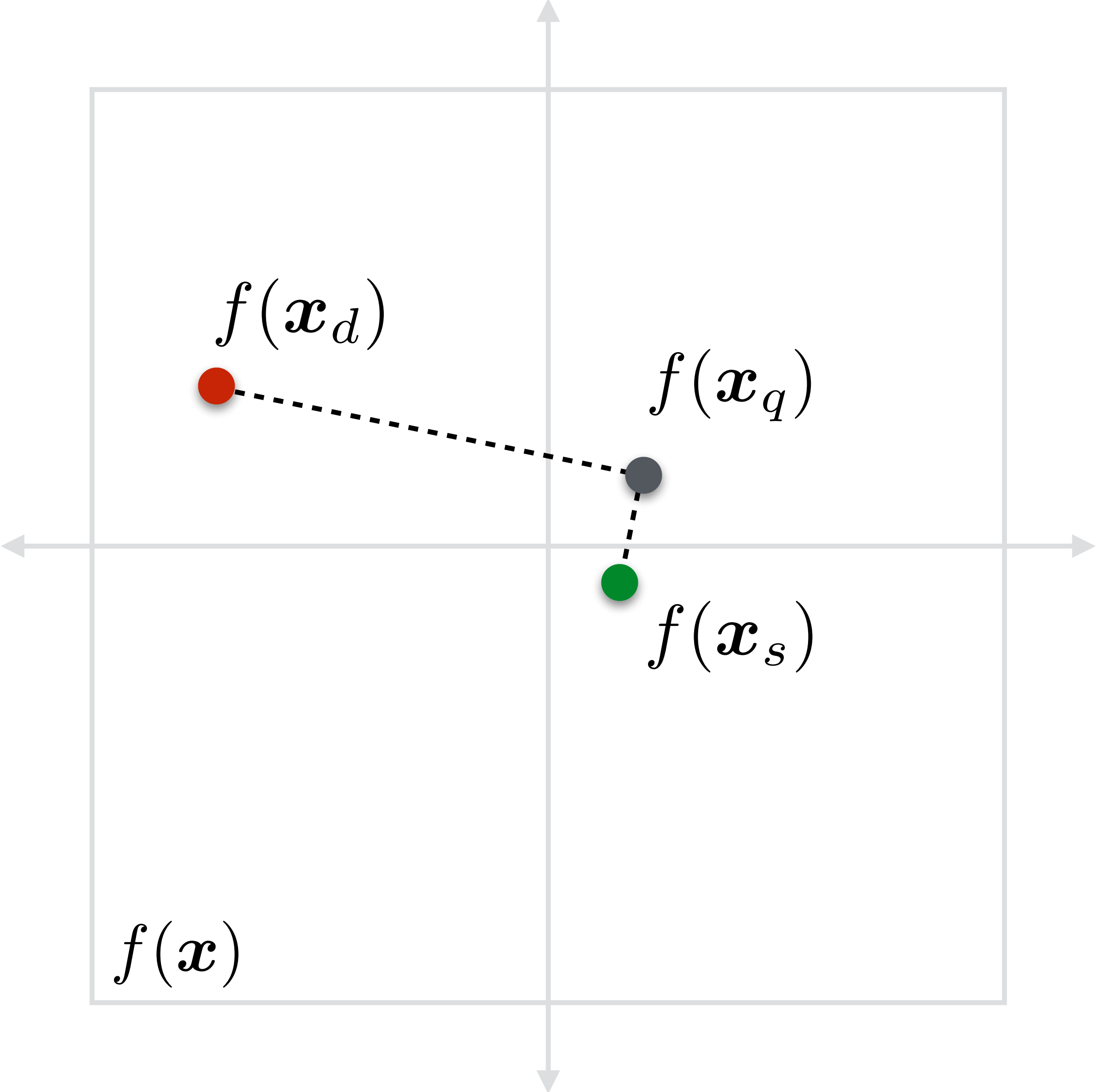
$$\mathcal{L}(\theta) = \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{X}} y D(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + (1-y)\Big[\alpha - D(\boldsymbol{x}_i, \boldsymbol{x}_j)\Big]_+^2$$

where

$$y = \begin{cases} 1 & \text{if } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{X}_S, \\ 0 & \text{if } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{X}_D \end{cases}$$

Supervision

"Semantic" Distance Measure
(Task appropriate)



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f(\boldsymbol{x}_i; \theta) - f(\boldsymbol{x}_j; \theta)\|_2$$

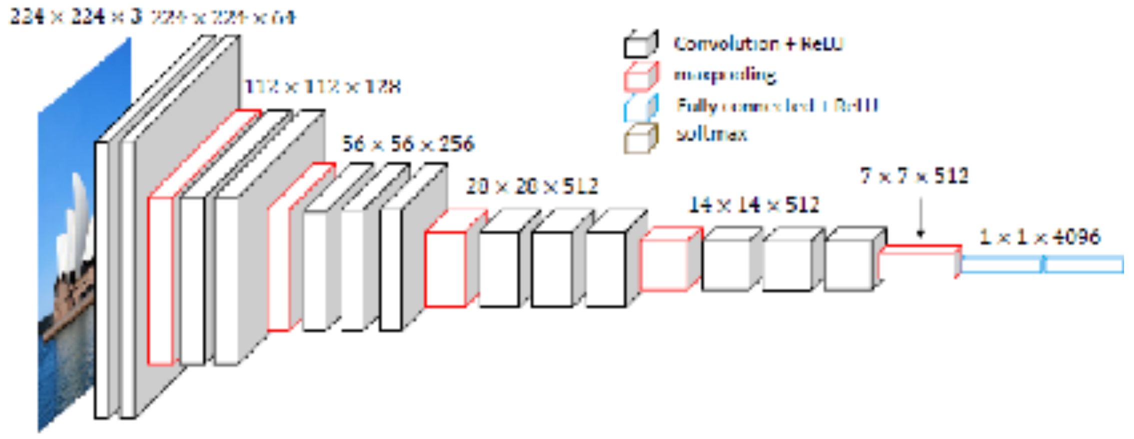
Learning a similarity metric discriminatively, with application to face verification
Chopra et al. 2005

SELF-SUPERVISED METRIC LEARNING FOR VISUAL PLACE-RECOGNITION

SELF-SUPERVISED METRIC LEARNING FOR VISUAL PLACE-RECOGNITION

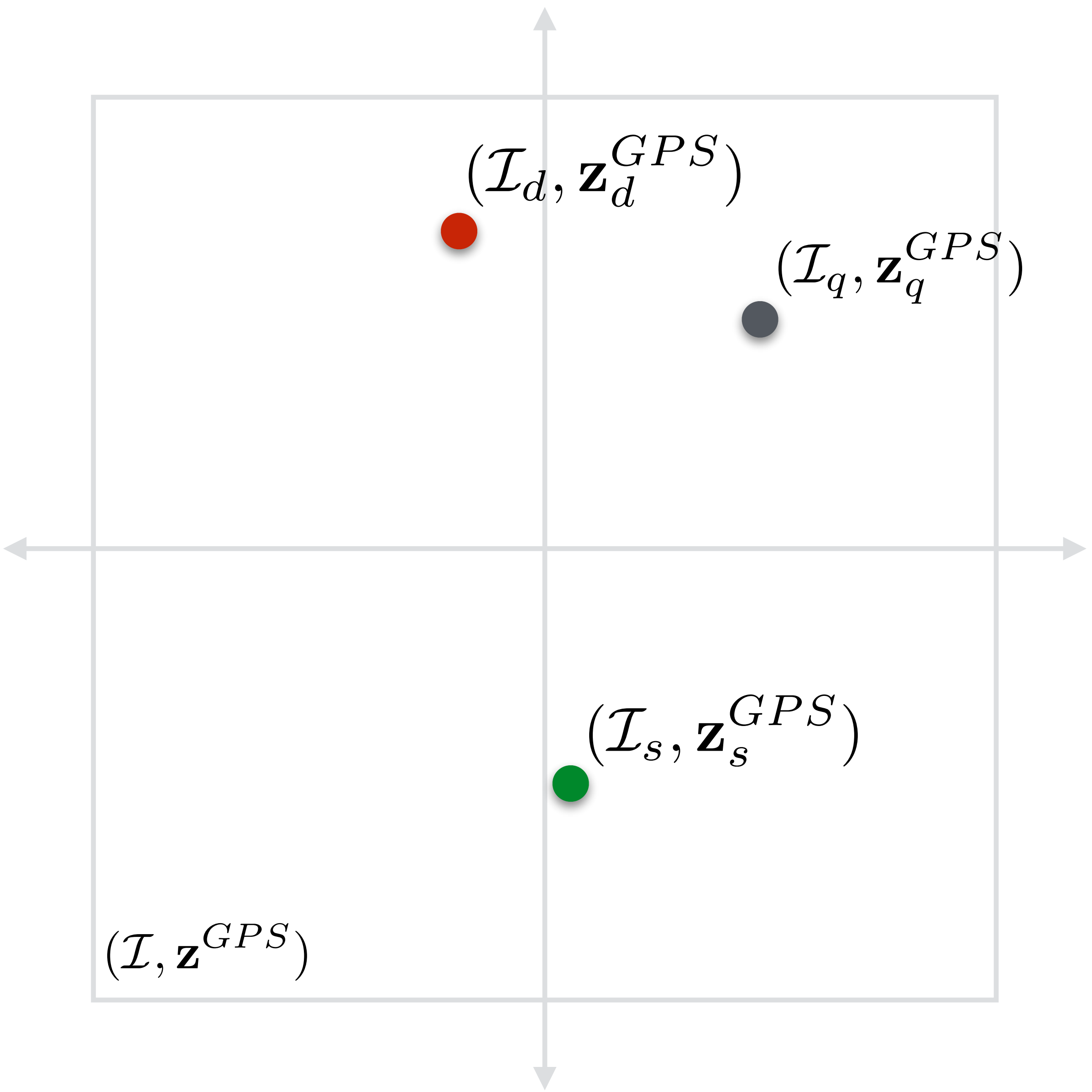
Determine

$$f(\mathcal{I};\theta)$$



SELF-SUPERVISED METRIC LEARNING FOR VISUAL PLACE-RECOGNITION

Cross-modal Image-GPS measurements

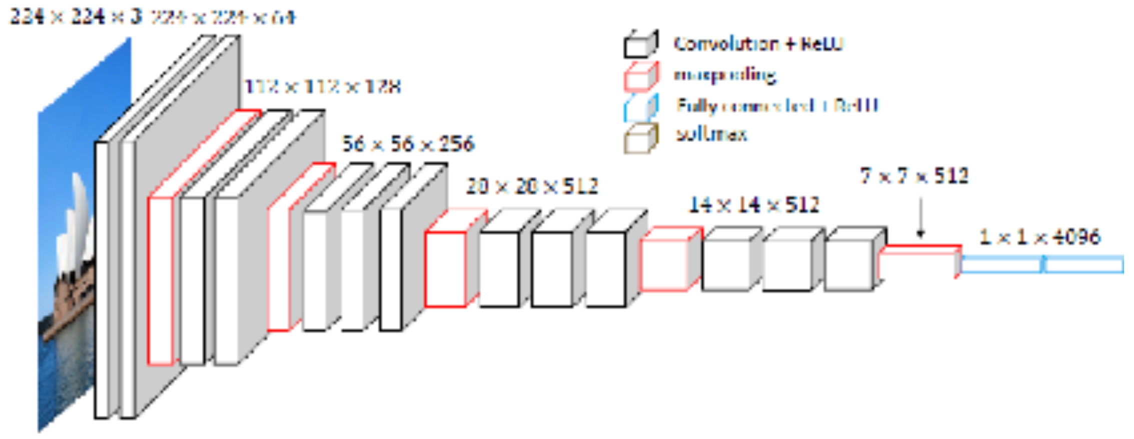


$$D(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) = \mathbf{z}_i^{GPS} \ominus \mathbf{z}_j^{GPS}$$

Distance on SE(2) manifold
(Relative pose transformation)

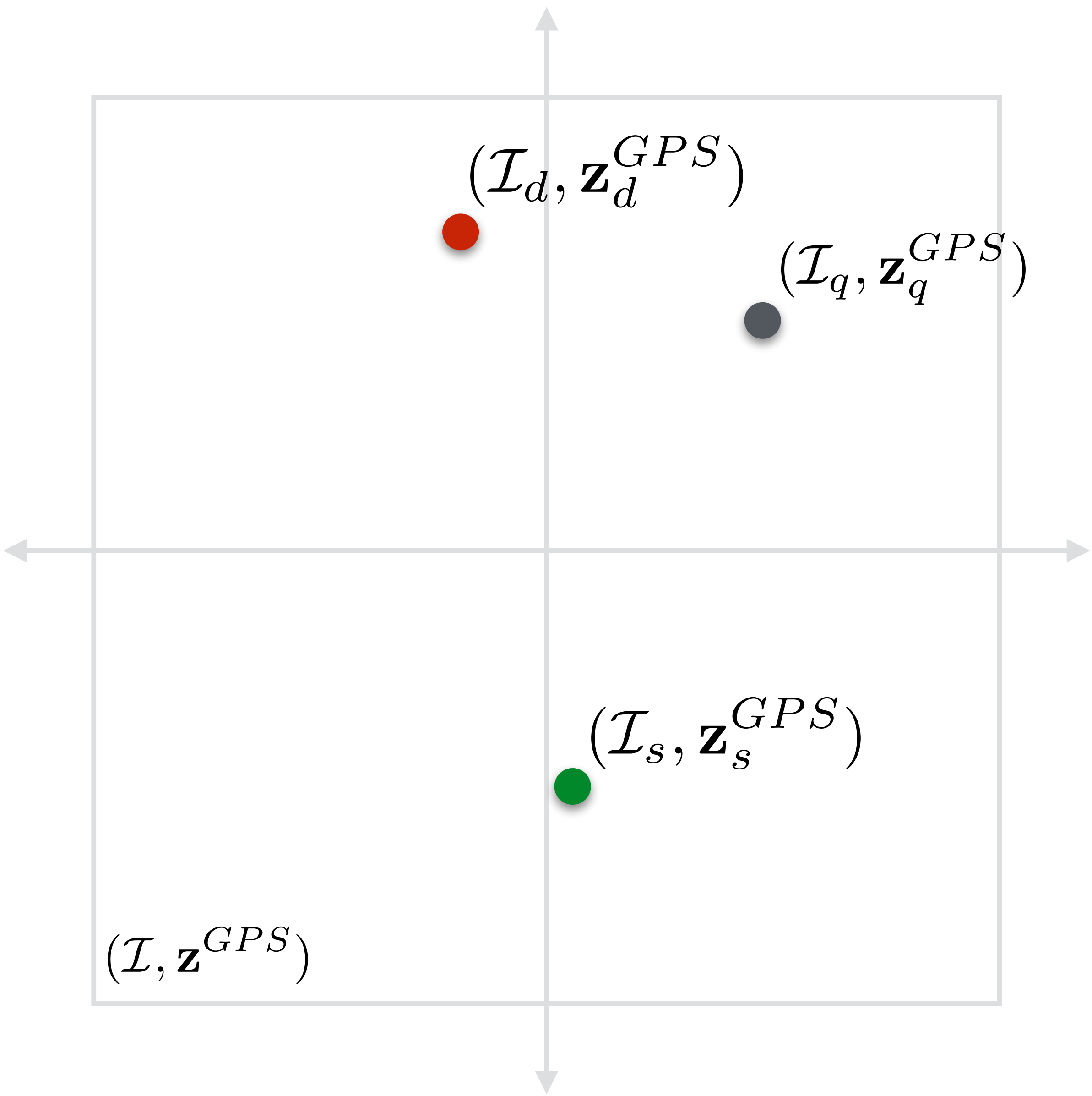
Determine

$$f(\mathcal{I}; \theta)$$



SELF-SUPERVISED METRIC LEARNING FOR VISUAL PLACE-RECOGNITION

Cross-modal Image-GPS measurements

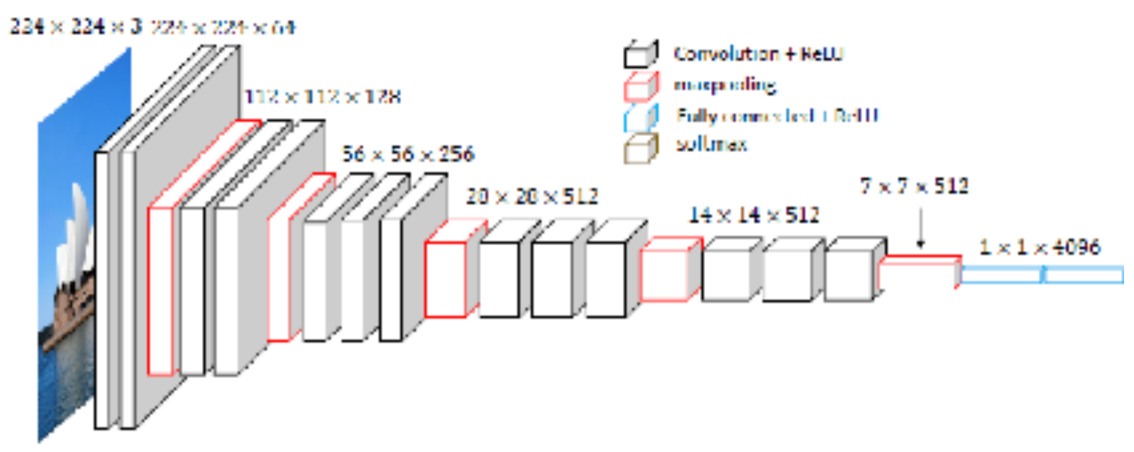


$$D(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) = \mathbf{z}_i^{GPS} \ominus \mathbf{z}_j^{GPS}$$

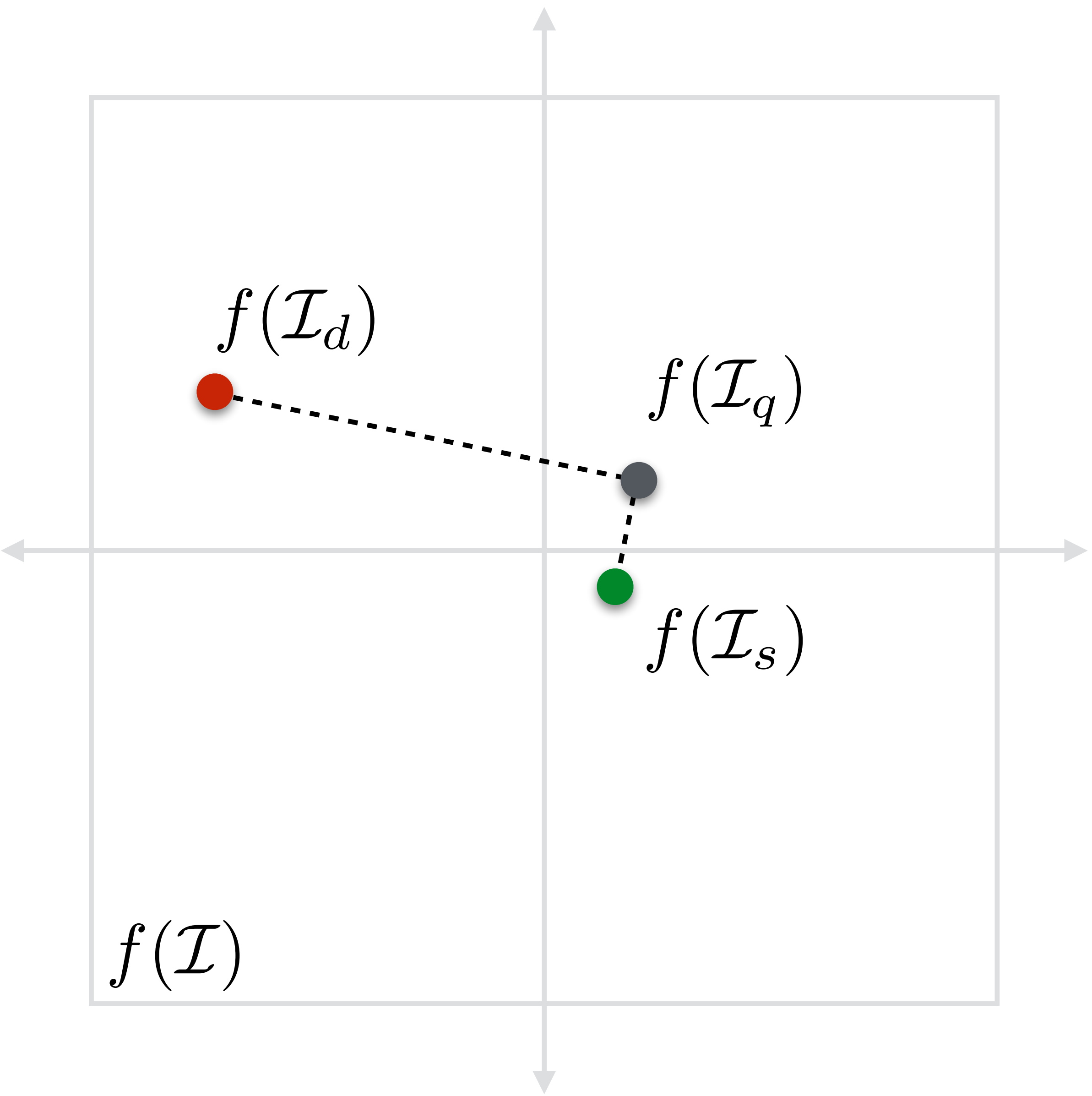
Distance on SE(2) manifold
(Relative pose transformation)

Determine

$$f(\mathcal{I}; \theta)$$



Embedding appropriate for
Visual Loop-closure Detection

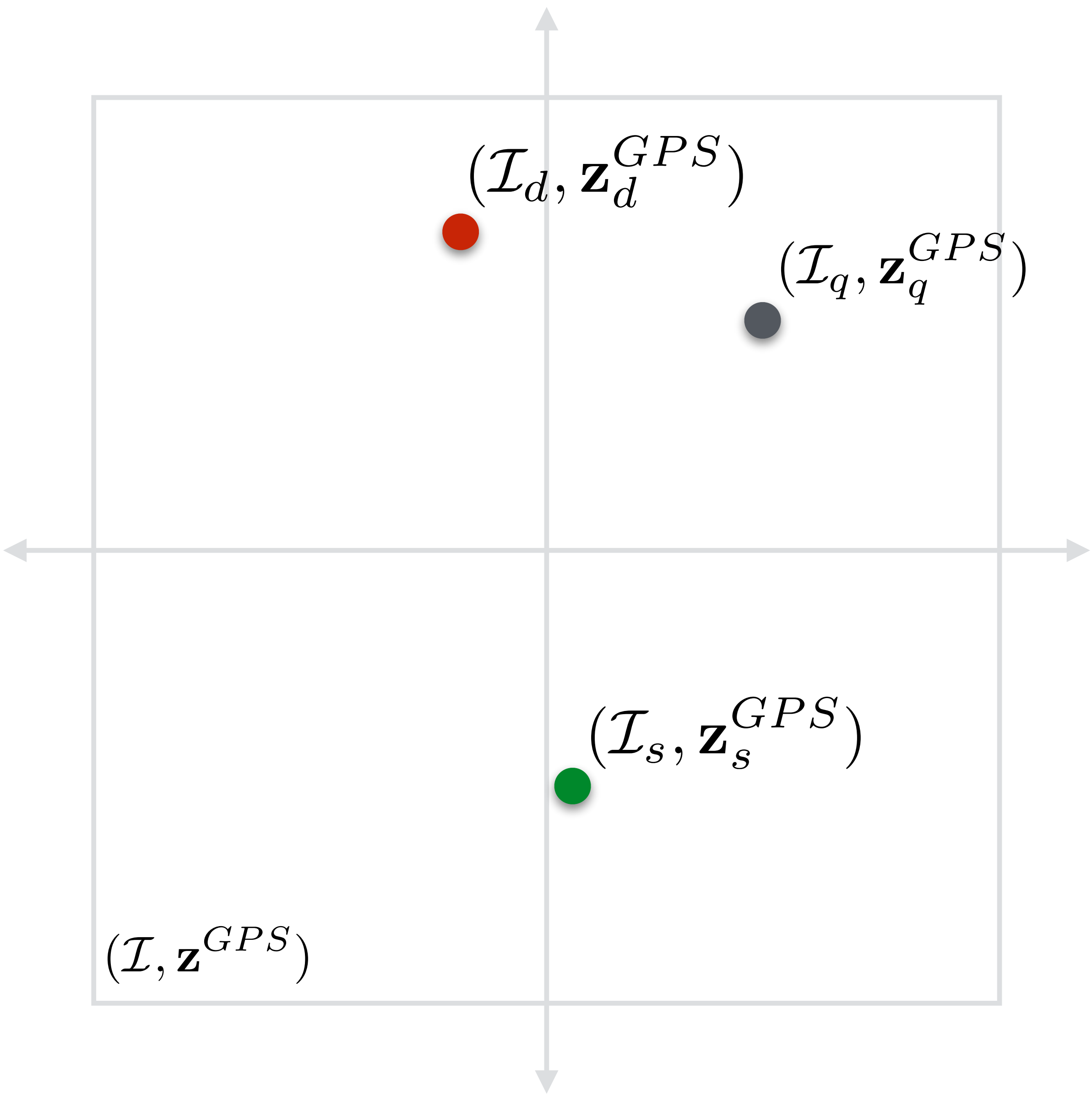


$$D(\mathcal{I}_i, \mathcal{I}_j) = \|f(\mathcal{I}_i) - f(\mathcal{I}_j)\|_2$$

"Semantic" distance in embedding

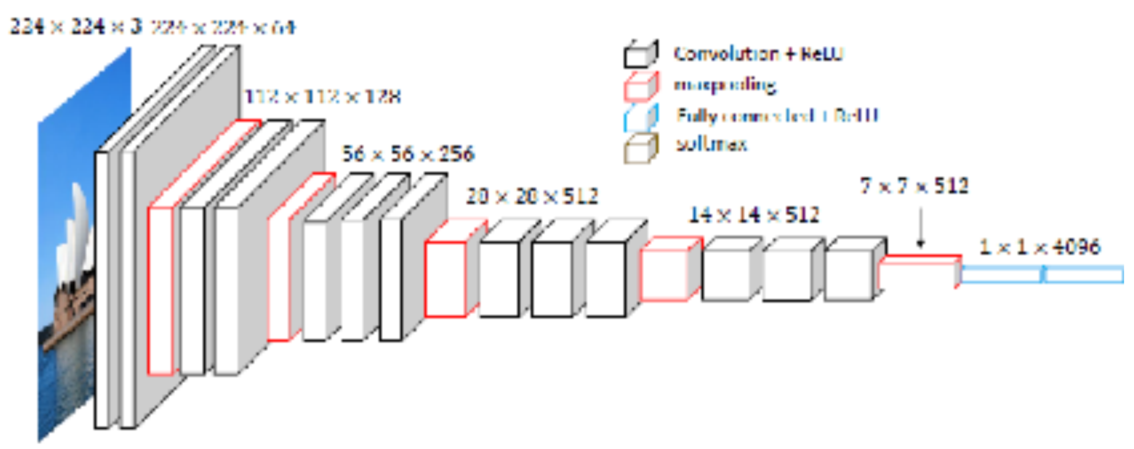
SELF-SUPERVISED METRIC LEARNING FOR VISUAL PLACE-RECOGNITION

Cross-modal Image-GPS measurements



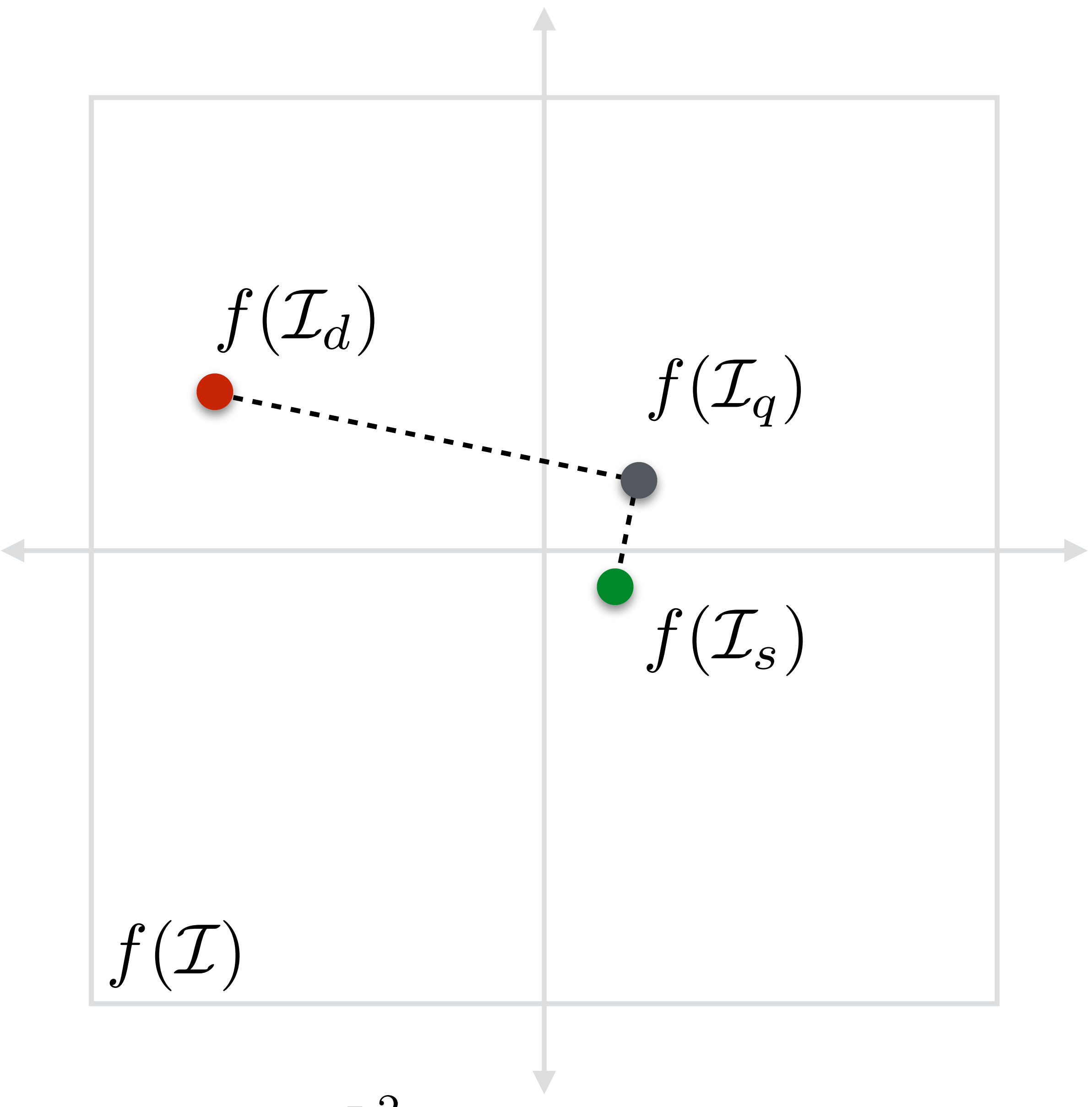
Determine

$$f(\mathcal{I}; \theta)$$



such that we minimize

Embedding appropriate for Visual Loop-closure Detection



$$\mathcal{L}(\theta^{loc}) = \sum_{((\mathcal{I}_i,\mathbf{z}_i),(\mathcal{I}_j,\mathbf{z}_j))\in\mathcal{X}} (\mathbb{1}_{GPS}) \cdot D(\mathcal{I}_i,\mathcal{I}_j)^2 + (1-\mathbb{1}_{GPS}) \cdot \Big[\alpha - D(\mathcal{I}_i,\mathcal{I}_j)\Big]_+^2$$

$$D(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) = \mathbf{z}_i^{GPS} \ominus \mathbf{z}_j^{GPS}$$

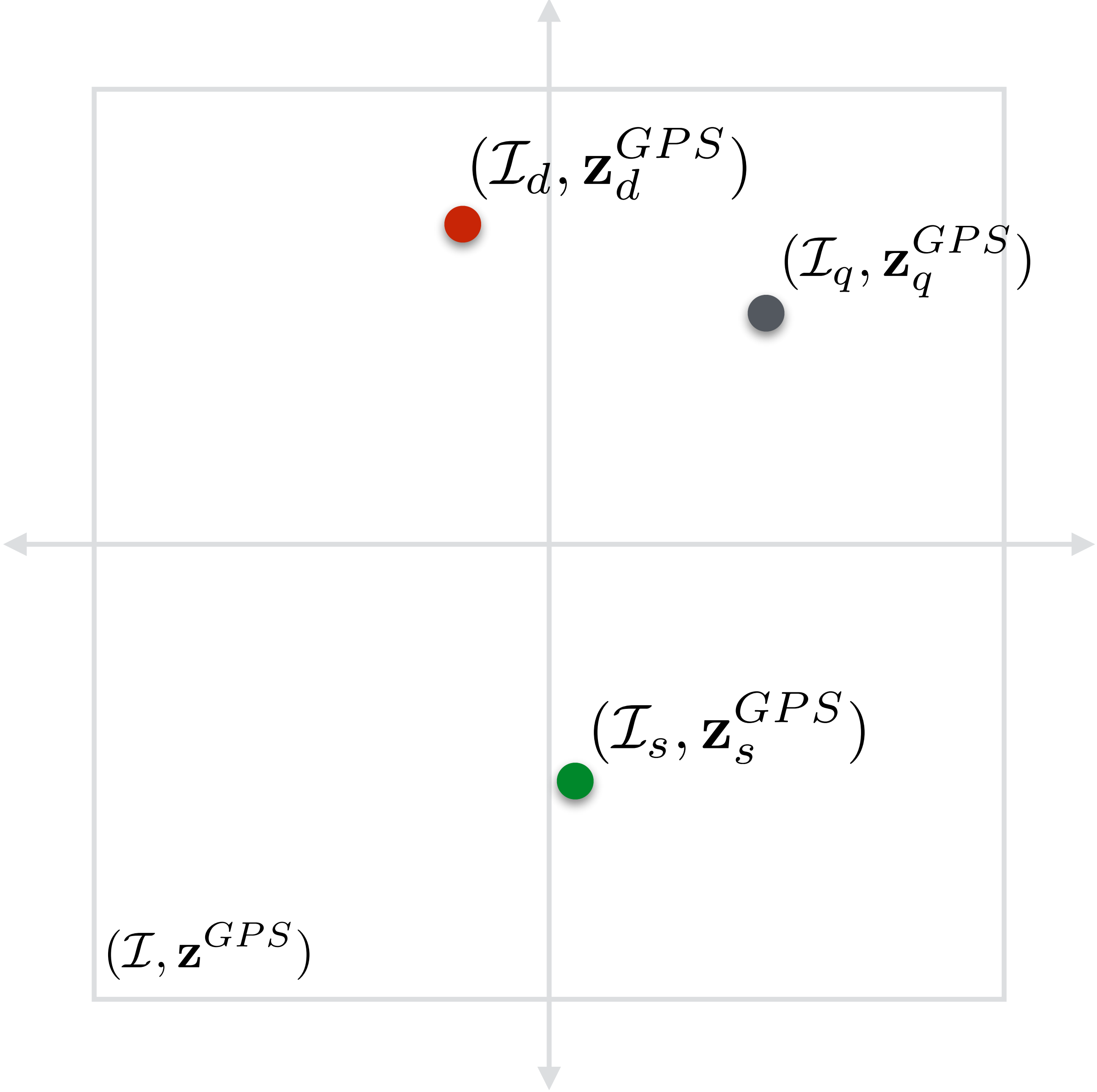
Distance on SE(2) manifold
(Relative pose transformation)

$$D(\mathcal{I}_i, \mathcal{I}_j) = \|f(\mathcal{I}_i) - f(\mathcal{I}_j)\|_2$$

“Semantic” distance in embedding

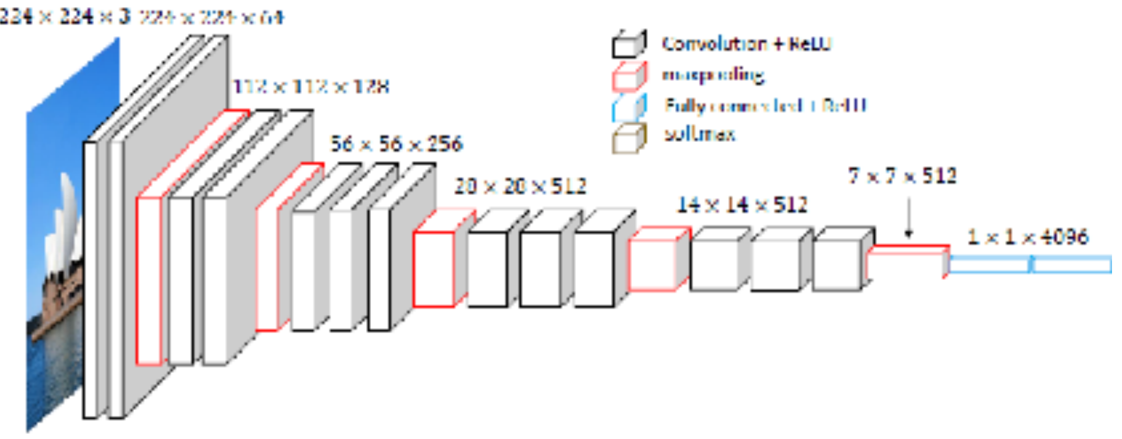
SELF-SUPERVISED METRIC LEARNING FOR VISUAL PLACE-RECOGNITION

Cross-modal Image-GPS measurements



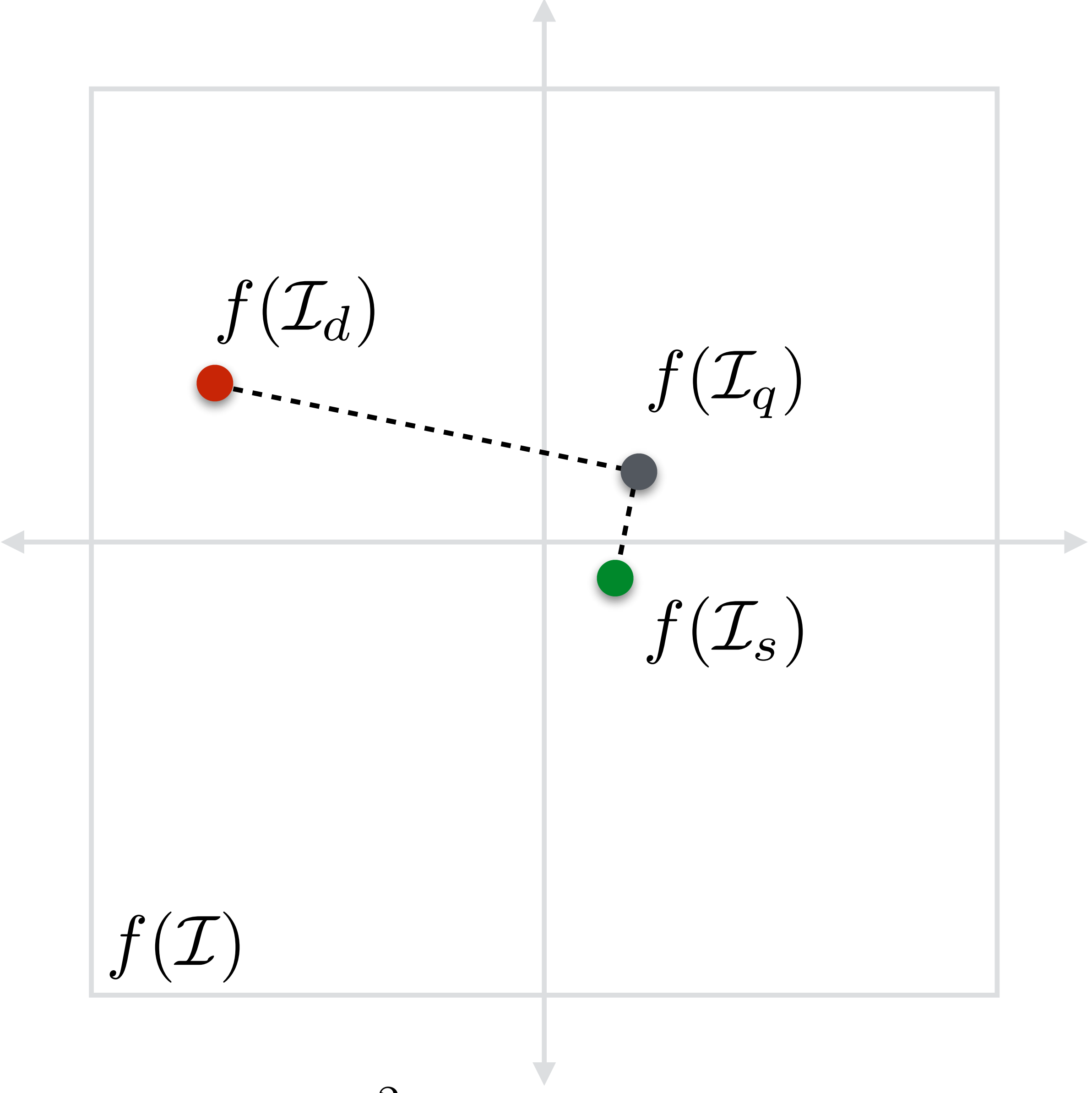
Determine

$$f(\mathcal{I}; \theta)$$



such that we minimize

Embedding appropriate for Visual Loop-closure Detection



$$\mathcal{L}(\theta^{loc}) = \sum_{((\mathcal{I}_i,\mathbf{z}_i),(\mathcal{I}_j,\mathbf{z}_j))\in\mathcal{X}} (\mathbb{1}_{GPS}) \cdot D(\mathcal{I}_i, \mathcal{I}_j)^2 + (1 - \mathbb{1}_{GPS}) \cdot \left[\alpha - D(\mathcal{I}_i, \mathcal{I}_j)\right]_+^2$$

$$D(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) = \mathbf{z}_i^{GPS} \ominus \mathbf{z}_j^{GPS}$$

Distance on SE(2) manifold
(Relative pose transformation)

$$\boxed{\mathbb{1}_{GPS} = \begin{cases} 1 & \text{if } \mathcal{K}(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) > \tau_p^{\mathbf{Rt}} \\ 0 & \text{if } \mathcal{K}(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) < \tau_n^{\mathbf{Rt}} \end{cases}}$$

Self-Supervision

$$D(\mathcal{I}_i, \mathcal{I}_j) = \|f(\mathcal{I}_i) - f(\mathcal{I}_j)\|_2$$

“Semantic” distance in embedding

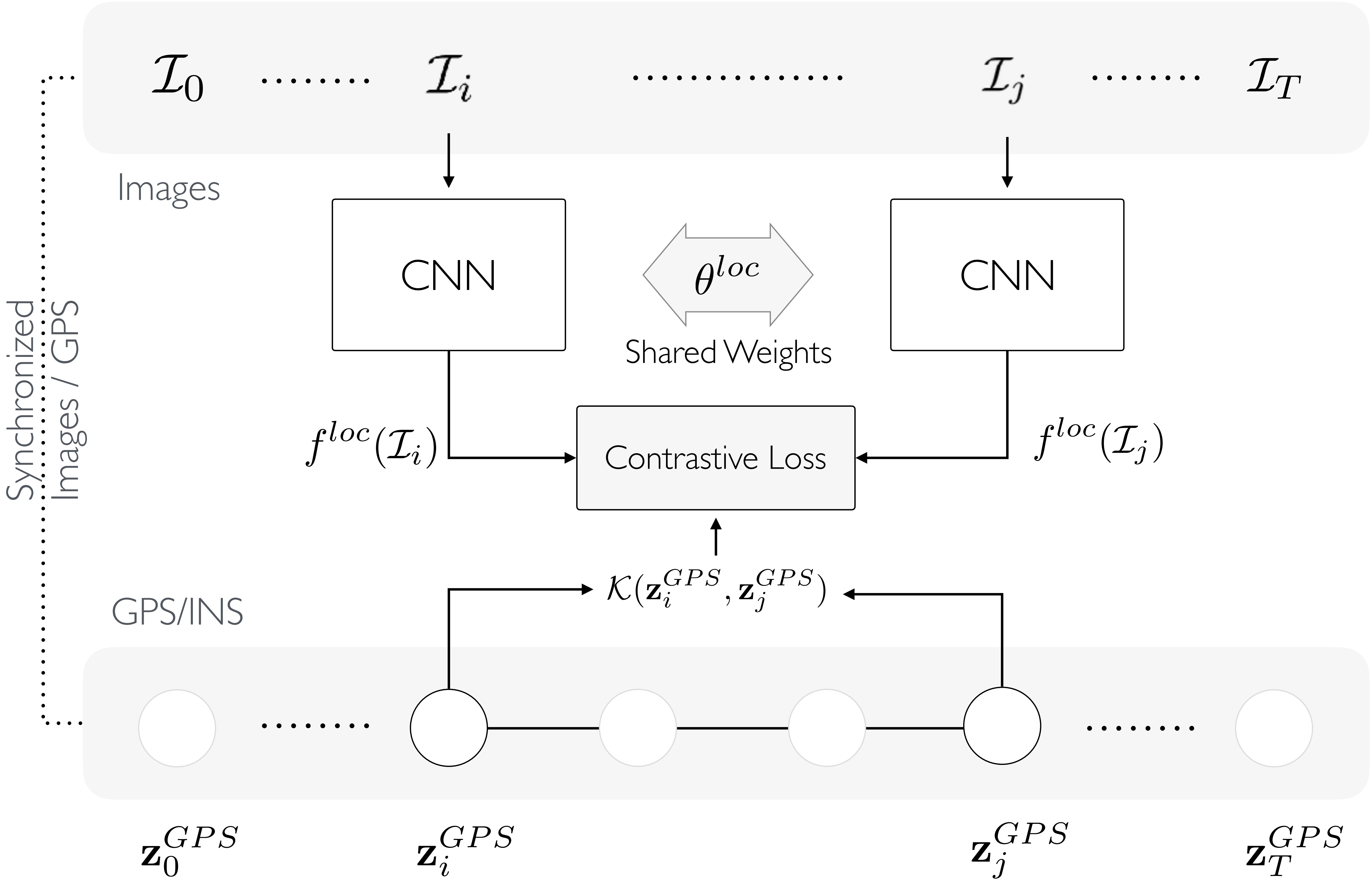
SELF-SUPERVISED METRIC LEARNING FOR VISUAL PLACE-RECOGNITION

- **Bootstrapped Visual Place Recognition Learning for Mobile Robots**
  - Self-supervised Siamese Net with Contrastive Loss
  - Calibrate distances for Loop-Closure Detection
  - Distance-weighted sampling for faster convergence

- **Siamese Place Recognition Model**
  - Pre-trained Places365-AlexNet with shared weights
  - *fc6*, *fc7* are fine-tuned, with remaining layers fixed

Self-Supervised Siamese Net with Contrastive Loss

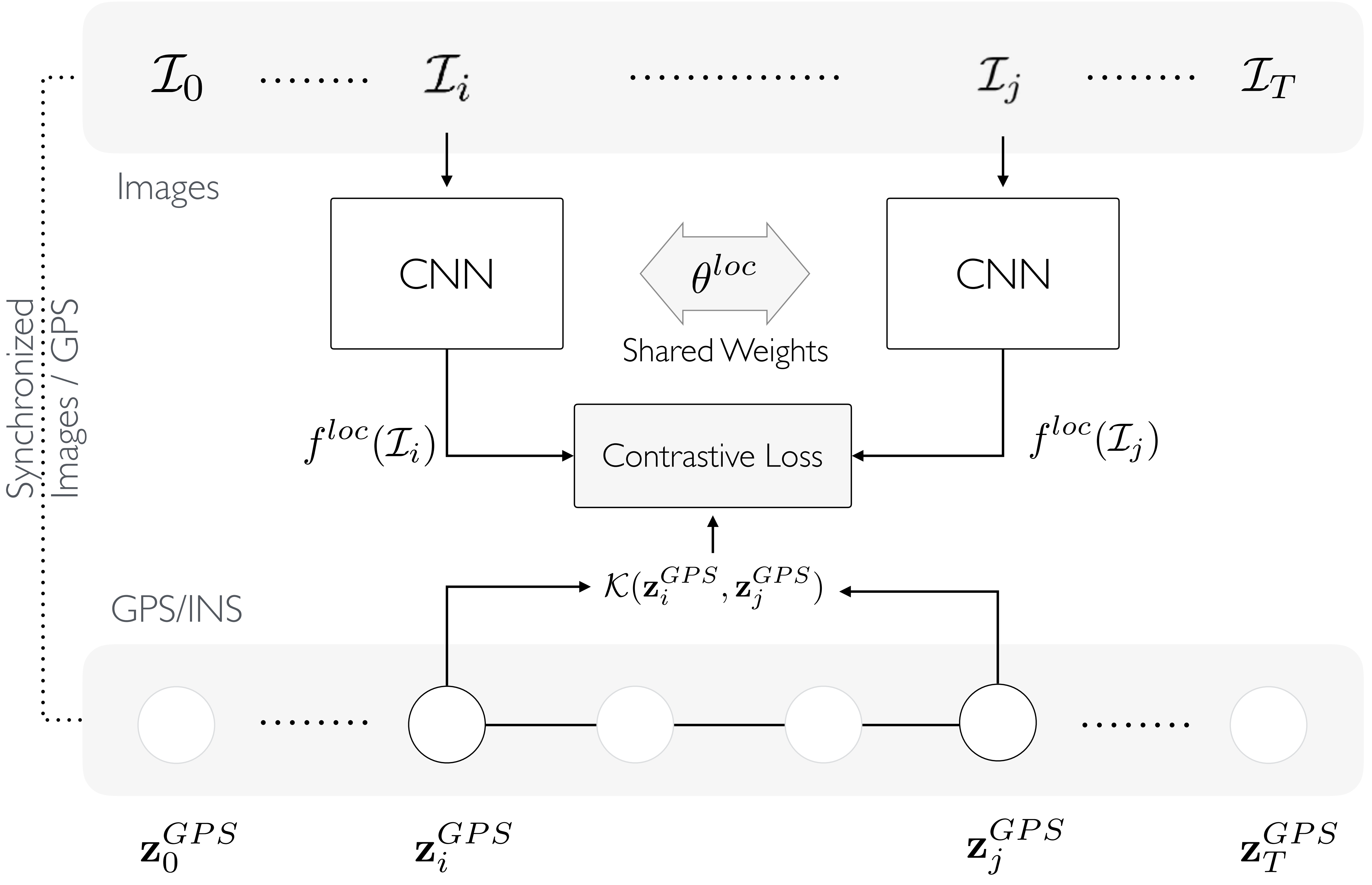
(Bootstrapped with synchronized GPS measurements)



SELF-SUPERVISED METRIC LEARNING FOR VISUAL PLACE-RECOGNITION

- **Bootstrapped Visual Place Recognition Learning for Mobile Robots**
  - Self-supervised Siamese Net with Contrastive Loss
  - Calibrate distances for Loop-Closure Detection
  - Distance-weighted sampling for faster convergence
- **Siamese Place Recognition Model**
  - Pre-trained Places365-AlexNet with shared weights
  - *fc6*, *fc7* are fine-tuned, with remaining layers fixed

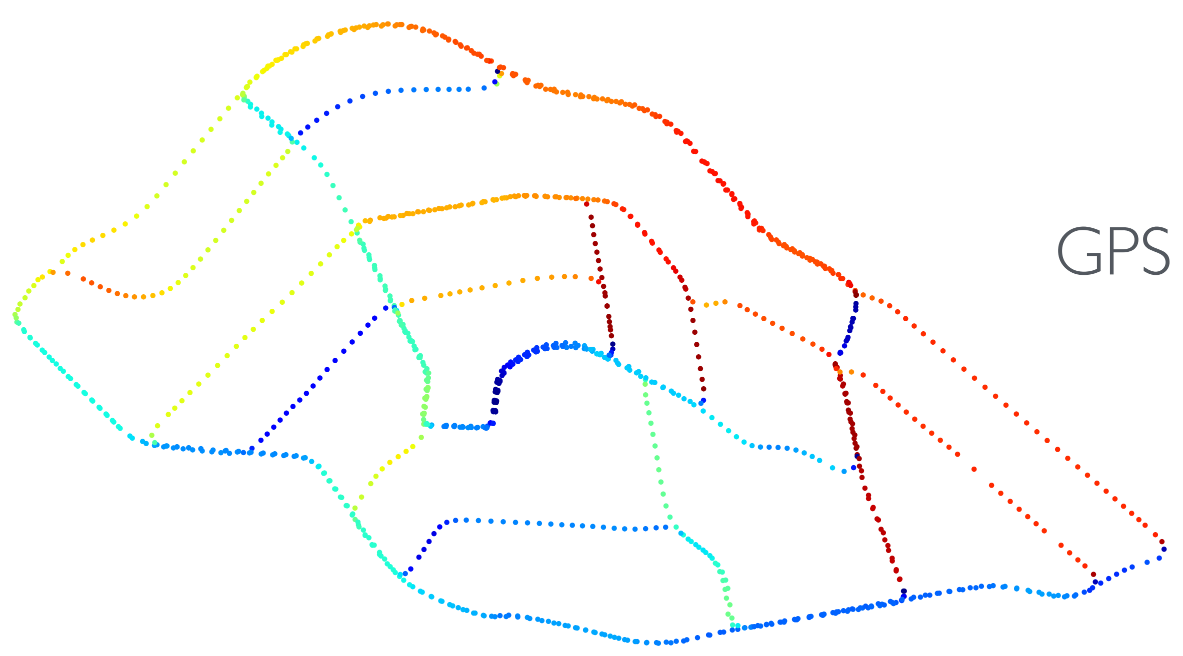
Self-Supervised Siamese Net with Contrastive Loss
(Bootstrapped with synchronized GPS measurements)



Self-Supervised Visual Place-Recognition Learning in Mobile Robots
Pillai et al. (Learning for Localization and Mapping Workshop, IROS 2017)

SELF-SUPERVISED LABELS FOR LOOP-CLOSURES

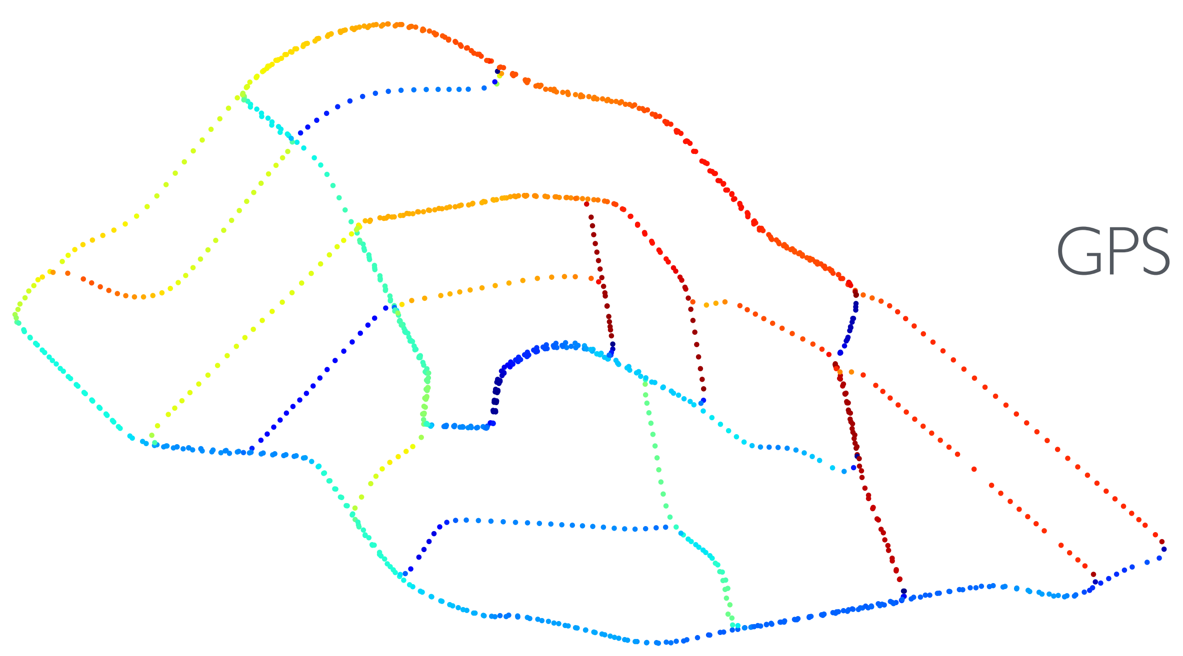
- **Self-supervision via cross-modal information**
  - Self-similarity for sequential pose measurements
  - Kernel with translation and rotational components



St Lucia Dataset
GPS measurements with colors indicating bearing

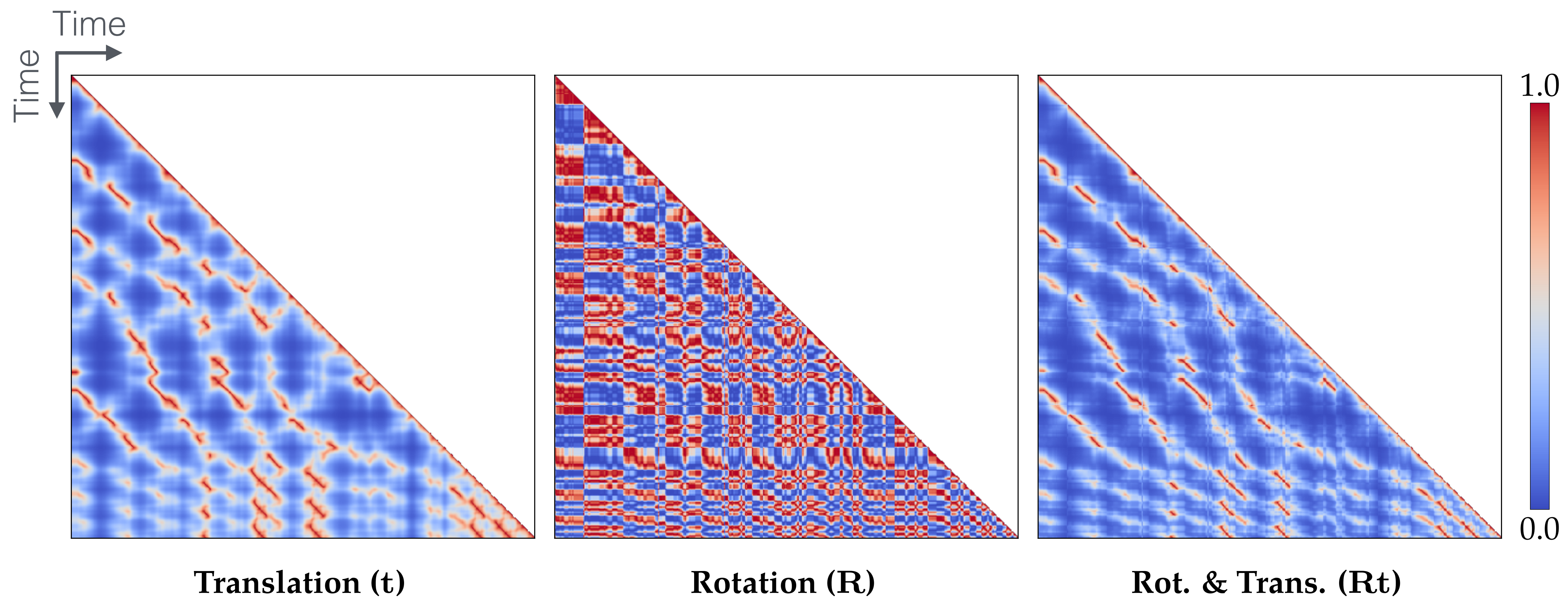
SELF-SUPERVISED LABELS FOR LOOP-CLOSURES

- **Self-supervision via cross-modal information**
  - Self-similarity for sequential pose measurements
  - Kernel with translation and rotational components



St Lucia Dataset
GPS measurements with colors indicating bearing

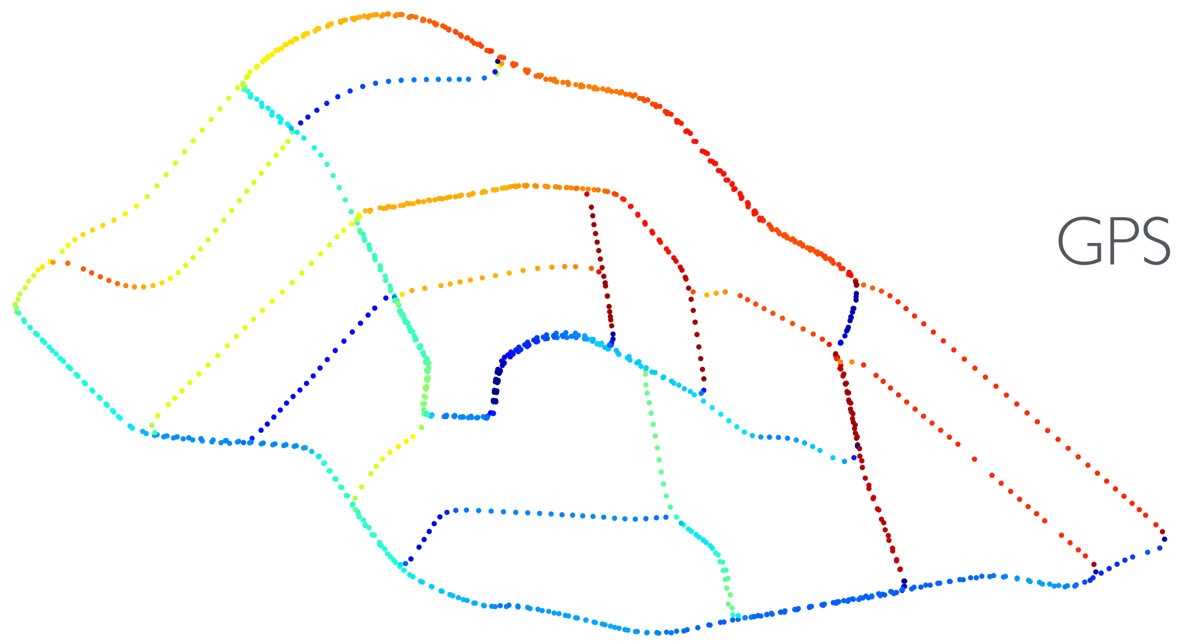
$$\mathcal{K}(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) = \underbrace{\exp(-\gamma^{\mathbf{t}} \left\|\mathbf{z}_i^{\mathbf{t}} - \mathbf{z}_j^{\mathbf{t}}\right\|_2^2)}_{\text{Translation similarity}} \cdot \underbrace{\exp(-\gamma^{\mathbf{R}} \left\|\mathbf{z}_i^{\mathbf{R}} \ominus \mathbf{z}_j^{\mathbf{R}}\right\|_2^2)}_{\text{Rotation similarity}}$$



Self-Similarity
(Kernel derived from GPS measurements)

SELF-SUPERVISED LABELS FOR LOOP-CLOSURES

- **Self-supervision via cross-modal information**
  - Self-similarity for sequential pose measurements
  - Kernel with translation and rotational components

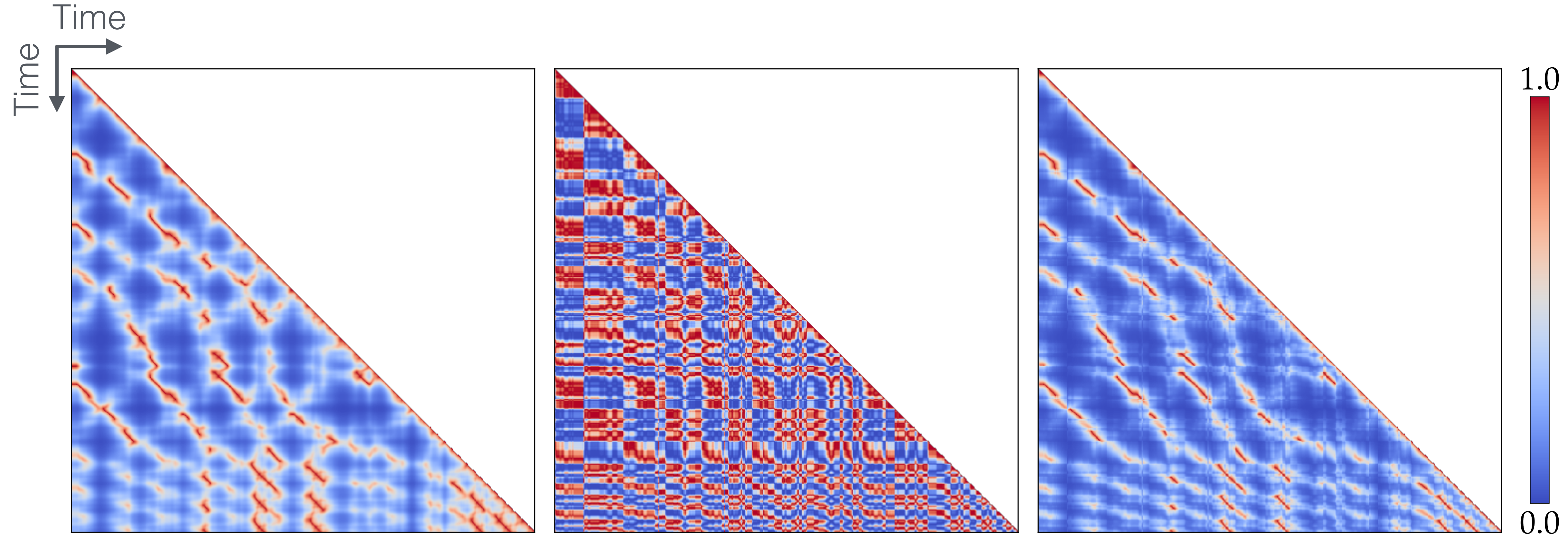


St Lucia Dataset
GPS measurements with colors indicating bearing

$$\mathcal{K}(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) = \underbrace{\exp(-\gamma^{\mathbf{t}} \left\|\mathbf{z}_i^{\mathbf{t}} - \mathbf{z}_j^{\mathbf{t}}\right\|_2^2)}_{\text{Translation similarity}} \cdot \underbrace{\exp(-\gamma^{\mathbf{R}} \left\|\mathbf{z}_i^{\mathbf{R}} \ominus \mathbf{z}_j^{\mathbf{R}}\right\|_2^2)}_{\text{Rotation similarity}}$$

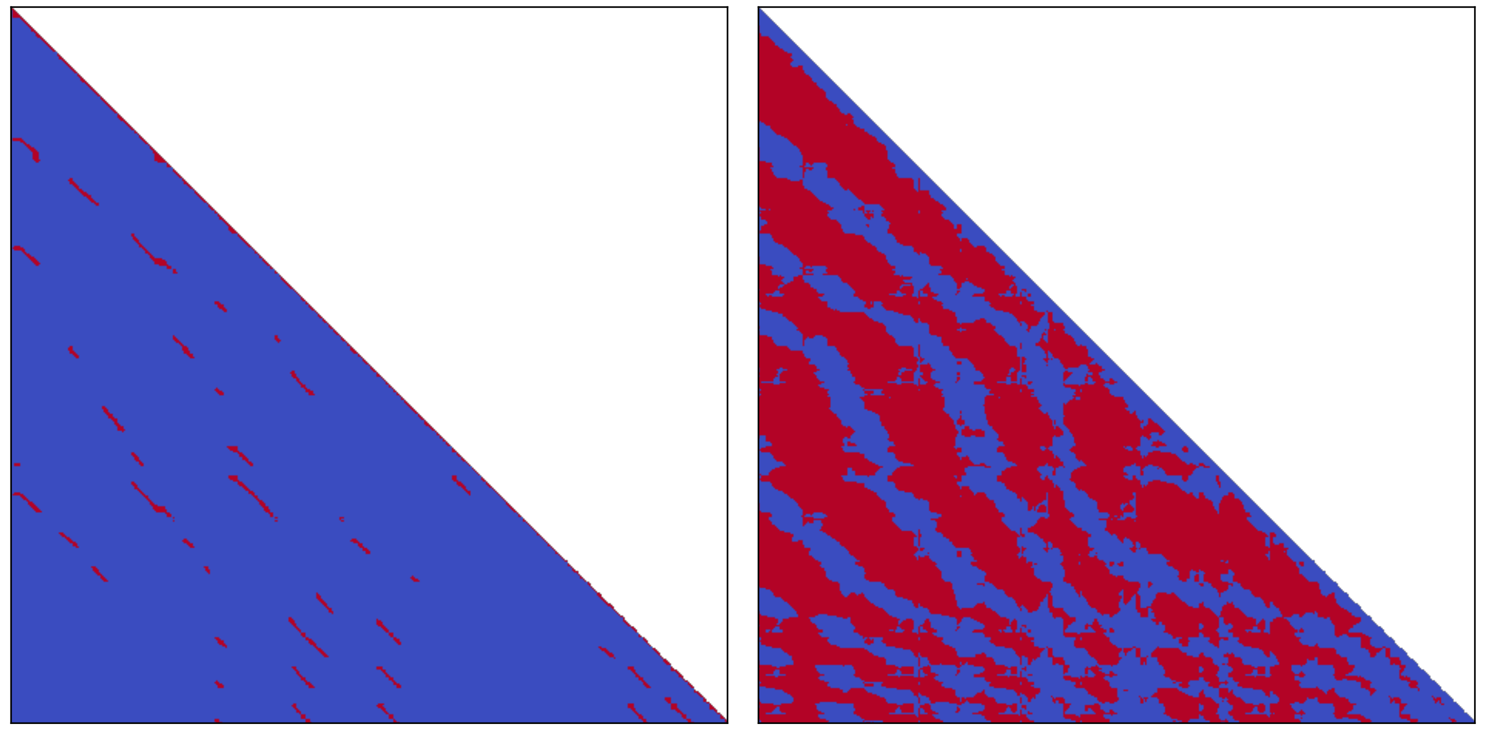
Positive/Negative Indicator

$$\mathbb{1}_{GPS} = \begin{cases} 1 & \text{if } \mathcal{K}(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) > \tau_p^{\mathbf{Rt}} \\ 0 & \text{if } \mathcal{K}(\mathbf{z}_i^{GPS}, \mathbf{z}_j^{GPS}) < \tau_n^{\mathbf{Rt}} \end{cases}$$



Translation (t) **Rotation (R)** **Rot. & Trans. (Rt)**

Self-Similarity
(Kernel derived from GPS measurements)



Positive Labels **Negative Labels**

Self-Supervised Positive/Negative Pairs
(Distance-weighted sampling)

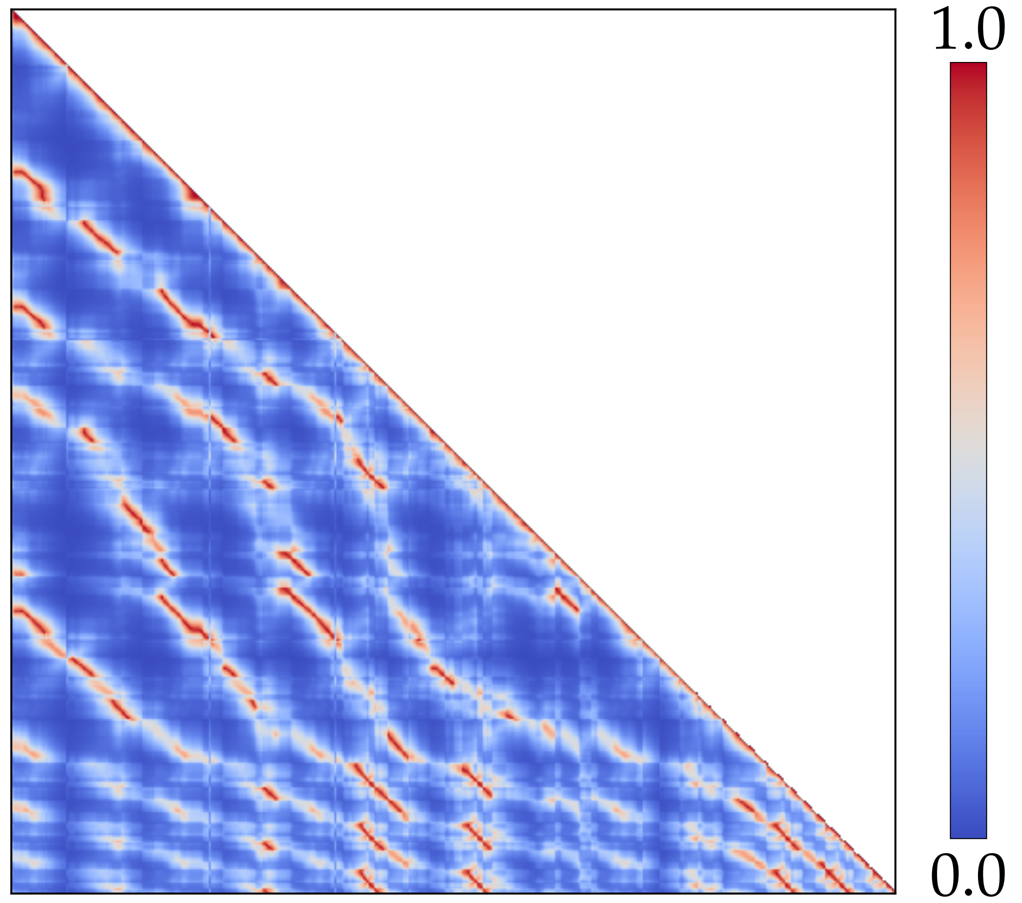
LEARNING A VISUAL-SIMILARITY METRIC

- **Self-supervised metric learning for place-recognition**
  - Calibrate / Fine-tune an appropriate metric for place-recognition
  - Learned embedding can be directly used for loop-closure detection

LEARNING A VISUAL-SIMILARITY METRIC

- **Self-supervised metric learning for place-recognition**
  - Calibrate / Fine-tune an appropriate metric for place-recognition
  - Learned embedding can be directly used for loop-closure detection

GPS Self-Similarity
(Rot & Trans.)



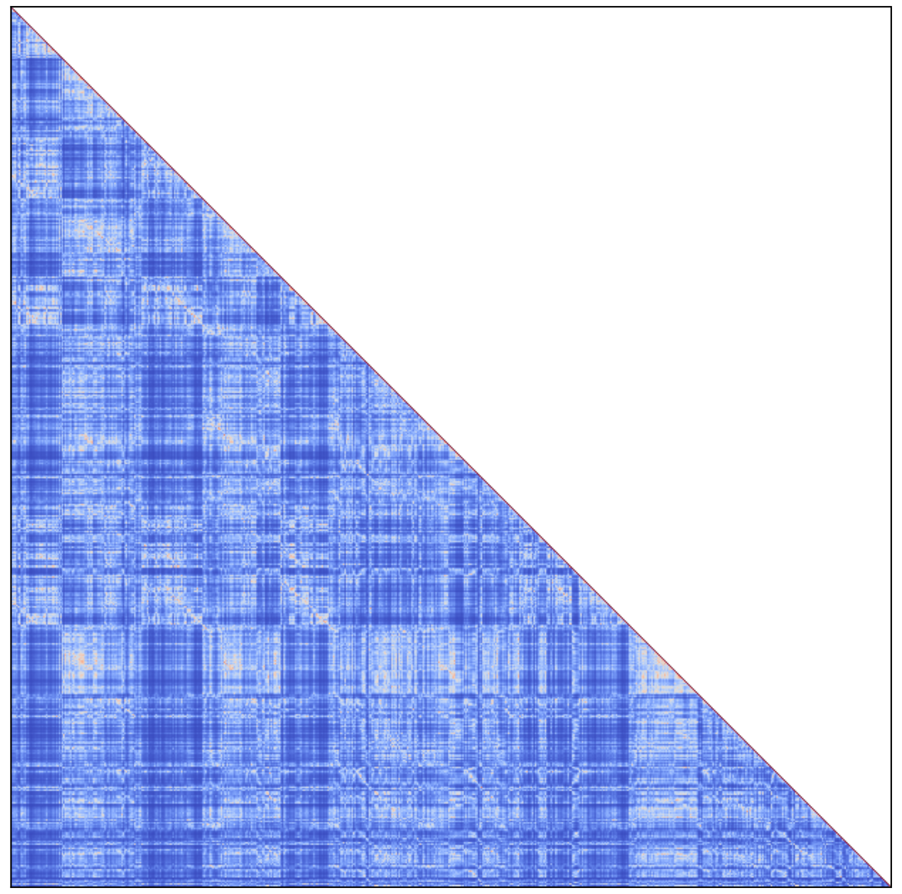
Desired

LEARNING A VISUAL-SIMILARITY METRIC

- **Self-supervised metric learning for place-recognition**
  - Calibrate / Fine-tune an appropriate metric for place-recognition
  - Learned embedding can be directly used for loop-closure detection

Image Self-Similarity

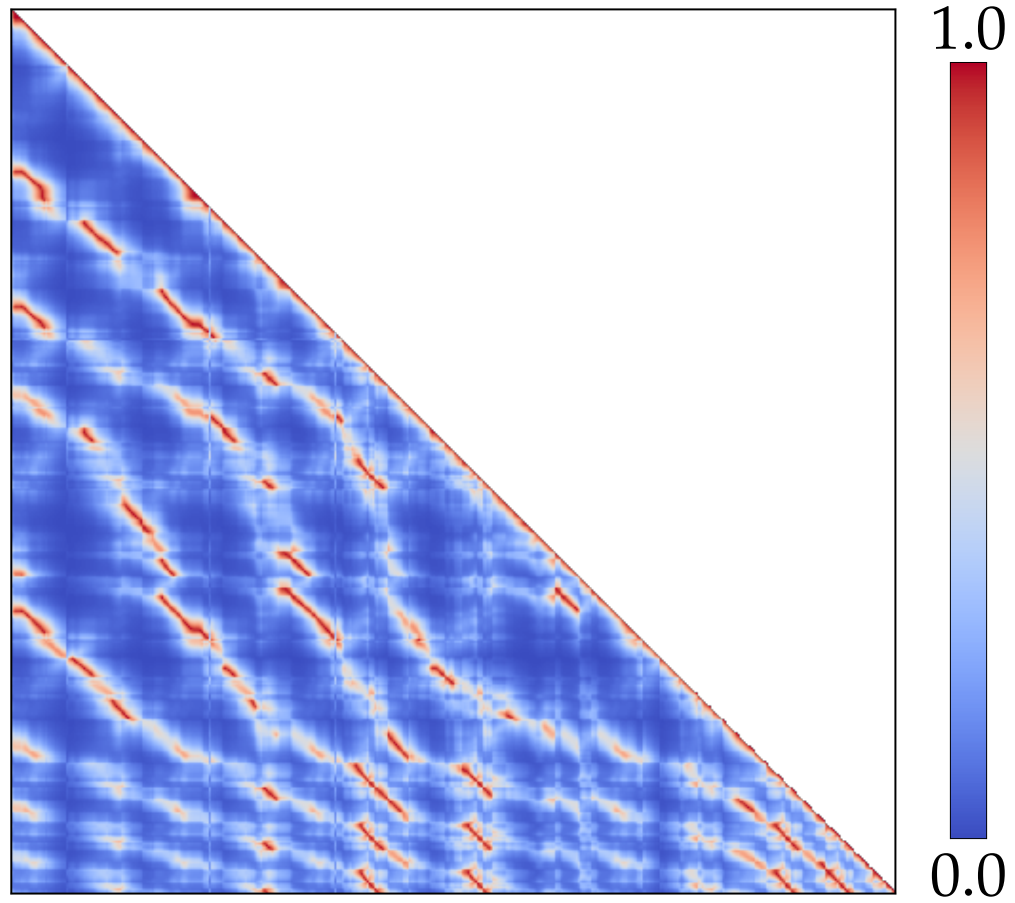
$$D(\mathcal{I}_i, \mathcal{I}_j) = \left\| f^{loc}(\mathcal{I}_i) - f^{loc}(\mathcal{I}_j) \right\|_2$$



Epoch 0

GPS Self-Similarity

(Rot & Trans.)



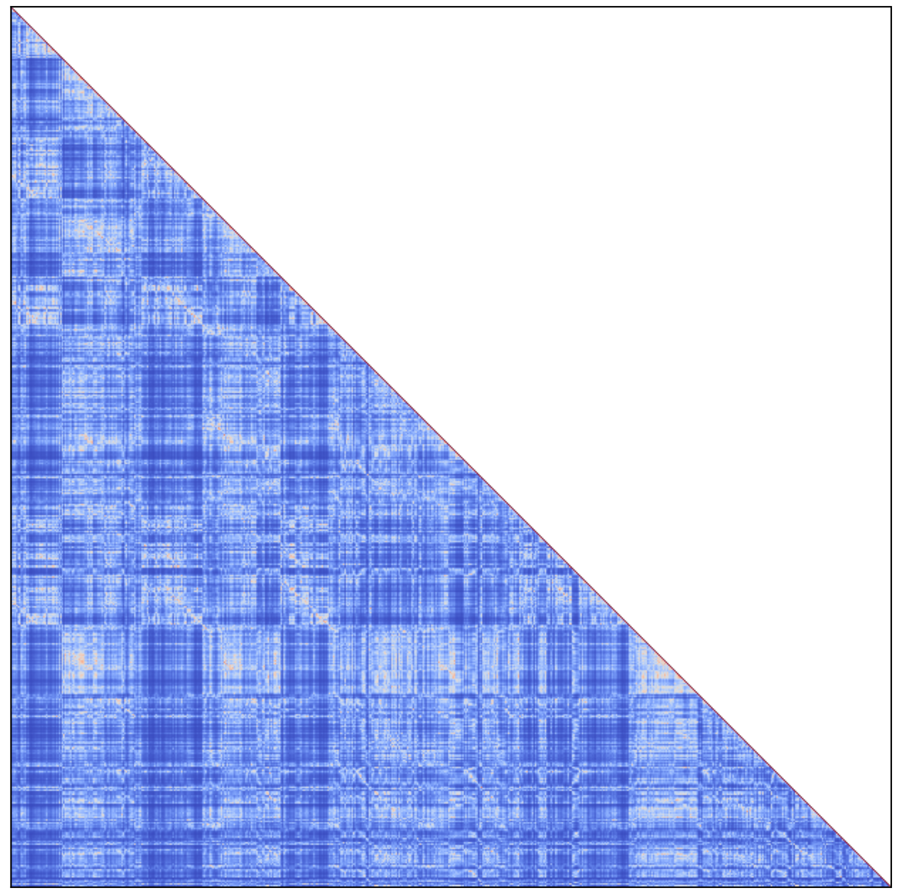
Desired

LEARNING A VISUAL-SIMILARITY METRIC

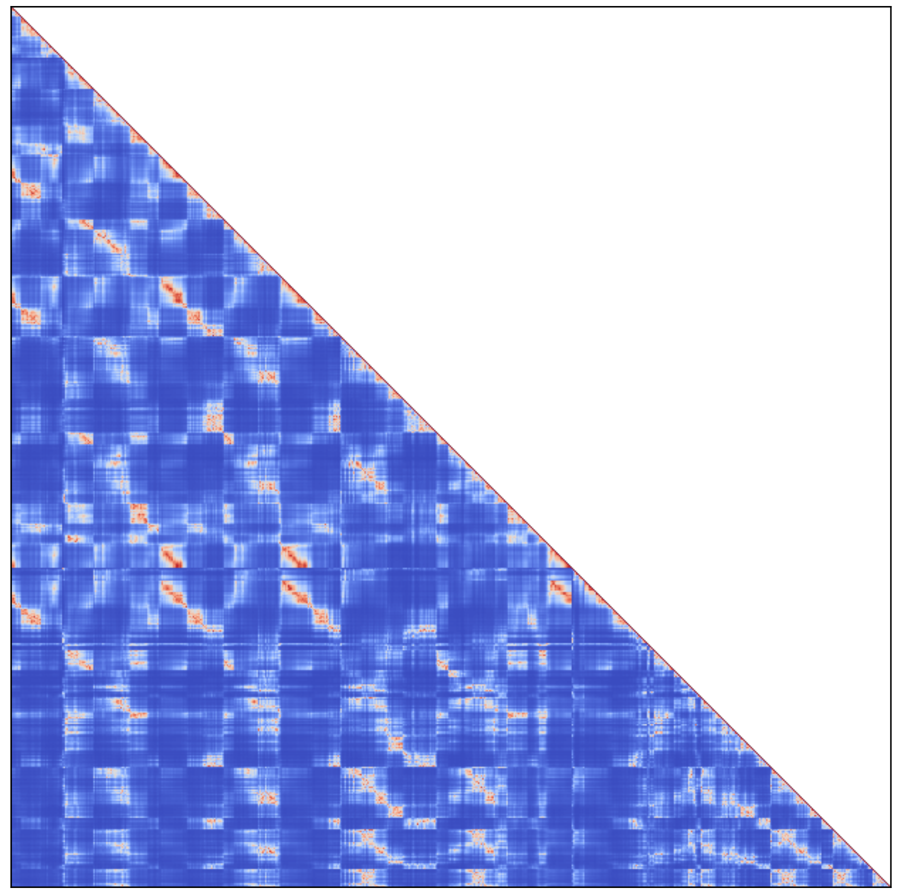
- **Self-supervised metric learning for place-recognition**
  - Calibrate / Fine-tune an appropriate metric for place-recognition
  - Learned embedding can be directly used for loop-closure detection

Image Self-Similarity

$D(\mathcal{I}_i, \mathcal{I}_j) = \left\| f^{loc}(\mathcal{I}_i) - f^{loc}(\mathcal{I}_j) \right\|_2$



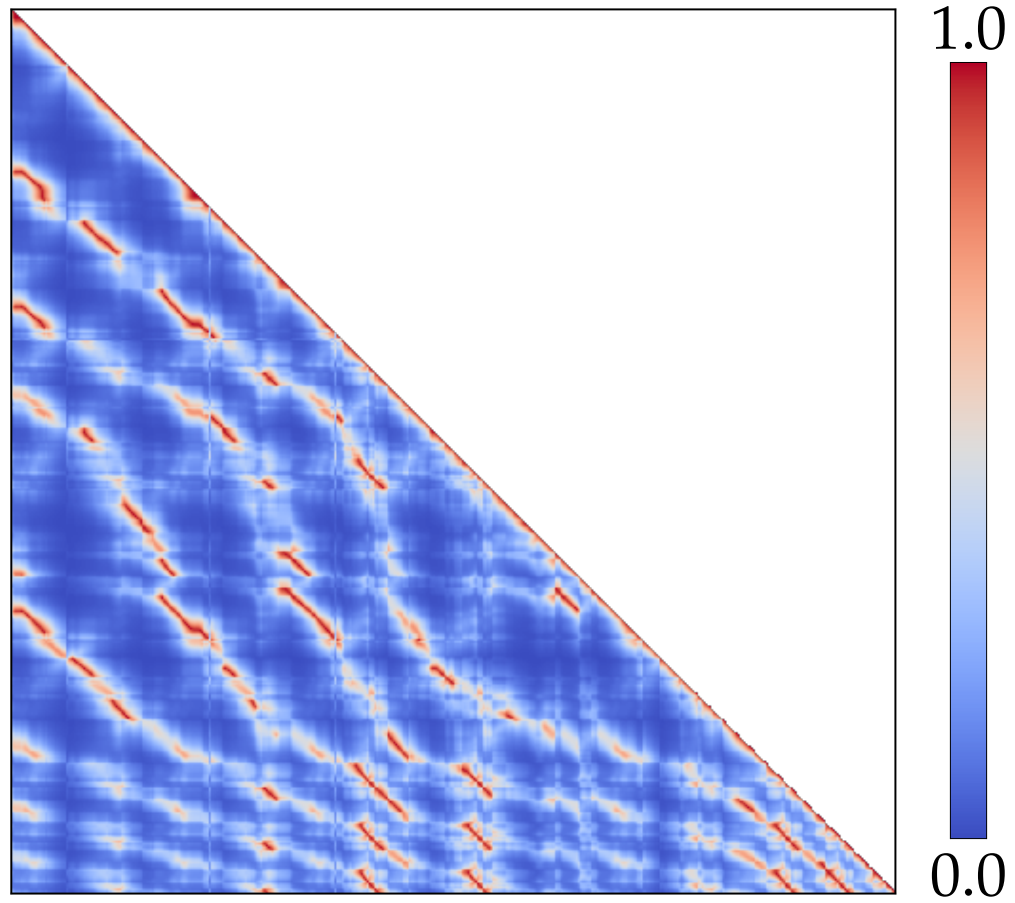
Epoch 0



Epoch 30

GPS Self-Similarity

(Rot & Trans.)



Desired

Self-supervised learning of a visual-similarity metric

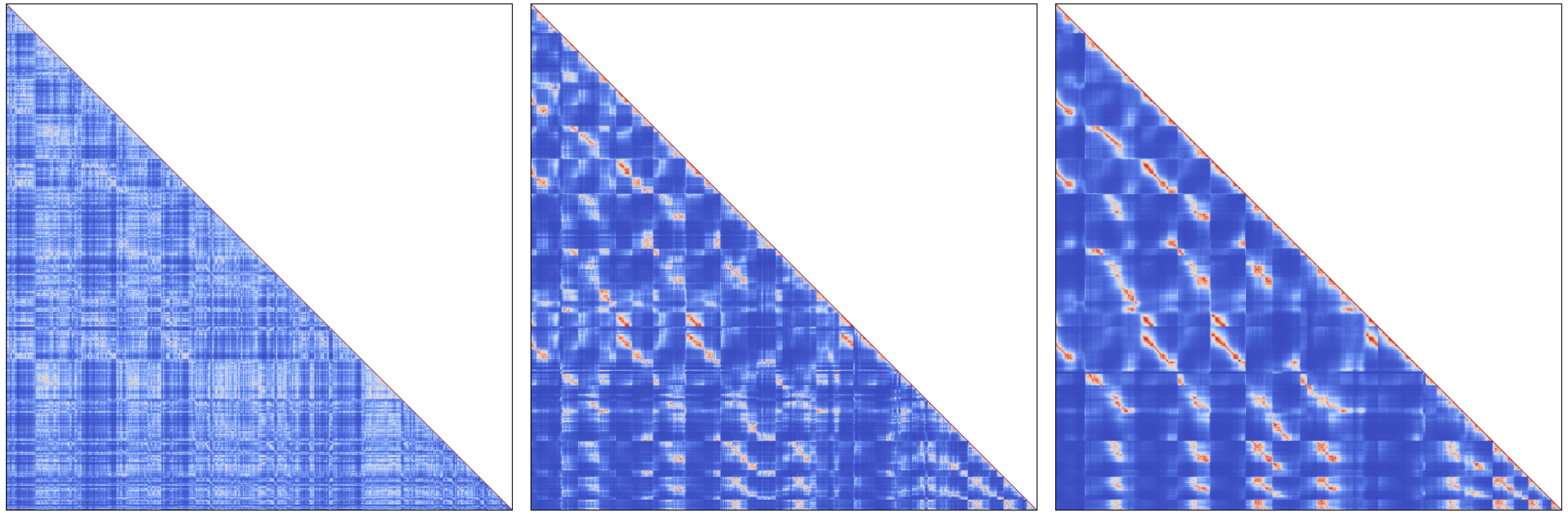
(Learning evolution)

LEARNING A VISUAL-SIMILARITY METRIC

- **Self-supervised metric learning for place-recognition**
  - Calibrate / Fine-tune an appropriate metric for place-recognition
  - Learned embedding can be directly used for loop-closure detection

Image Self-Similarity

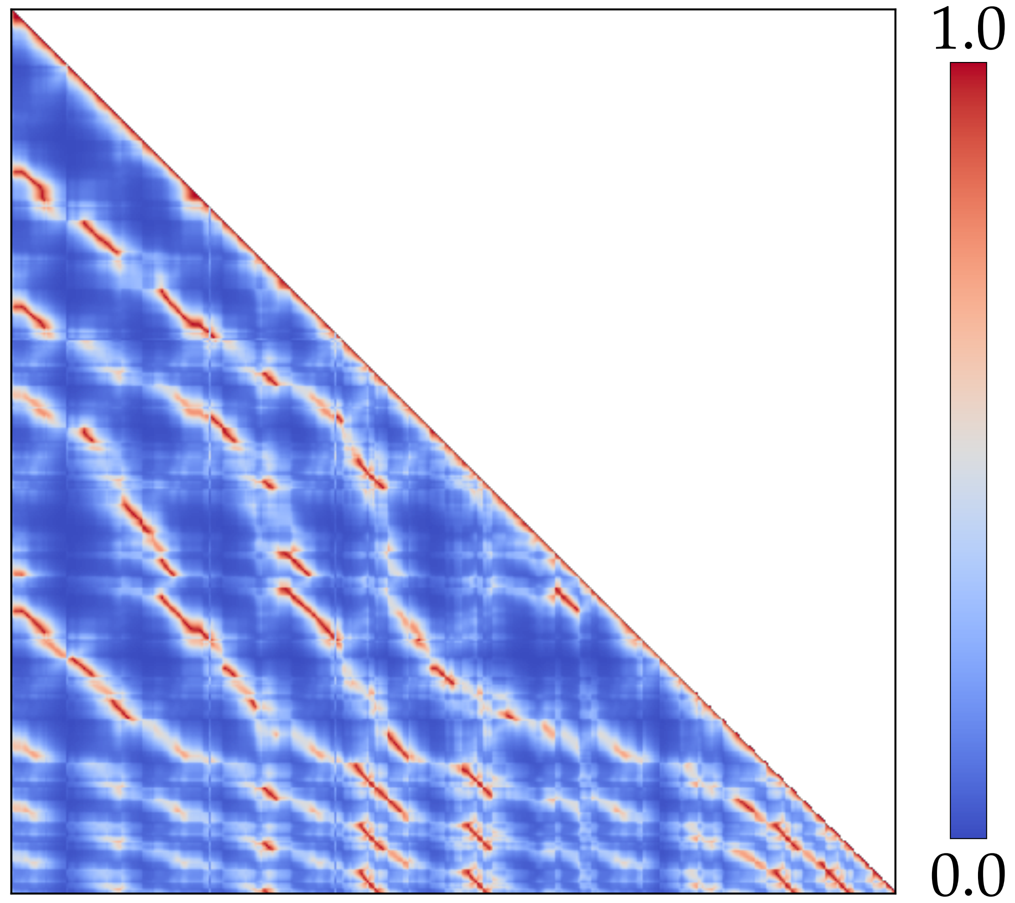
$D(\mathcal{I}_i, \mathcal{I}_j) = \left\| f^{loc}(\mathcal{I}_i) - f^{loc}(\mathcal{I}_j) \right\|_2$



Epoch 0 Epoch 30 Epoch 180

Self-supervised learning of a visual-similarity metric
(Learning evolution)

GPS Self-Similarity
(Rot & Trans.)



Desired

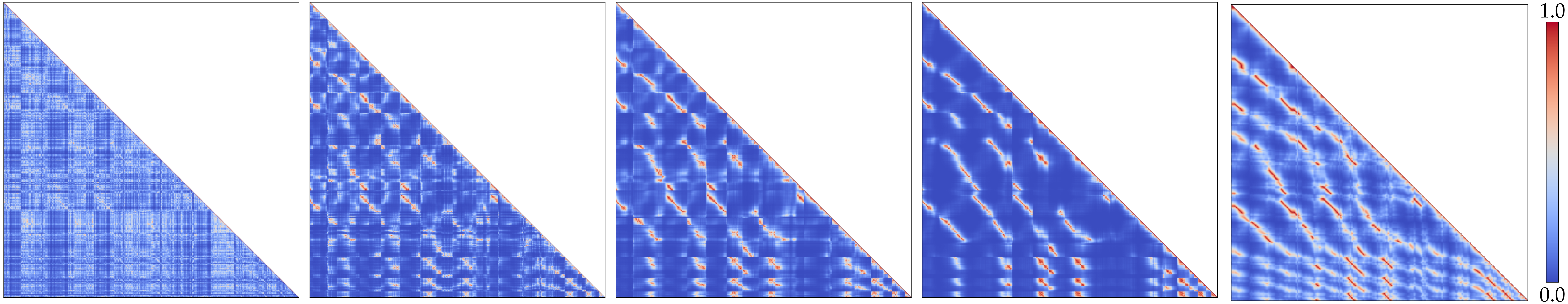
LEARNING A VISUAL-SIMILARITY METRIC

- **Self-supervised metric learning for place-recognition**
  - Calibrate / Fine-tune an appropriate metric for place-recognition
  - Learned embedding can be directly used for loop-closure detection

Image Self-Similarity

$$D(\mathcal{I}_i, \mathcal{I}_j) = \left\| f^{loc}(\mathcal{I}_i) - f^{loc}(\mathcal{I}_j) \right\|_2$$

GPS Self-Similarity
(Rot & Trans.)

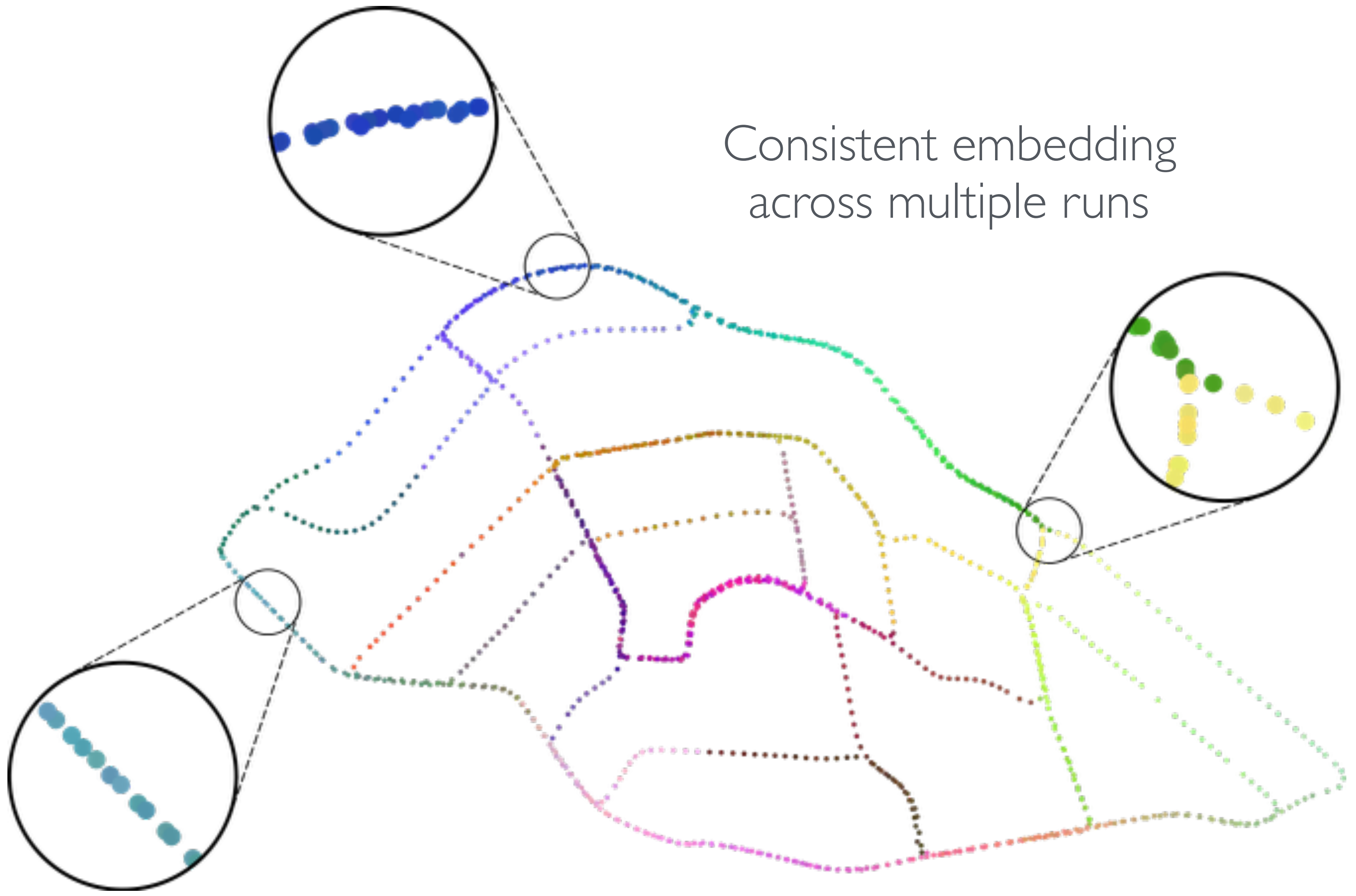


Epoch 0 | Epoch 30 | Epoch 180 | Final Epoch | Desired

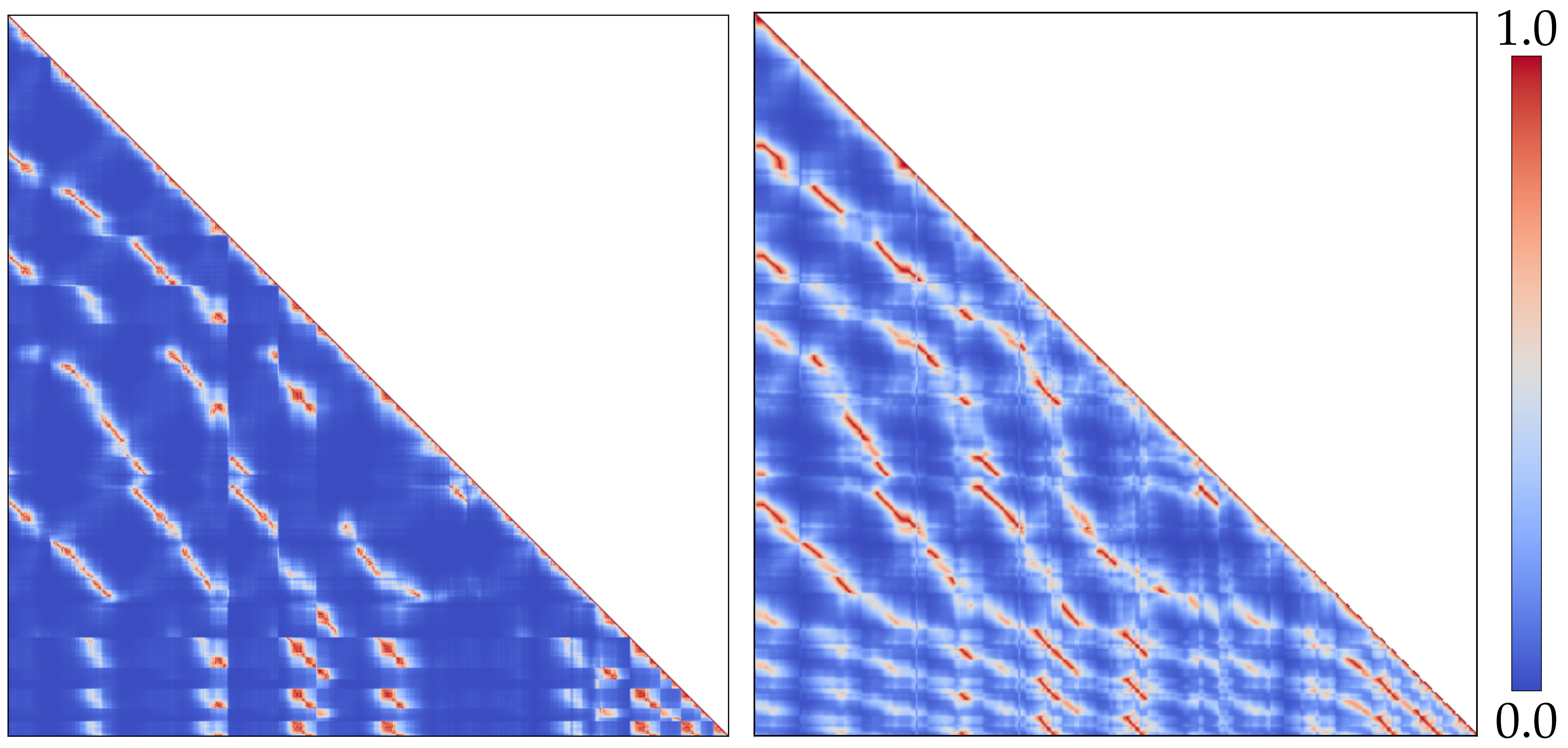
Self-supervised learning of a visual-similarity metric
(Learning evolution)

LEARNING A VISUAL-SIMILARITY METRIC

- **Self-supervised metric learning for place-recognition**
  - Calibrate / Fine-tune an appropriate metric for place-recognition
  - Learned embedding can be directly used for loop-closure detection



Trajectory with embedded CNN features
Colored with T-SNE
(St. Lucia Dataset)

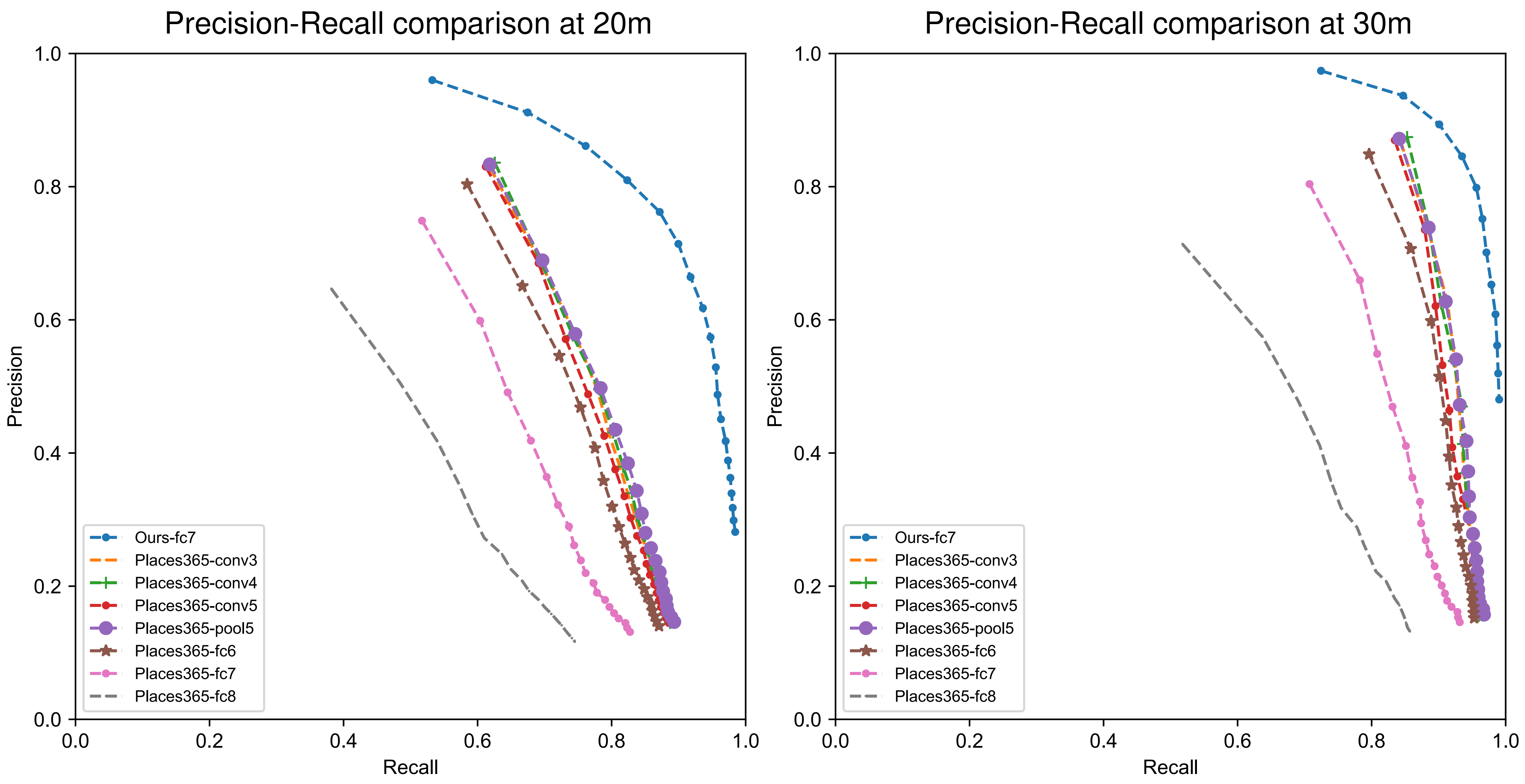


Learned Self-Similarity
(Image embedding)

Target Self-Similarity
(GPS measurements)

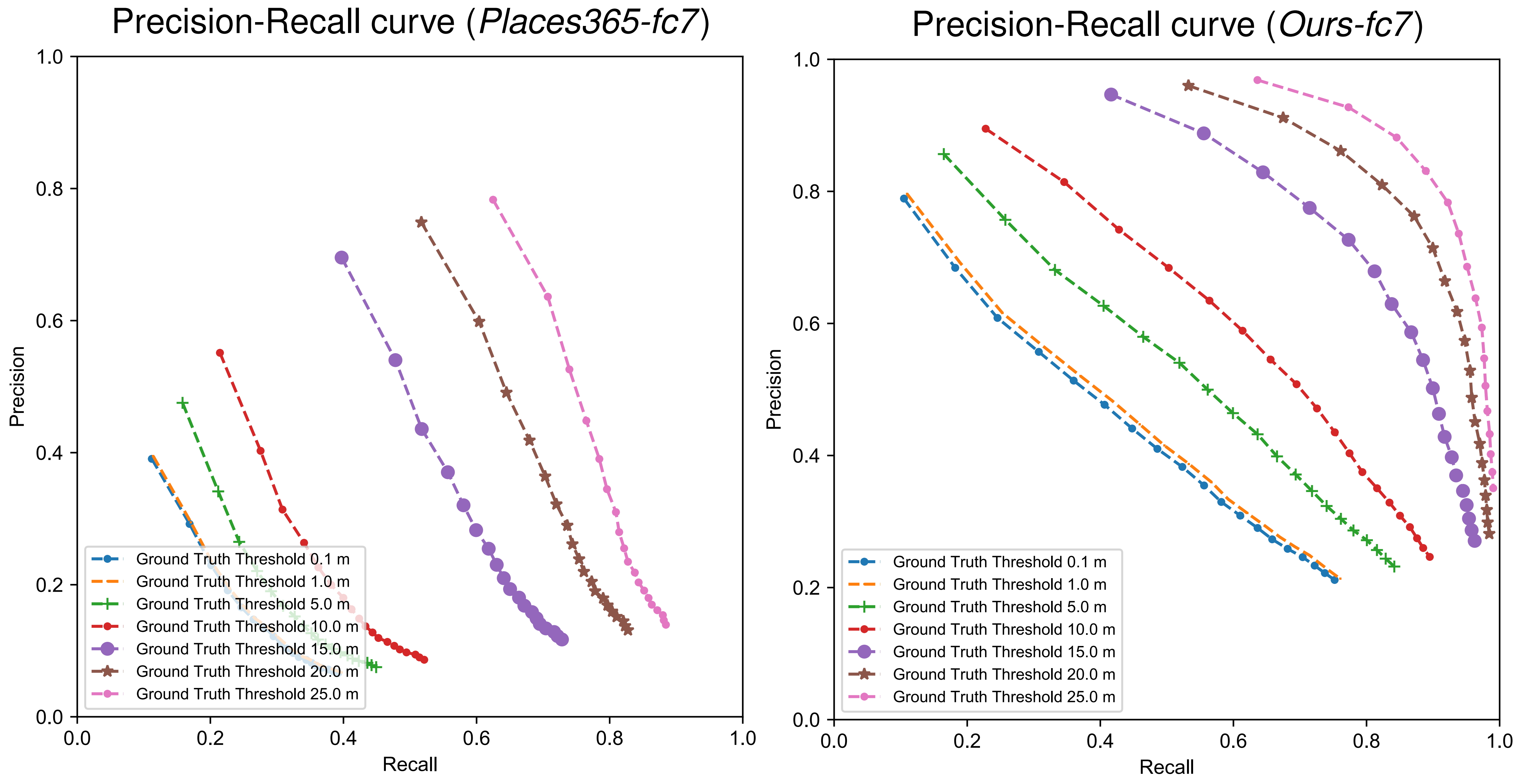
SELF-SUPERVISED PLACE RECOGNITION PERFORMANCE

Precision-Recall for Loop-Closure Recognition
(Comparing ***Places365 AlexNet layers*** and ***Ours-fc7*** learned embedding)



(k-NN: Considering top 20 nearest neighbors)

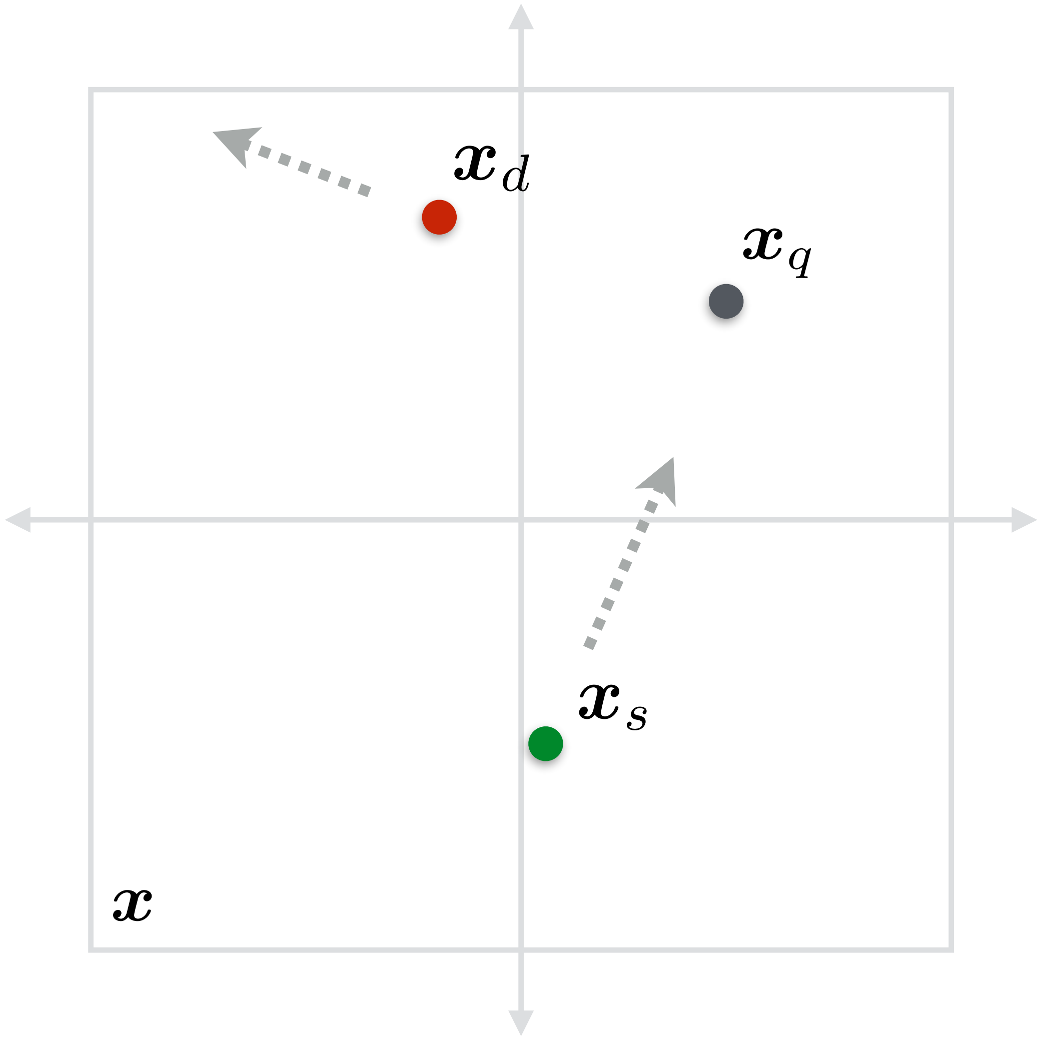
Precision-Recall for Loop-Closure Recognition
(Comparing ***Places365 AlexNet fc7*** and ***Ours-fc7*** learned embedding)



(k-NN: Considering top 20 nearest neighbors)

SELF-SUPERVISED PLACE RECOGNITION PERFORMANCE

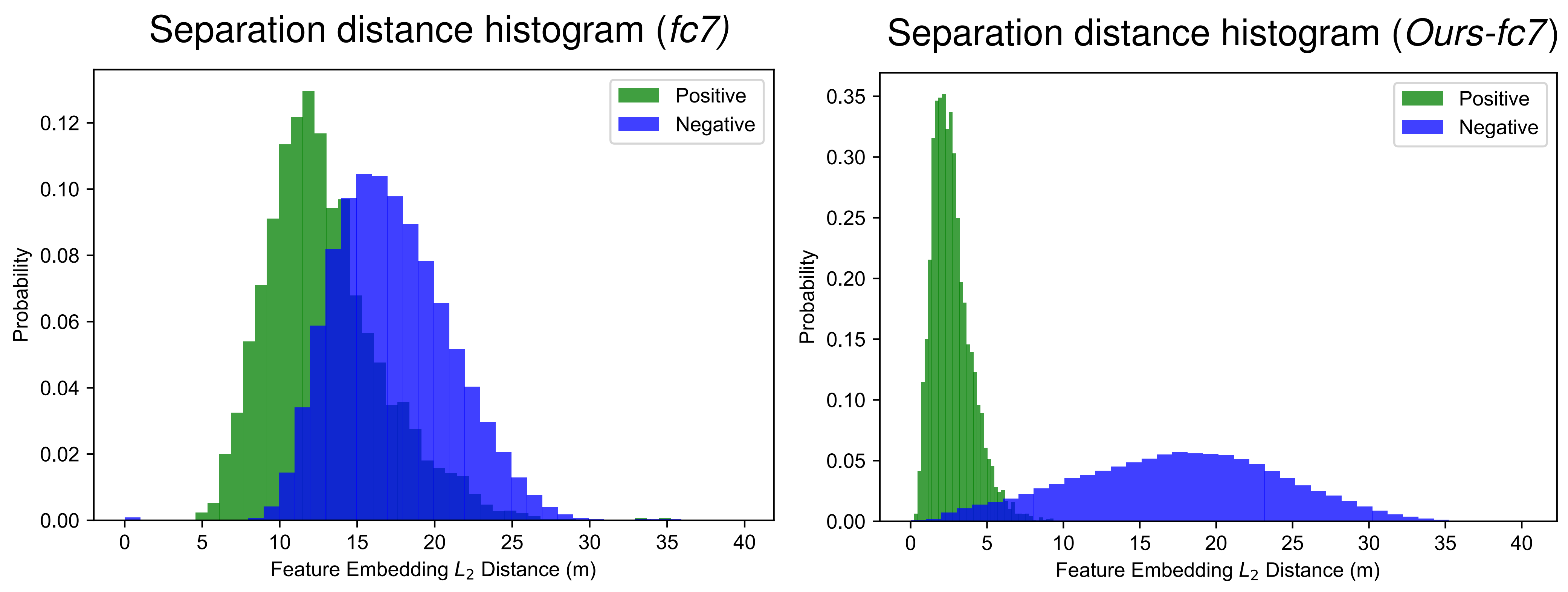
Arbitrarily-defined Distance Measure
(Meaningless)



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$$

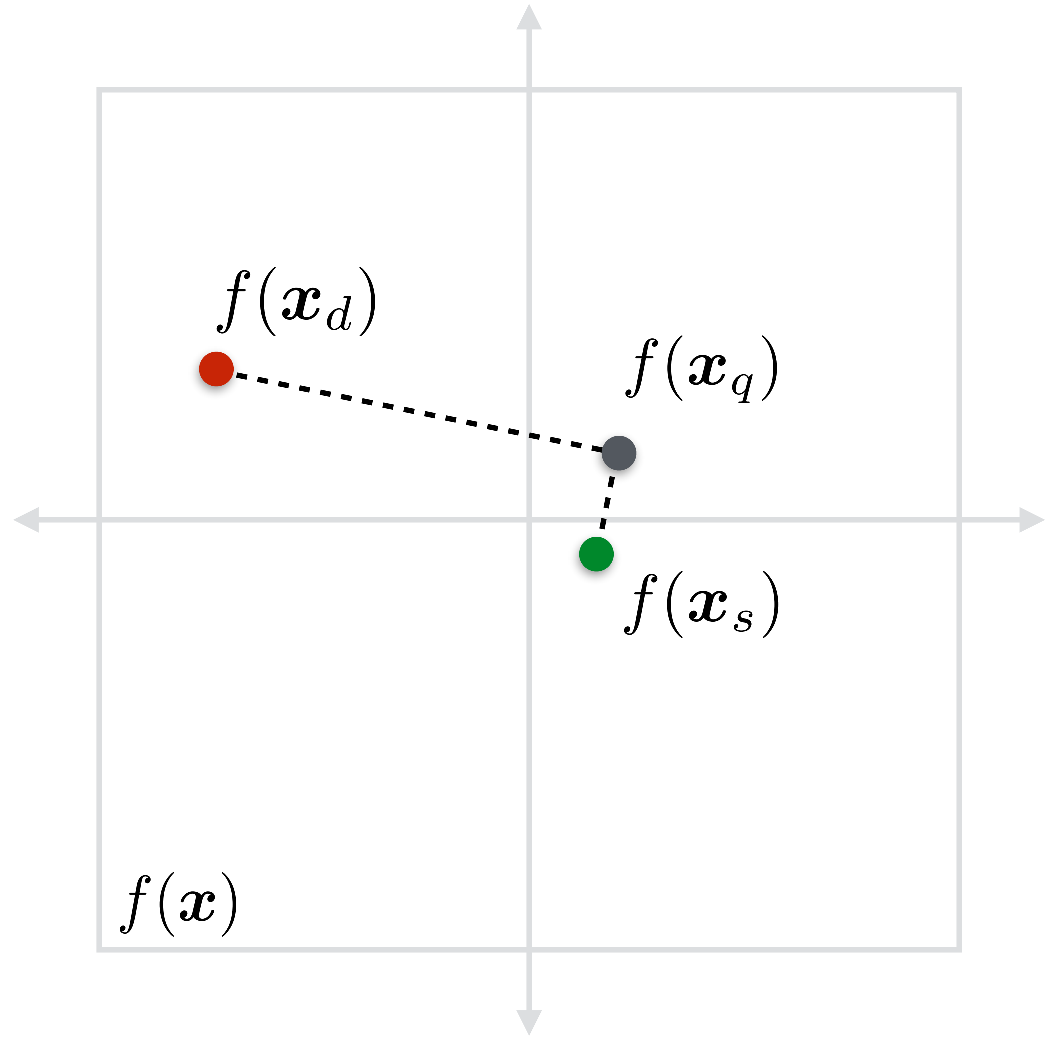
Separation Distance Calibration

(Comparing ***Places365 AlexNet fc7*** and ***Ours-fc7*** learned embedding)



Plot shows the histograms of L2 distances between similar and dissimilar examples. The distances are well-separated in the learned embedding.

"Semantic" Distance Measure
(Task appropriate)



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f(\boldsymbol{x}_i; \theta) - f(\boldsymbol{x}_j; \theta)\|_2$$

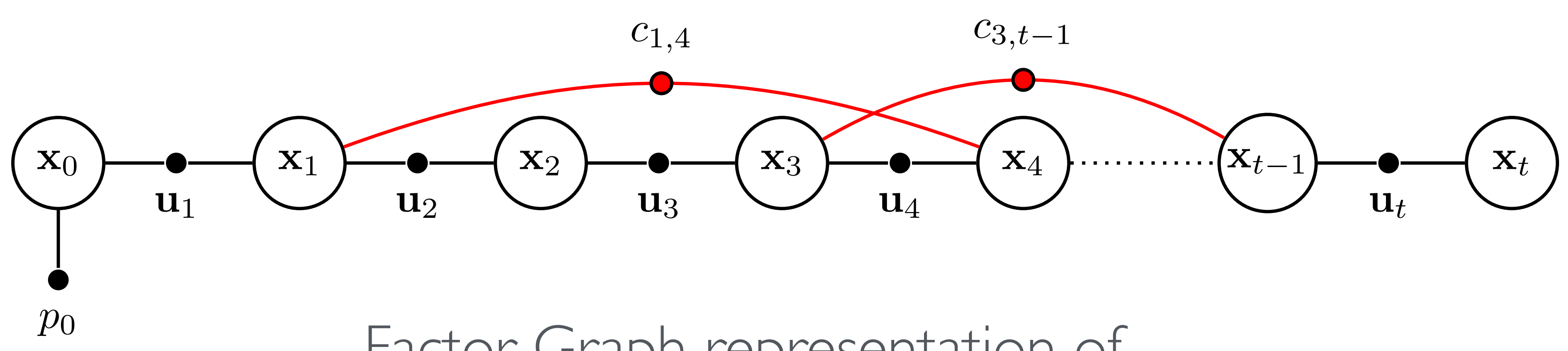
SELF-SUPERVISED LOOP-CLOSURE DETECTION

- **Learned similarity metric for loop-closure detection**
  - Fixed-radius NN on learned embedding reduces false positives
  - Fine-tuning only requires collecting data
  - Works with any real-valued descriptor
  - Learned embedding can be used for indexing, querying, quantization

SELF-SUPERVISED LOOP-CLOSURE DETECTION

- **Learned similarity metric for loop-closure detection**
  - Fixed-radius NN on learned embedding reduces false positives
  - Fine-tuning only requires collecting data
  - Works with any real-valued descriptor
  - Learned embedding can be used for indexing, querying, quantization

- **Vision-based Pose-Graph SLAM**
  - Self-supervised loop-closure identification with learned embedding

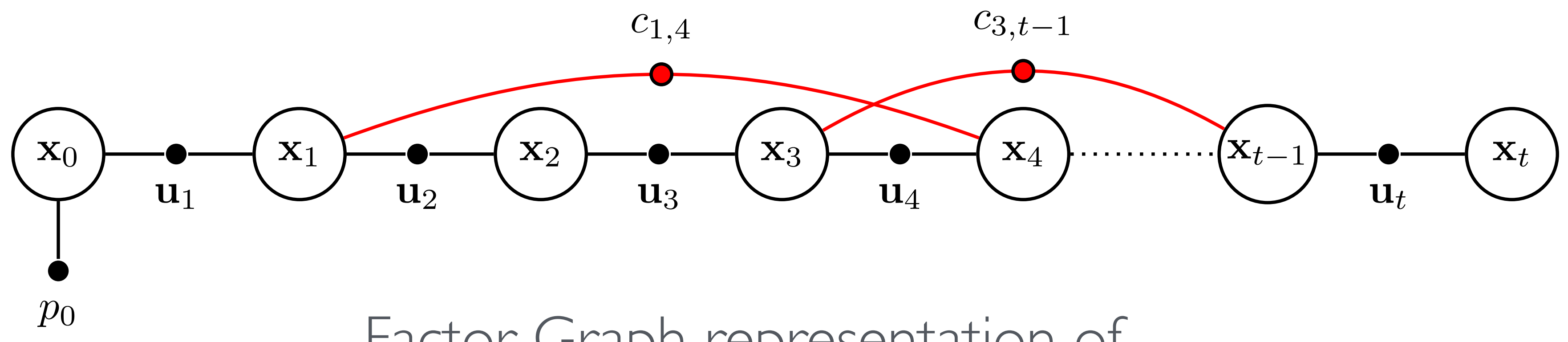


Factor Graph representation of Pose-Graph SLAM

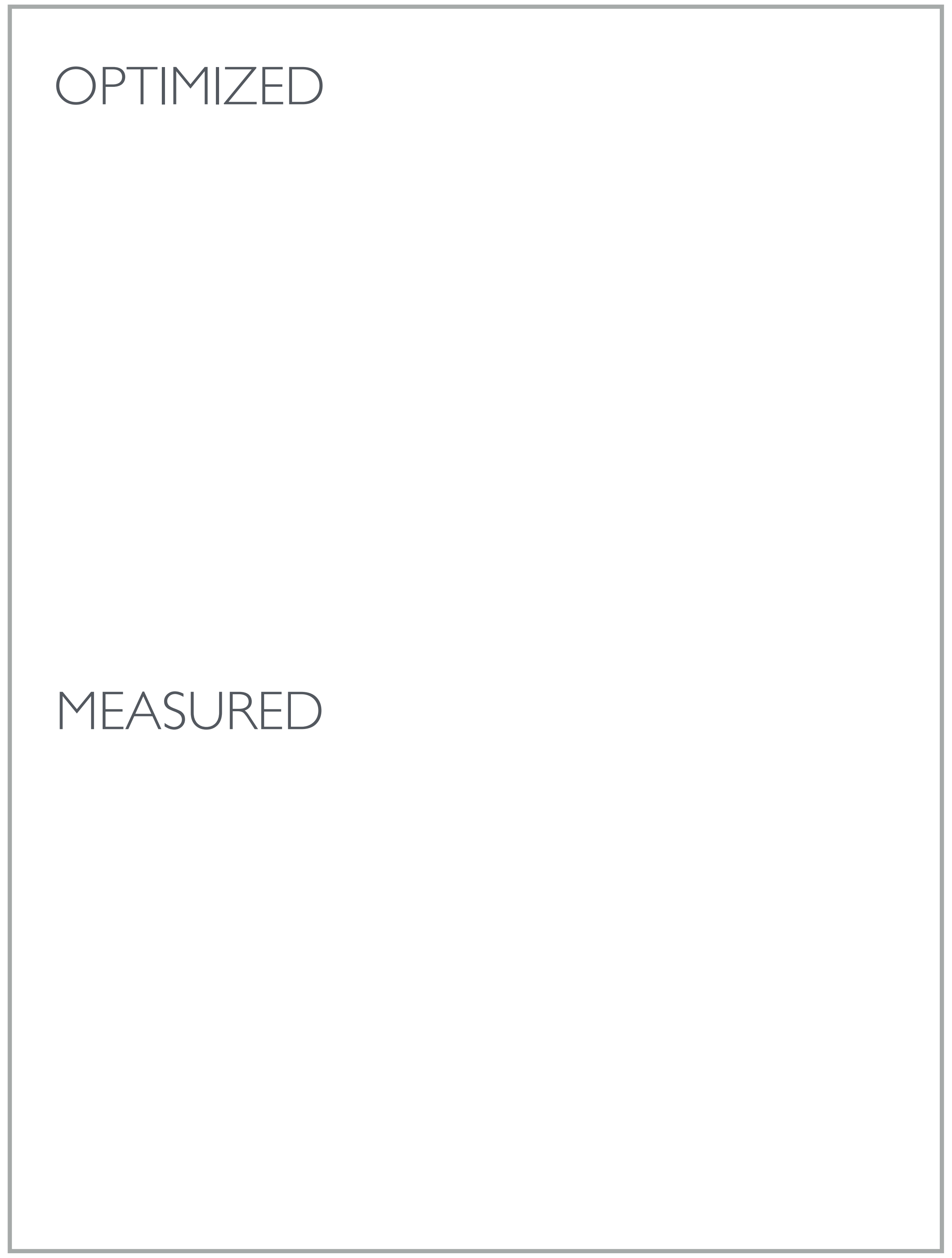
SELF-SUPERVISED LOOP-CLOSURE DETECTION

- **Learned similarity metric for loop-closure detection**
  - Fixed-radius NN on learned embedding reduces false positives
  - Fine-tuning only requires collecting data
  - Works with any real-valued descriptor
  - Learned embedding can be used for indexing, querying, quantization

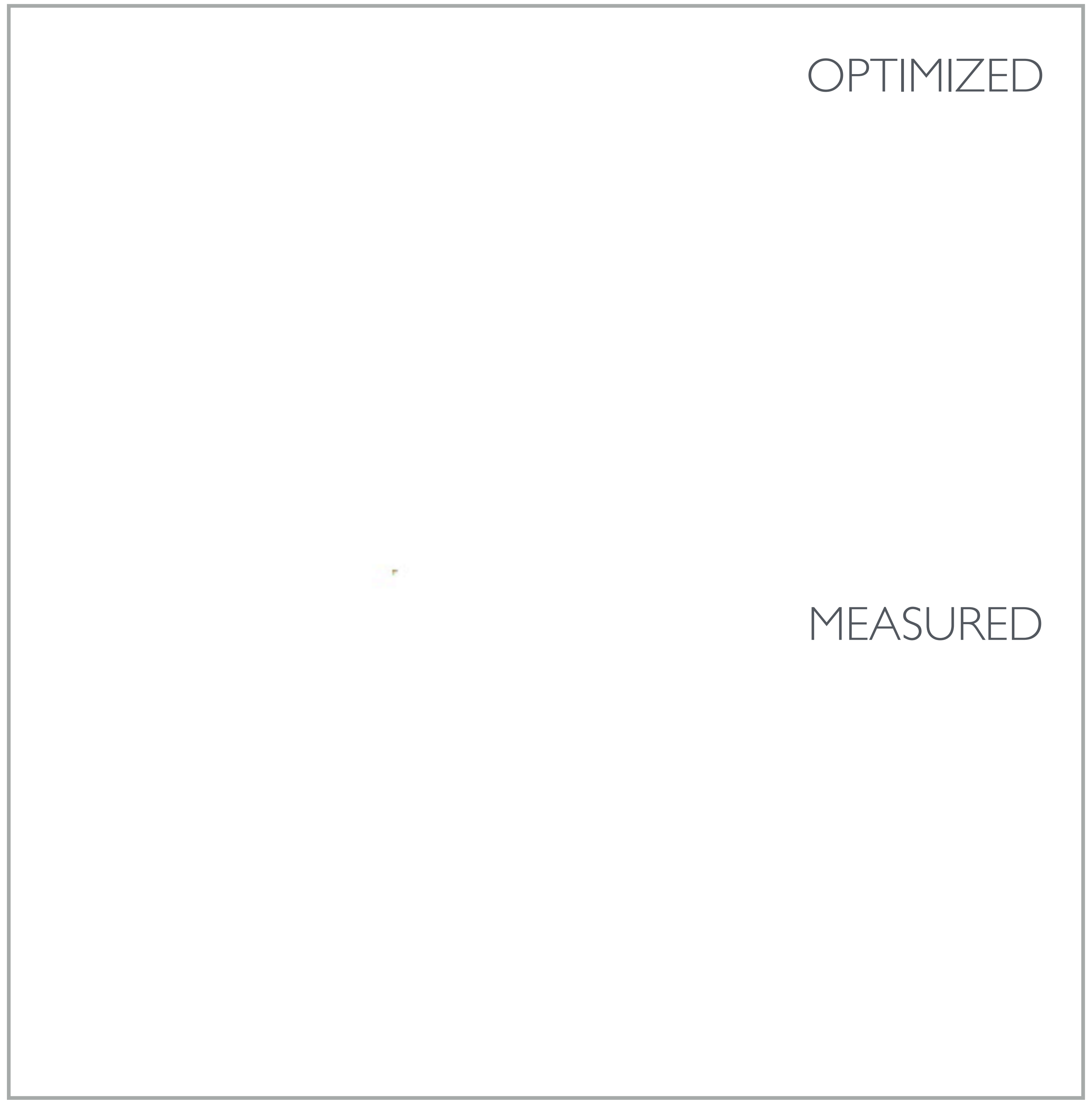
- **Vision-based Pose-Graph SLAM**
  - Self-supervised loop-closure identification with learned embedding



Factor Graph representation of Pose-Graph SLAM



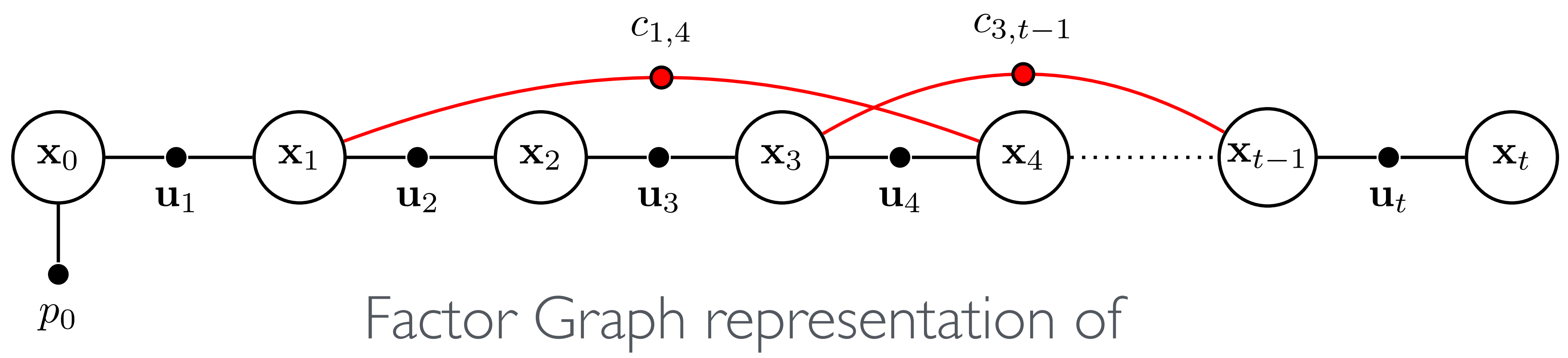
KITTI Dataset
(With learned metric)



St. Lucia Dataset
(With learned metric)

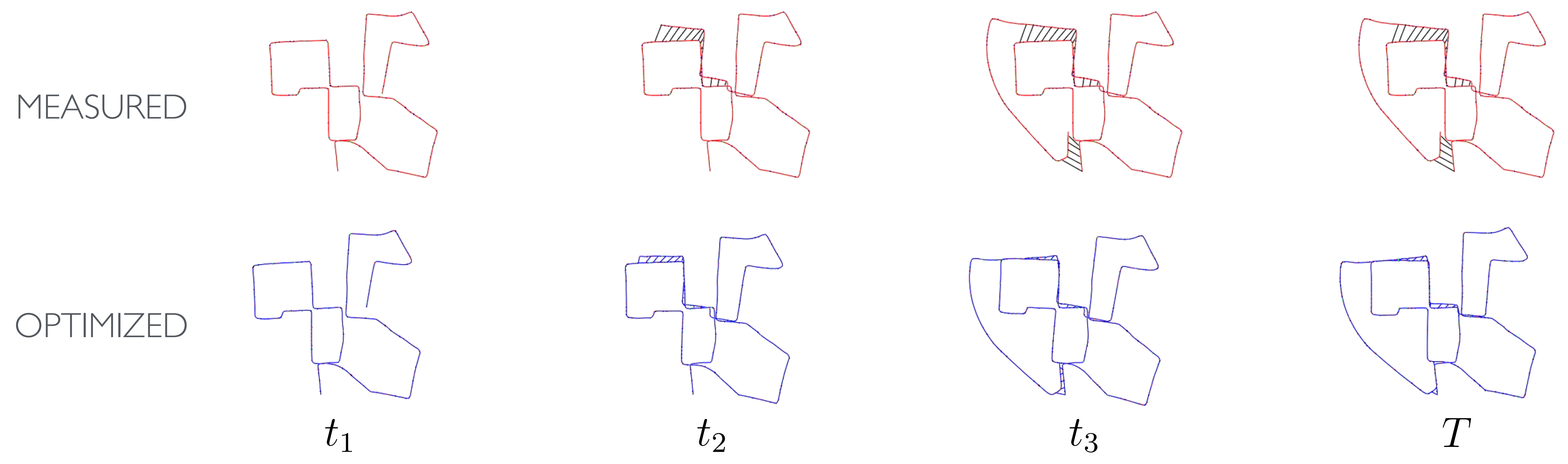
SELF-SUPERVISED LOOP-CLOSURE DETECTION

- **Vision-based Pose-Graph SLAM**
  - Self-supervised loop-closure identification with learned embedding

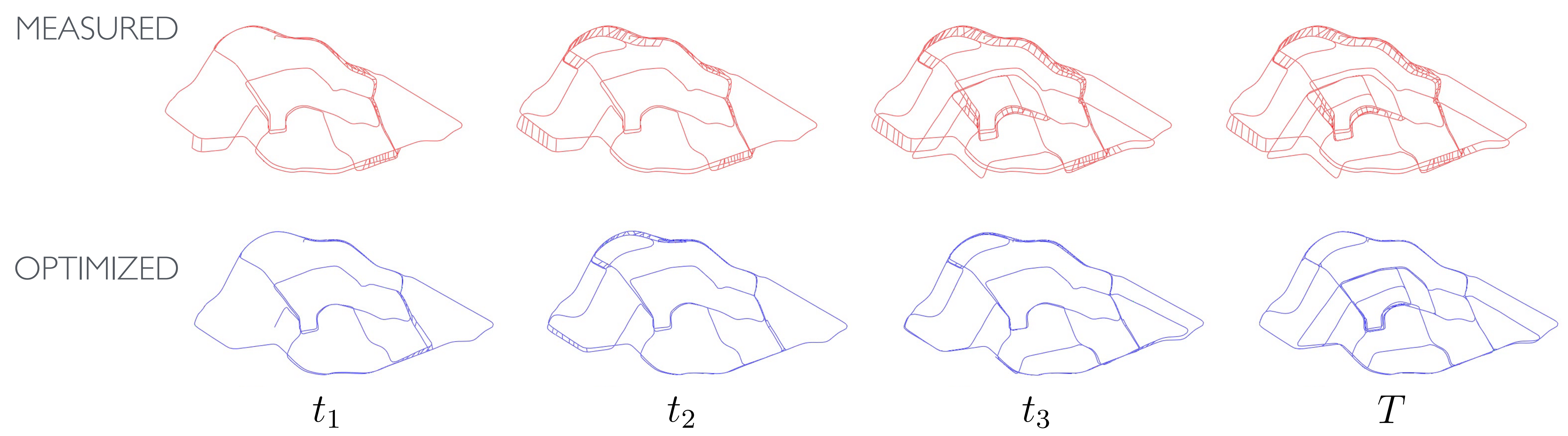


Factor Graph representation of Pose-Graph SLAM

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} p(\mathbf{X} \mid \mathbf{U}, \mathbf{Z_c})$$
$$= \arg\min_{\mathbf{X}} \left\{ \underbrace{\sum_{i=1}^{M} \|f_u(\mathbf{x}_{i-1}, \mathbf{u}_i) - \mathbf{x}_i\|^2_{\Sigma_u}}_{\text{Odometry Measurement Factors}} + \underbrace{\sum_{(j,k)\in\mathcal{C}} \|h_c(\mathbf{x}_j, \mathbf{x}_k) - \mathbf{z}_{jk}\|^2_{\Sigma_c}}_{\text{Loop-Closure Constraint Factors}} \right\}$$



KITTI Dataset

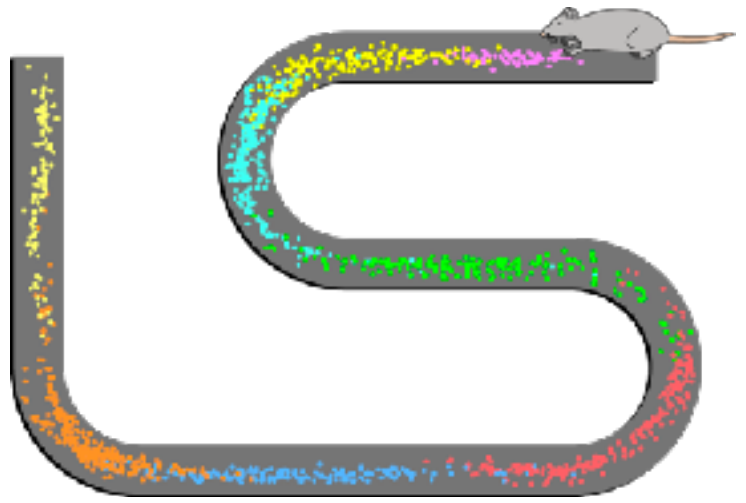


St. Lucia Dataset

SLAM-SUPERVISED SCENE EMBEDDINGS

SLAM-SUPERVISED SCENE EMBEDDINGS

LEARNING TO LOCALIZE



Place-cells

2014 Nobel Prize in Physiology or Medicine

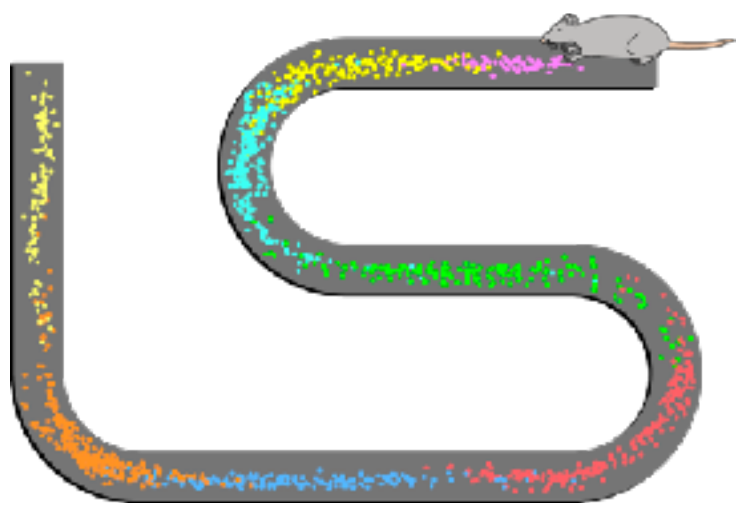
Spatial Cells in the Hippocampal Formation

John O'Keefe, May-Britt Moser, Edvard I. Moser

SLAM-SUPERVISED SCENE EMBEDDINGS

- Learning location-specific scene embeddings

LEARNING TO LOCALIZE



Place-cells

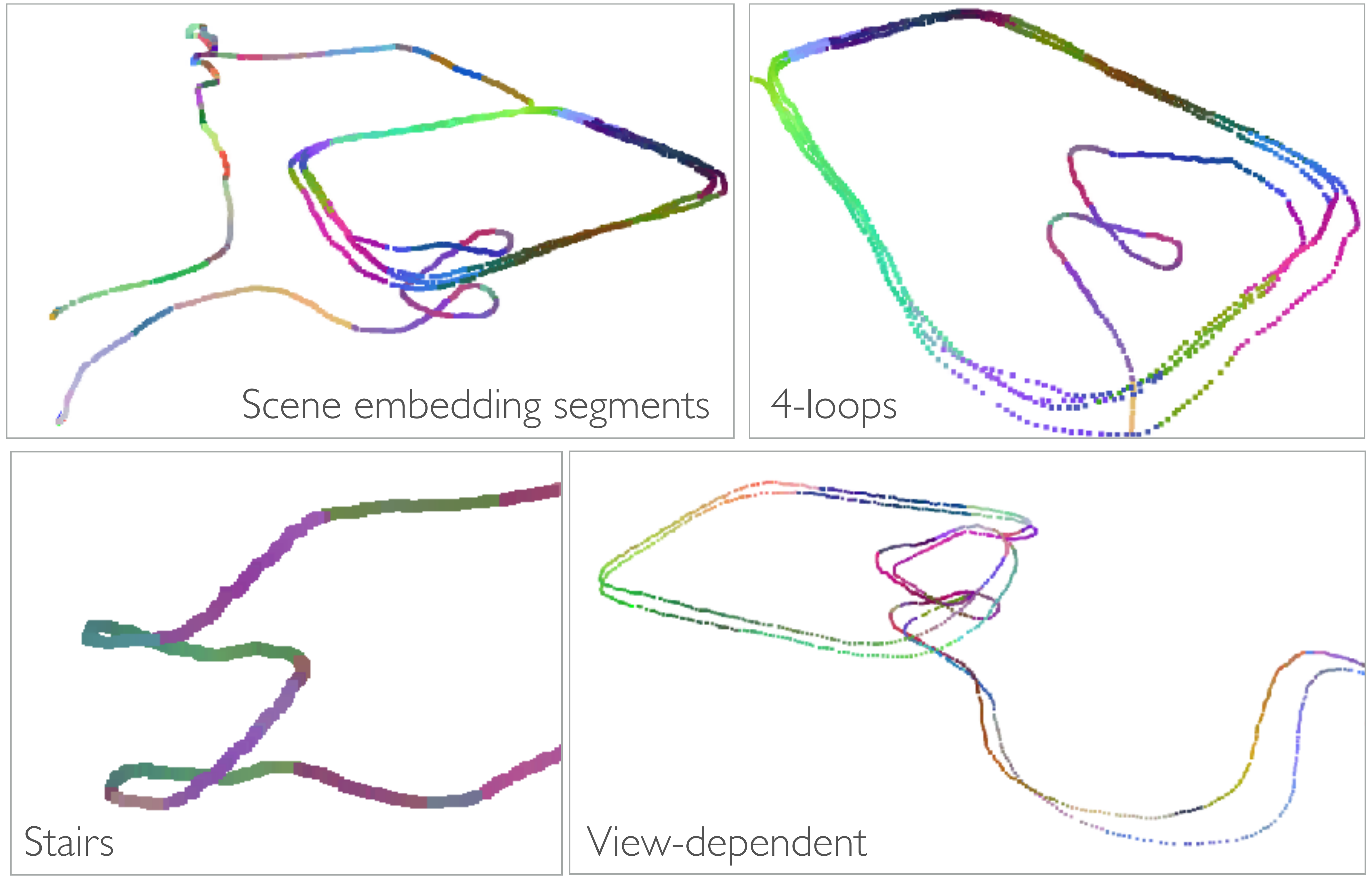
2014 Nobel Prize in Physiology or Medicine

Spatial Cells in the Hippocampal Formation

John O'Keefe, May-Britt Moser, Edvard I. Moser

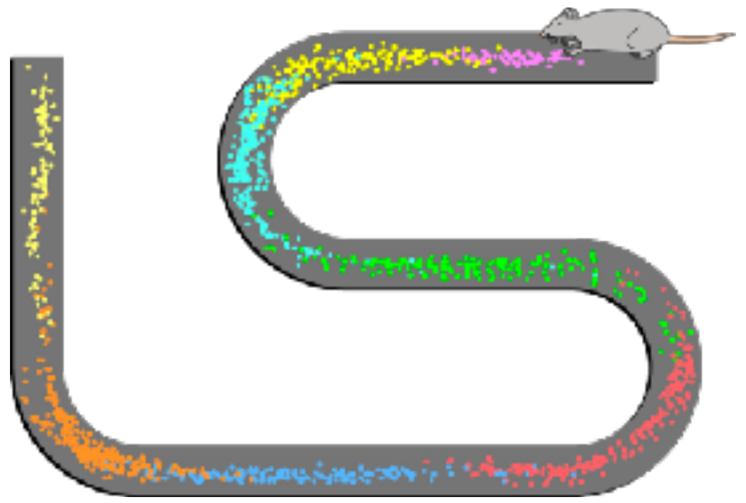
SLAM-SUPERVISED SCENE EMBEDDINGS

- **Learning location-specific scene embeddings**
  - Learned embedding powerful in discriminating visual scene instances



SLAM-Supervised Scene Embeddings
Consistent scene embeddings for same location
(Colors obtained via T-SNE embedding of learned metric)

LEARNING TO LOCALIZE

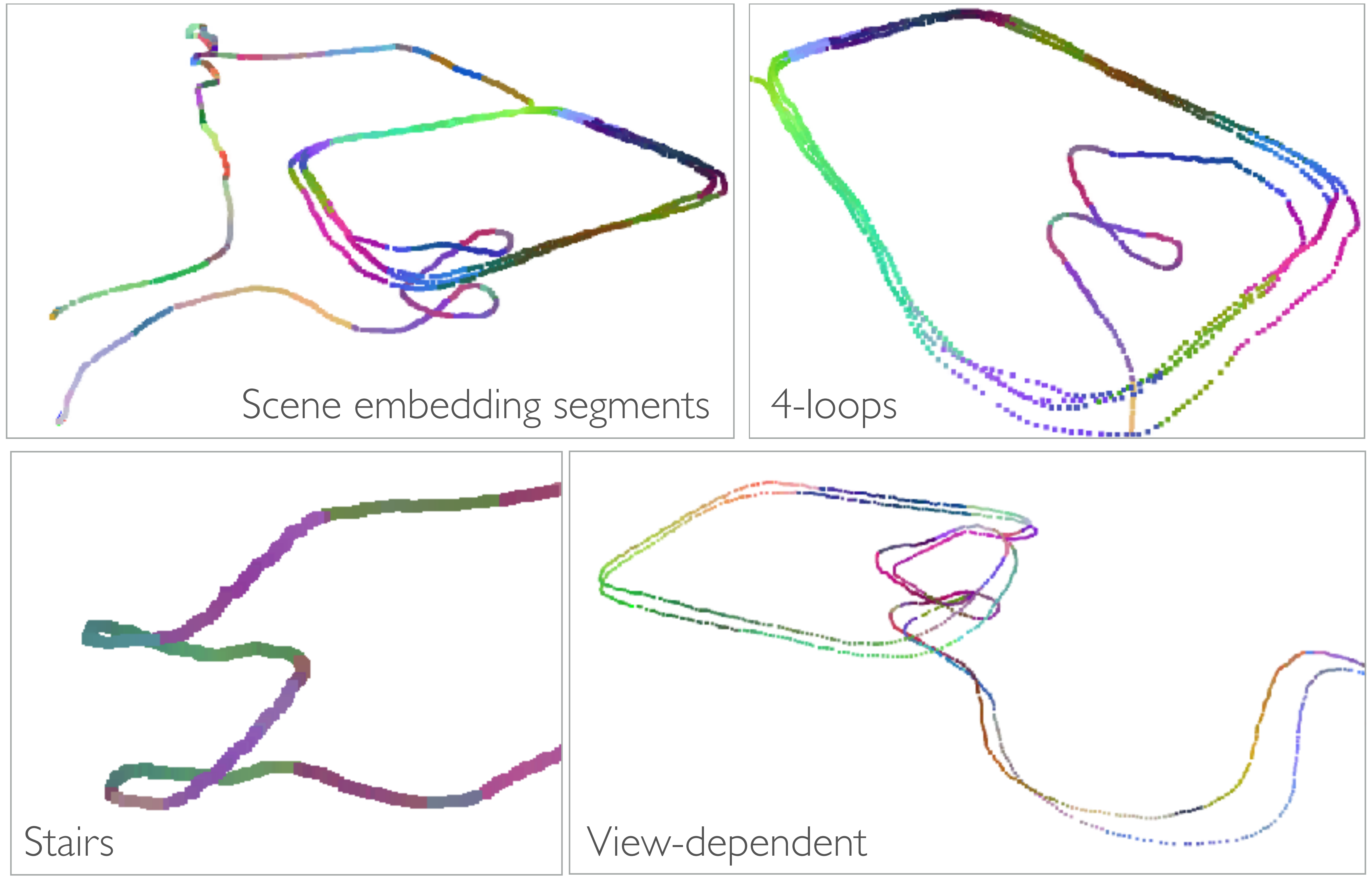


Place-cells

2014 Nobel Prize in Physiology or Medicine
Spatial Cells in the Hippocampal Formation
John O'Keefe, May-Britt Moser, Edvard I. Moser

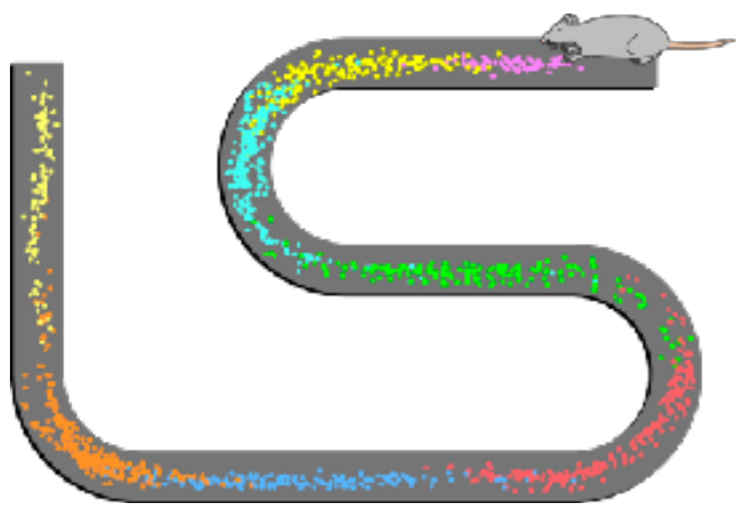
SLAM-SUPERVISED SCENE EMBEDDINGS

- **Learning location-specific scene embeddings**
  - Learned embedding powerful in discriminating visual scene instances
  - Weak-supervision under uncertainty (SLAM)



SLAM-Supervised Scene Embeddings
Consistent scene embeddings for same location
(Colors obtained via T-SNE embedding of learned metric)

LEARNING TO LOCALIZE

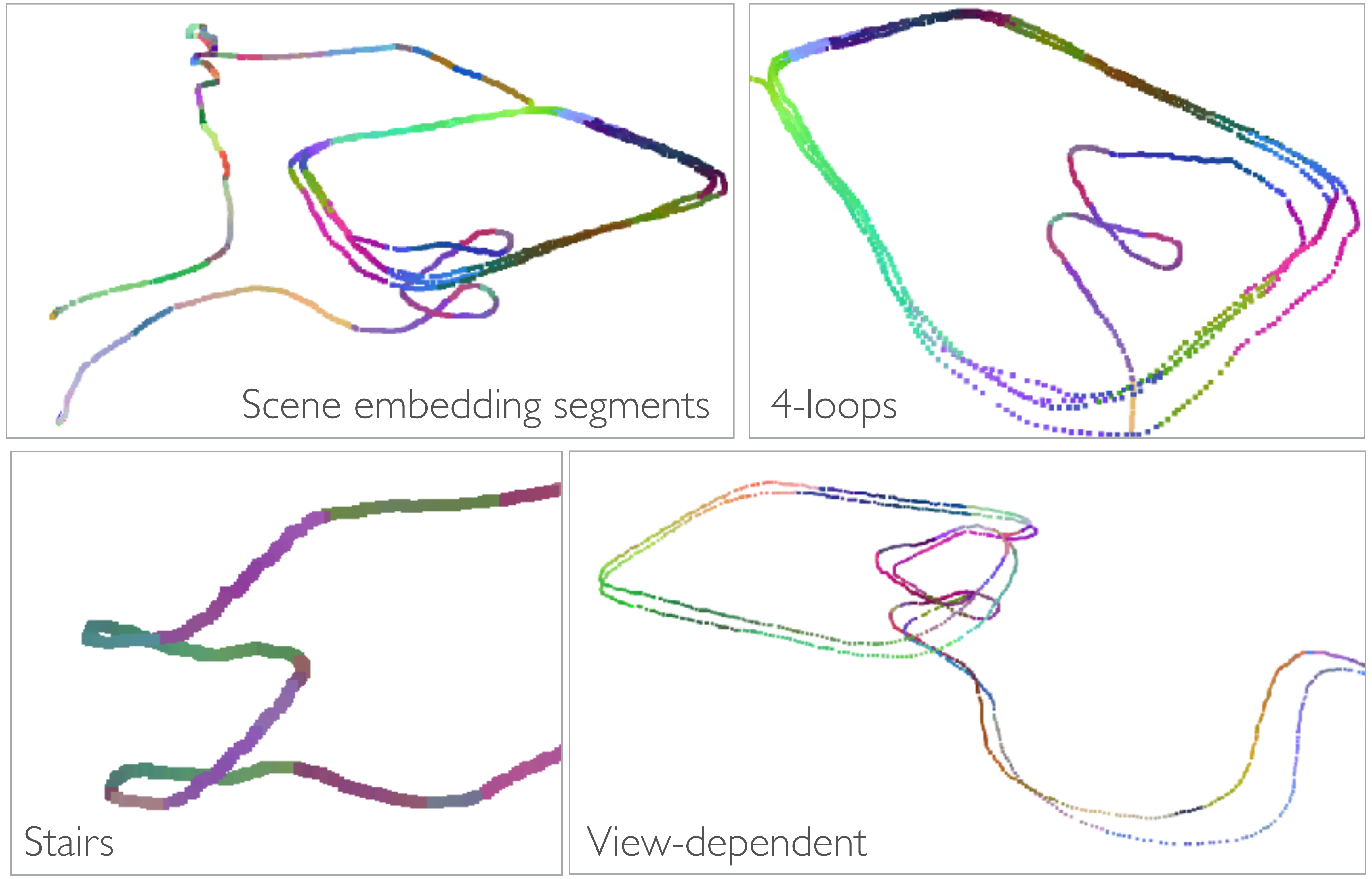


Place-cells

2014 Nobel Prize in Physiology or Medicine
Spatial Cells in the Hippocampal Formation
John O'Keefe, May-Britt Moser, Edvard I. Moser

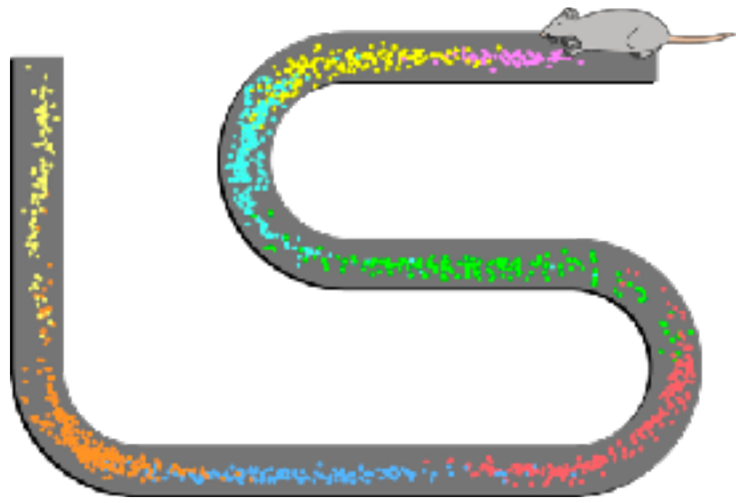
SLAM-SUPERVISED SCENE EMBEDDINGS

- **Learning location-specific scene embeddings**
  - Learned embedding powerful in discriminating visual scene instances
  - Weak-supervision under uncertainty (SLAM)
  - On-the-fly fine-tuning



SLAM-Supervised Scene Embeddings
Consistent scene embeddings for same location
(Colors obtained via T-SNE embedding of learned metric)

LEARNING TO LOCALIZE

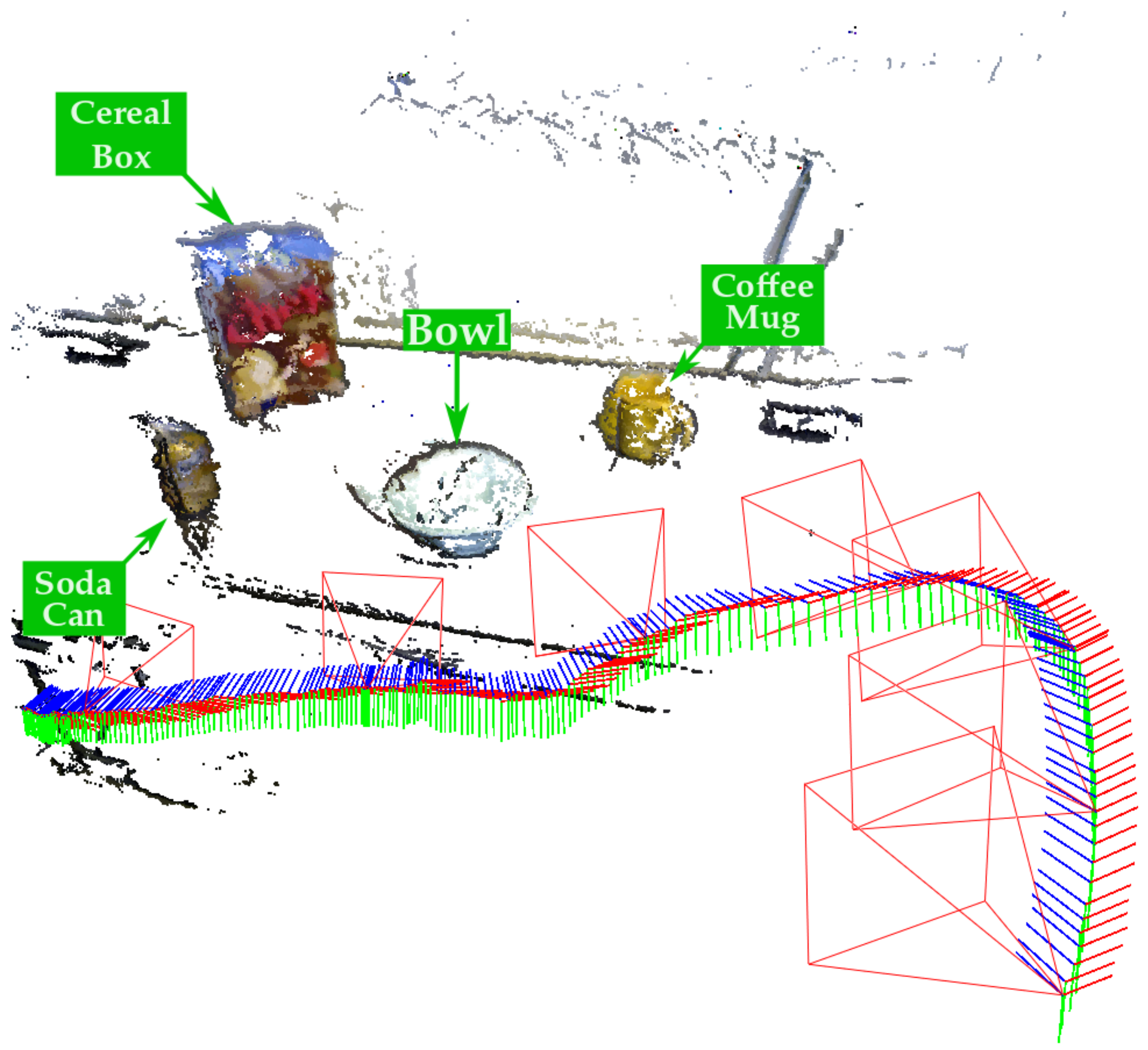


Place-cells

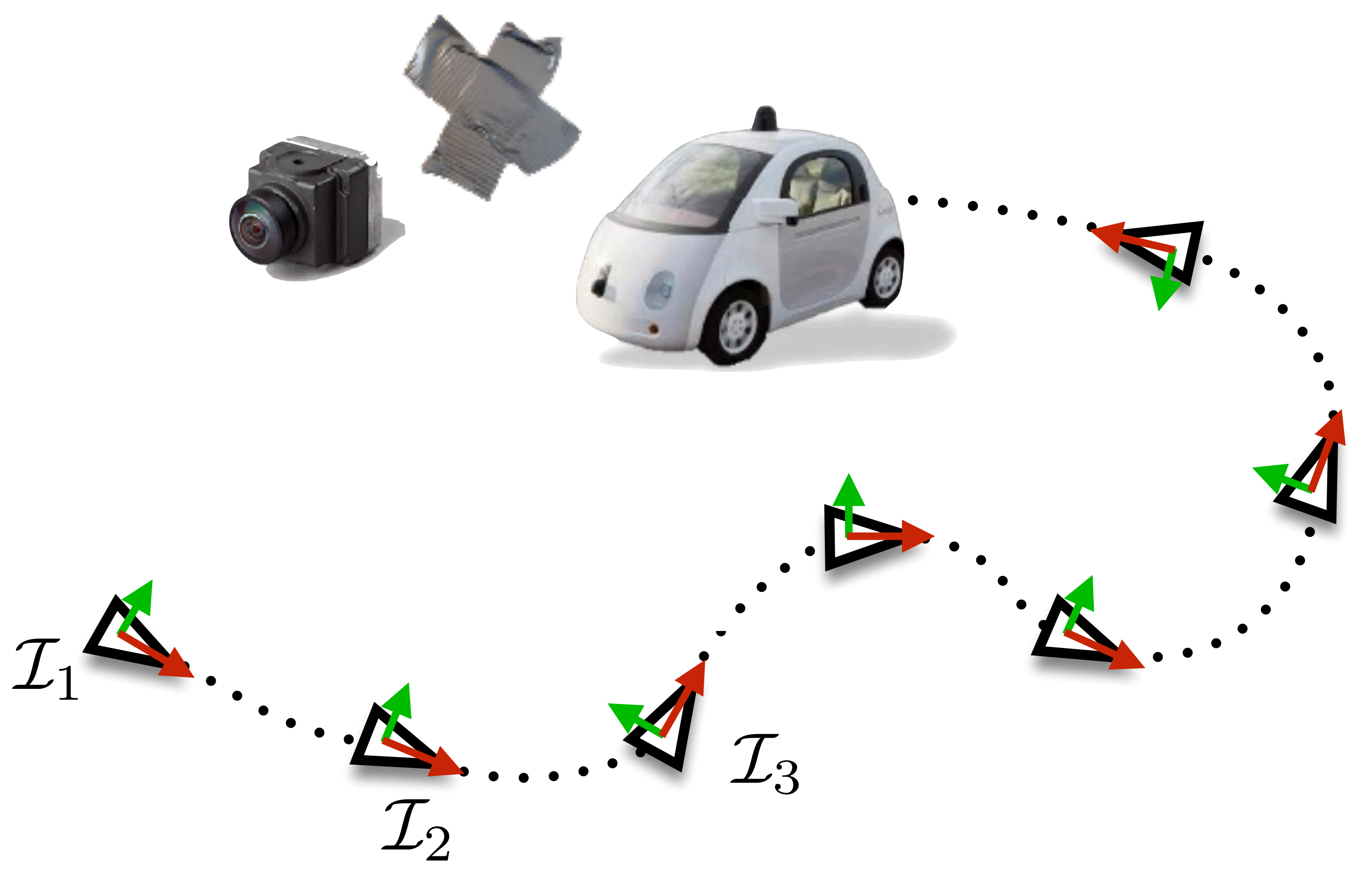
2014 Nobel Prize in Physiology or Medicine
Spatial Cells in the Hippocampal Formation
John O'Keefe, May-Britt Moser, Edvard I. Moser

CENTRAL THEME: SLAM AS A SUPERVISORY SIGNAL

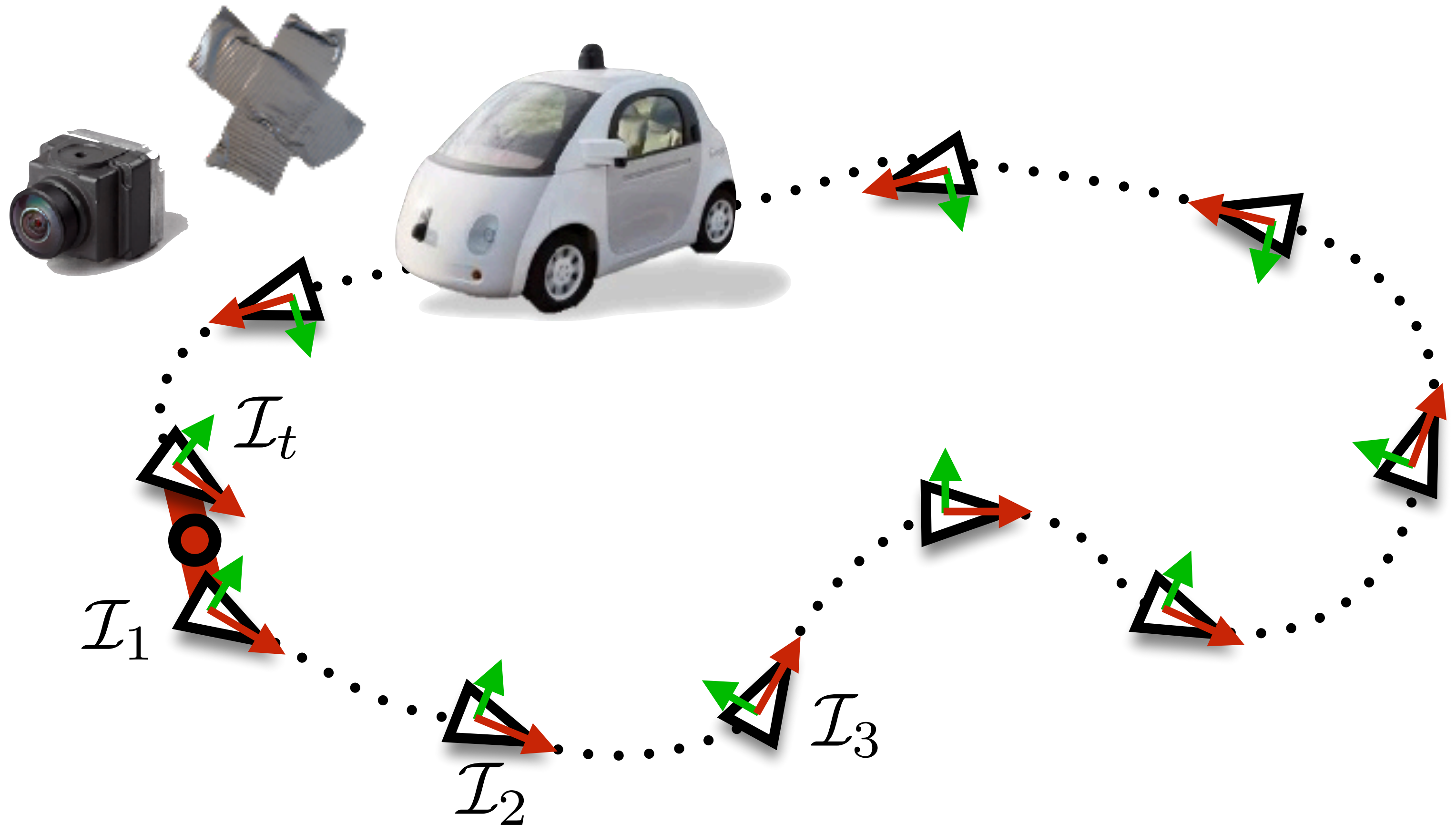
Monocular SLAM-Supported Object Recognition



Self-Supervised Visual Ego-motion Learning



Self-Supervised Visual Place Recognition Learning



SUPERVISION & SELF-SUPERVISION IN MOBILE ROBOTS with SLAM

Correspondence Engine
(Geometric data association)

Self-Supervision
(SLAM-aided supervision)

Knowledge Transfer
(Bootstrapping)

CONTRIBUTIONS

CONTRIBUTIONS

- Spatially Cognizant Perception

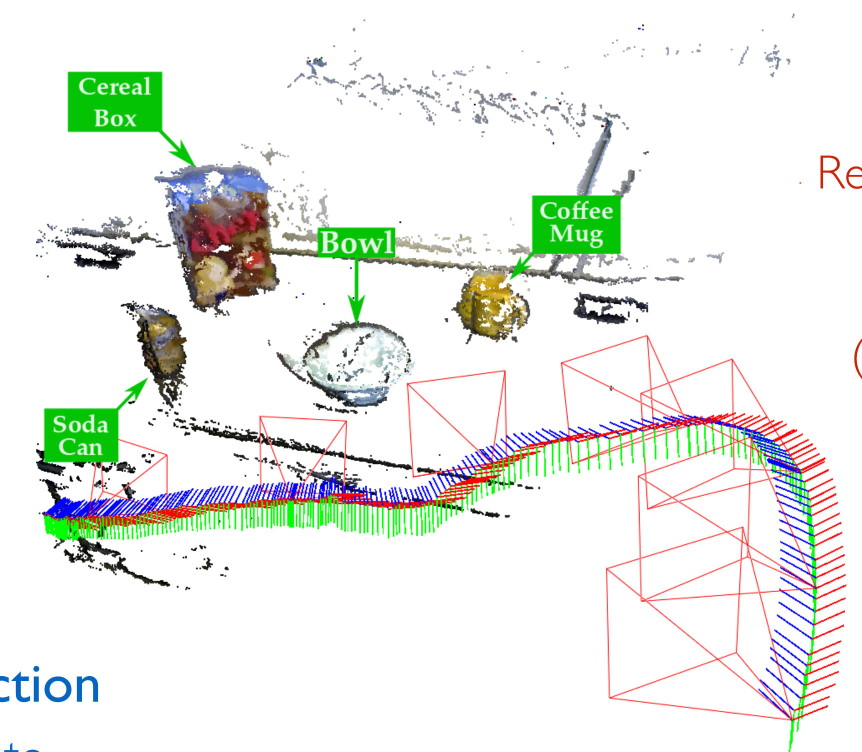
CONTRIBUTIONS

- Spatially Cognizant Perception
  - **SLAM-Supported Object Recognition**: Leverage SLAM capabilities to bolster classical object recognition in spatially-situated scenes

Single RGB Camera

Monocular SLAM supports improved recognition

(Semi-Dense Mapping Backend)



Robust

Reduced false positives via view correspondence from SLAM

(Multi-view prediction)

Multi-view Object Detection

Objects easily tease apart to enable better proposals

(Proposals from Semi-Dense Maps)

Scalable

Box-encoding / RoI Pooling

(FLAIR/Fast R-CNN)

Monocular SLAM-Supported Object Recognition

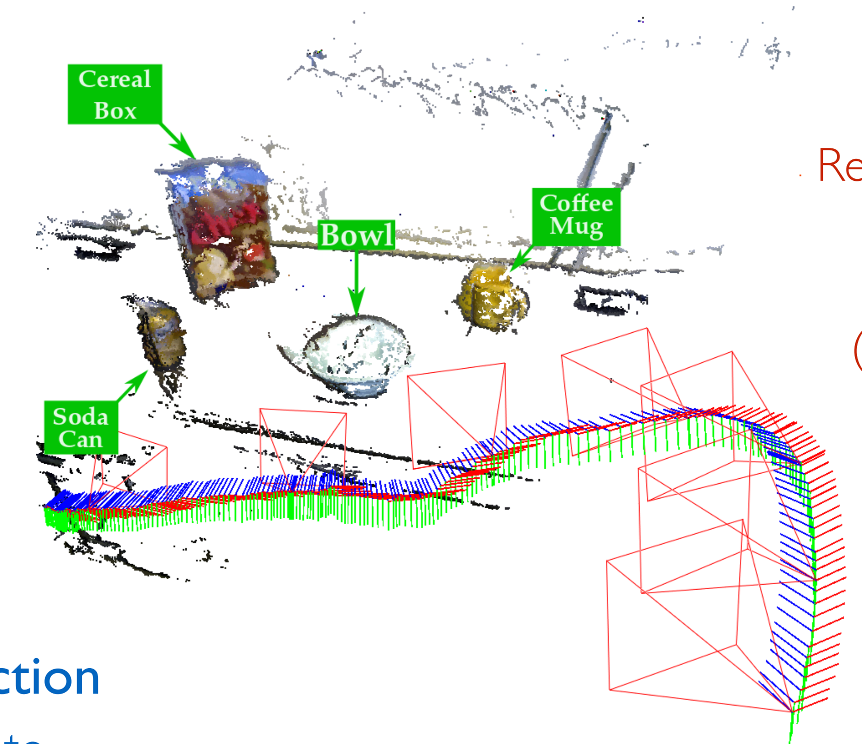
Pillai et al. (RSS 2015)

CONTRIBUTIONS

- Spatially Cognizant Perception
  - **SLAM-Supported Object Recognition**: Leverage SLAM capabilities to bolster classical object recognition in spatially-situated scenes
  - **SLAM-aware Few-shot Object Learning:** Use SLAM as a correspondence-engine for spatially-consistent and occlusion-aware label propagation, and learn object detectors from considerably fewer training examples

Single RGB Camera
Monocular SLAM supports improved recognition
(Semi-Dense Mapping Backend)

Robust
Reduced false positives via view correspondence from SLAM
(Multi-view prediction)



Multi-view Object Detection
Objects easily tease apart to enable better proposals
(Proposals from Semi-Dense Maps)

Scalable
Box-encoding / RoI Pooling
(FLAIR/Fast R-CNN)

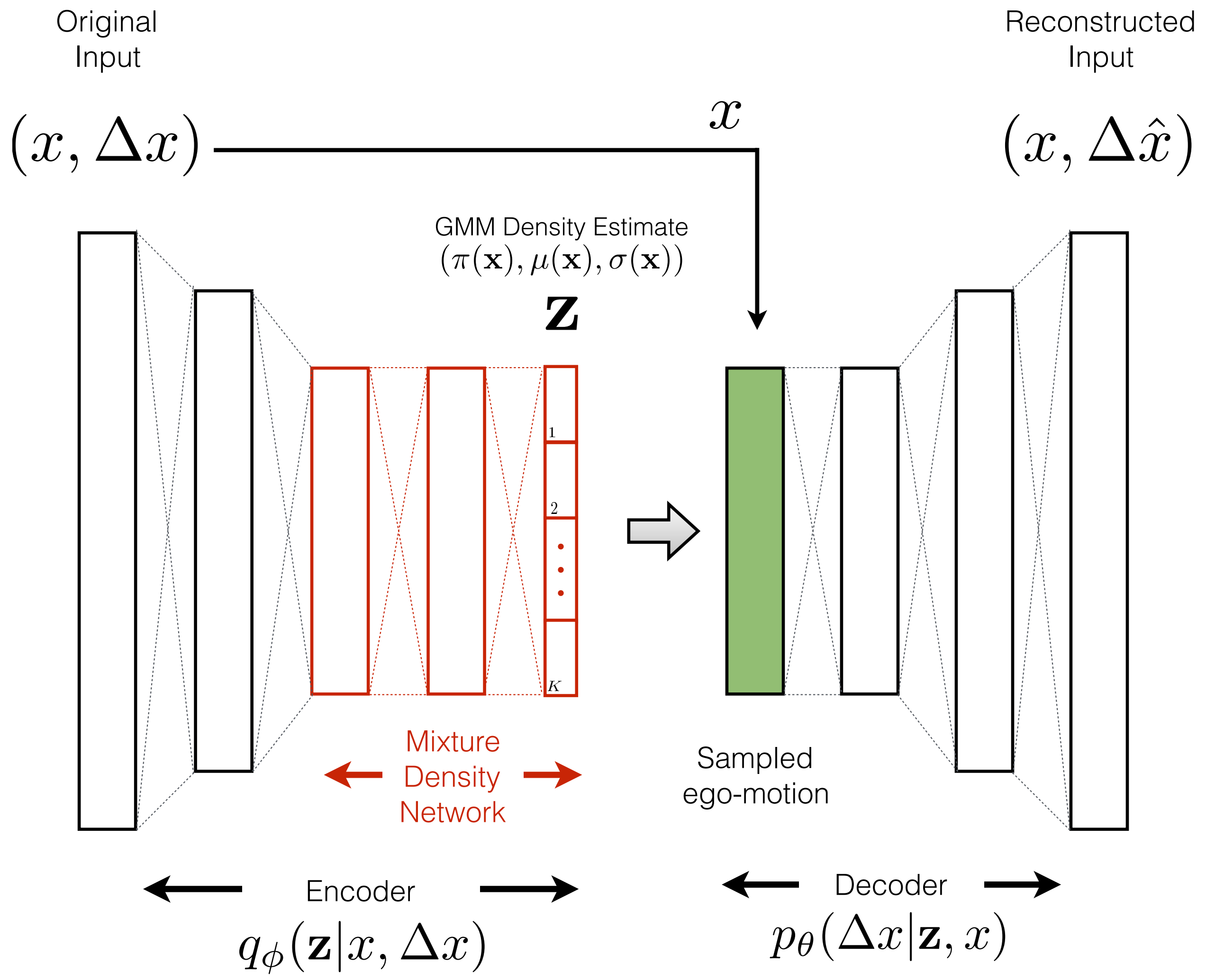
Monocular SLAM-Supported Object Recognition
Pillai et al. (RSS 2015)

CONTRIBUTIONS

- Life-Long Learning in Mobile Robots
  - **Self-supervised Ego-motion and Visual Place Recognition Learning:** By bootstrapping the robot's ability to perform GPS-aided SLAM, we develop a self-supervised visual SLAM front-end capable of performing visual ego-motion, and vision-based loop-closure recognition

CONTRIBUTIONS

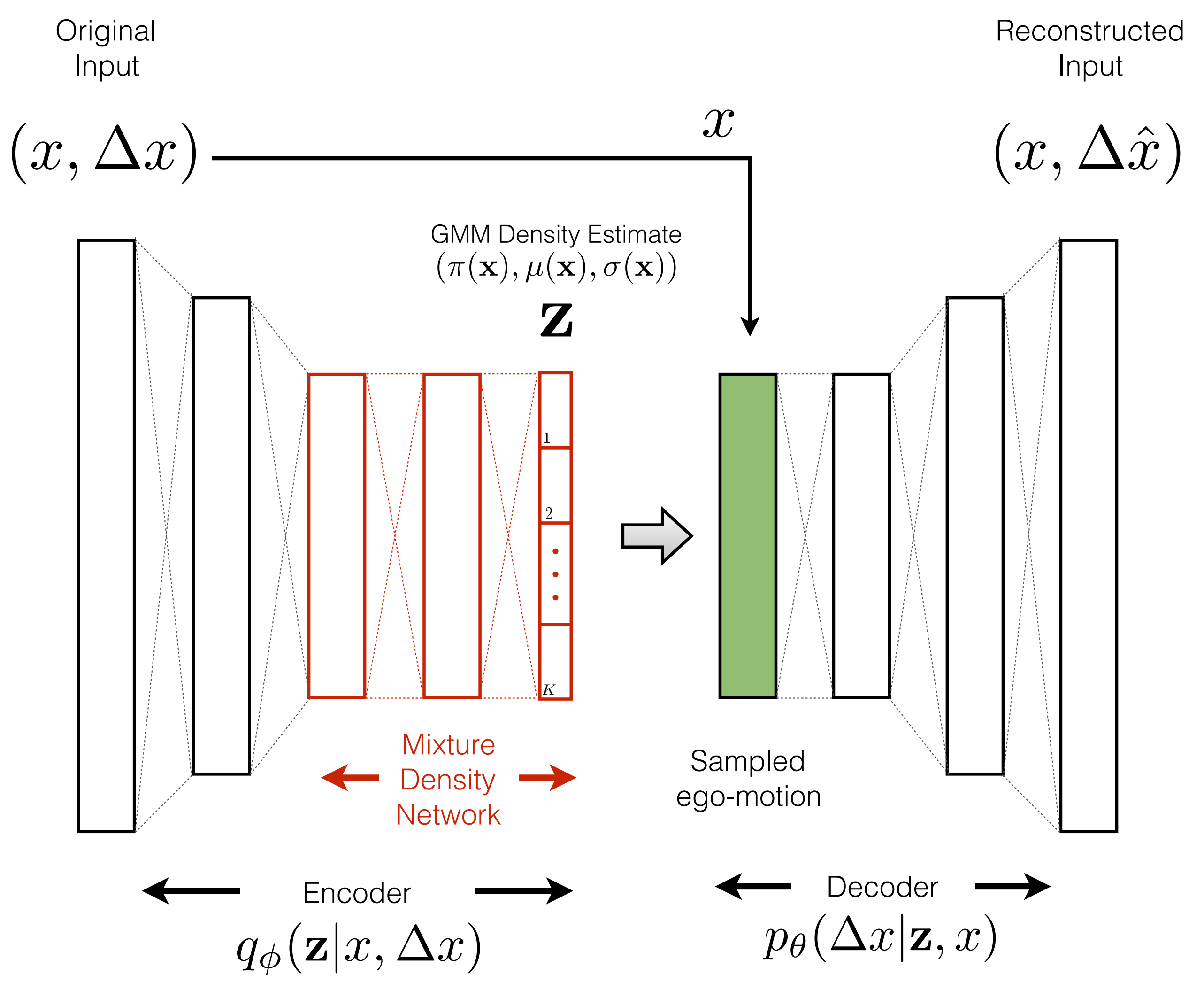
- Life-Long Learning in Mobile Robots
  - **Self-supervised Ego-motion and Visual Place Recognition Learning:** By bootstrapping the robot's ability to perform GPS-aided SLAM, we develop a self-supervised visual SLAM front-end capable of performing visual ego-motion, and vision-based loop-closure recognition



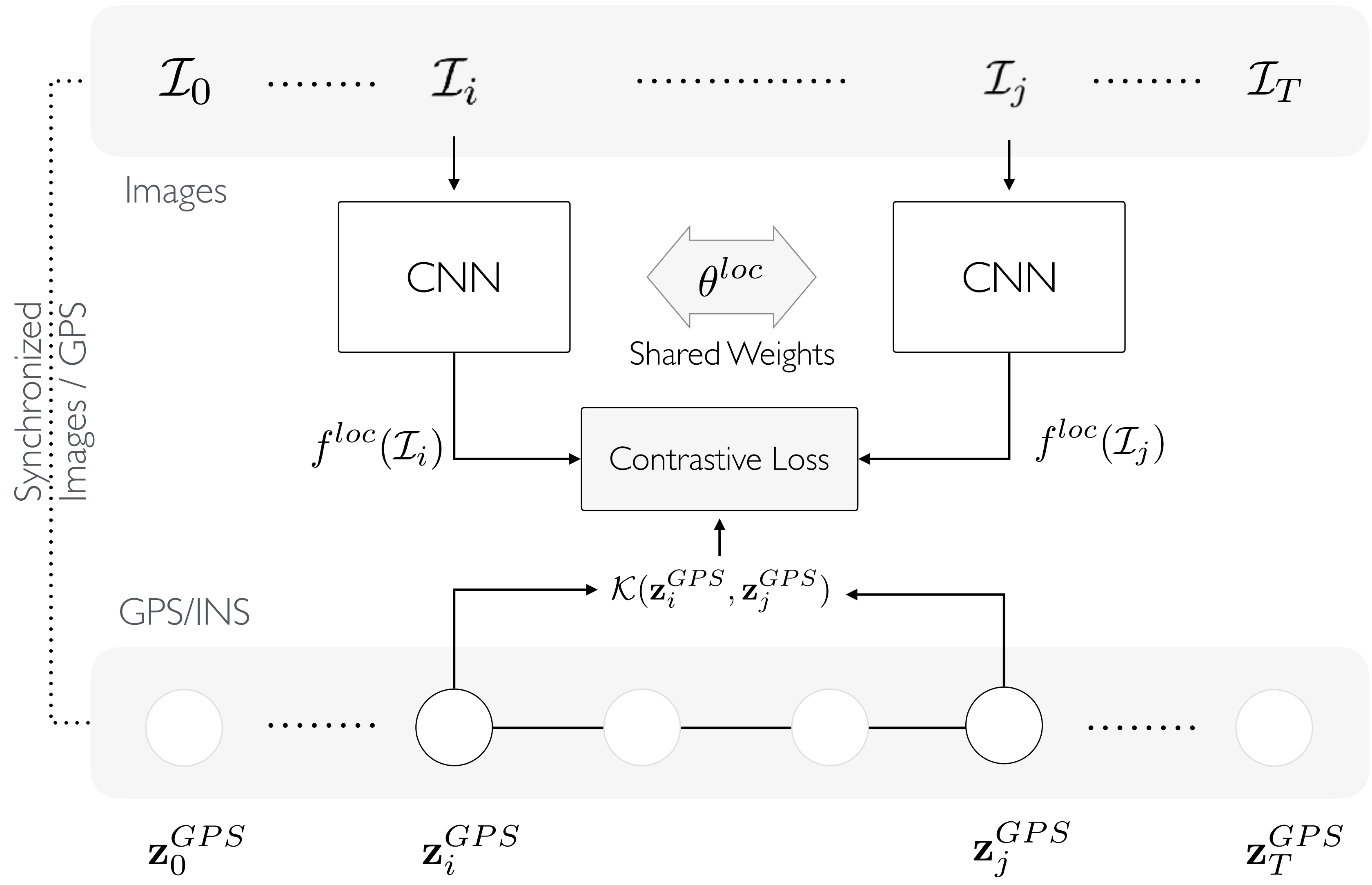
Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)

CONTRIBUTIONS

- Life-Long Learning in Mobile Robots
  - **Self-supervised Ego-motion and Visual Place Recognition Learning:** By bootstrapping the robot's ability to perform GPS-aided SLAM, we develop a self-supervised visual SLAM front-end capable of performing visual ego-motion, and vision-based loop-closure recognition



Towards Visual Ego-motion Learning in Robots
Pillai et al. (IROS 2017)



Self-Supervised Visual Place Recognition in Mobile Robots
Pillai et al. (Learning for Localization and Mapping Workshop, IROS 2017)

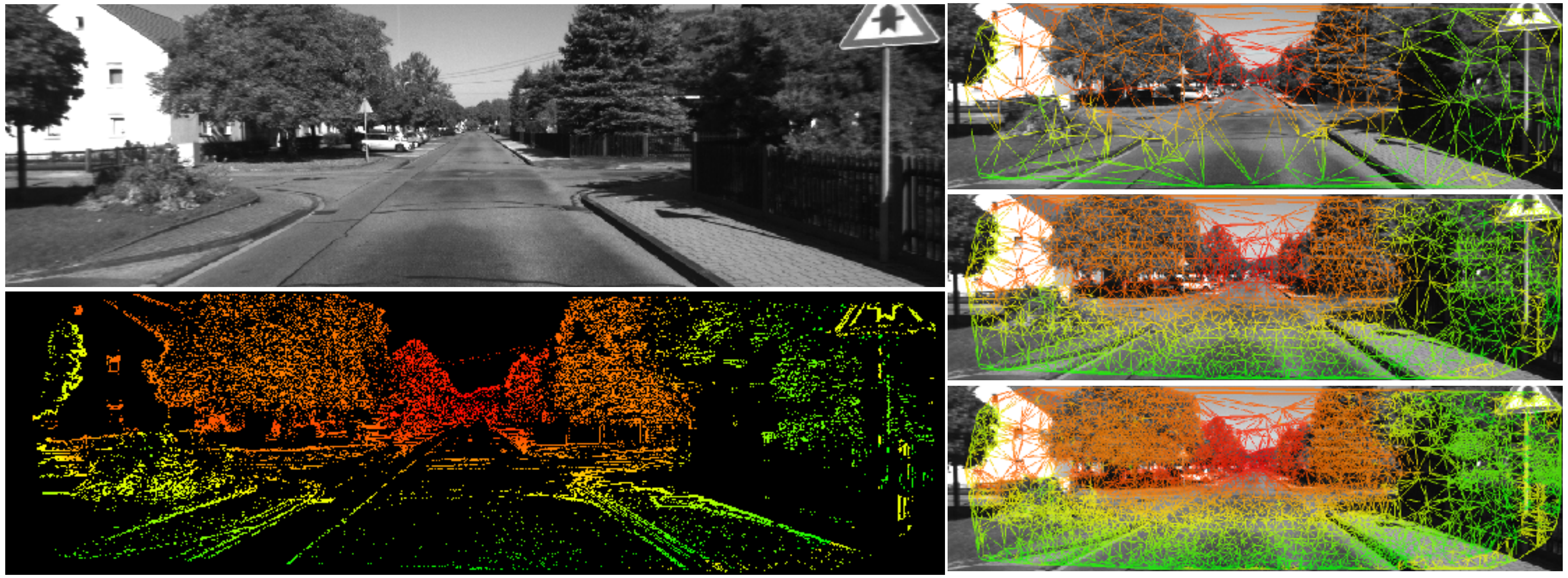
CONTRIBUTIONS

CONTRIBUTIONS

- Map Representations for Vision-Based Navigation

CONTRIBUTIONS

- Map Representations for Vision-Based Navigation
  - **High-Performance and Tunable Stereo Reconstruction:** Develop an any-time, iteratively refine-able, mesh reconstruction algorithm for stereo imagery that can be potentially used in planning, obstacle avoidance etc.



High-Performance And Tunable Stereo Reconstruction
Pillai et al. (ICRA 2016)

FUTURE DIRECTIONS

FUTURE DIRECTIONS

- SLAM as a supervisory signal

FUTURE DIRECTIONS

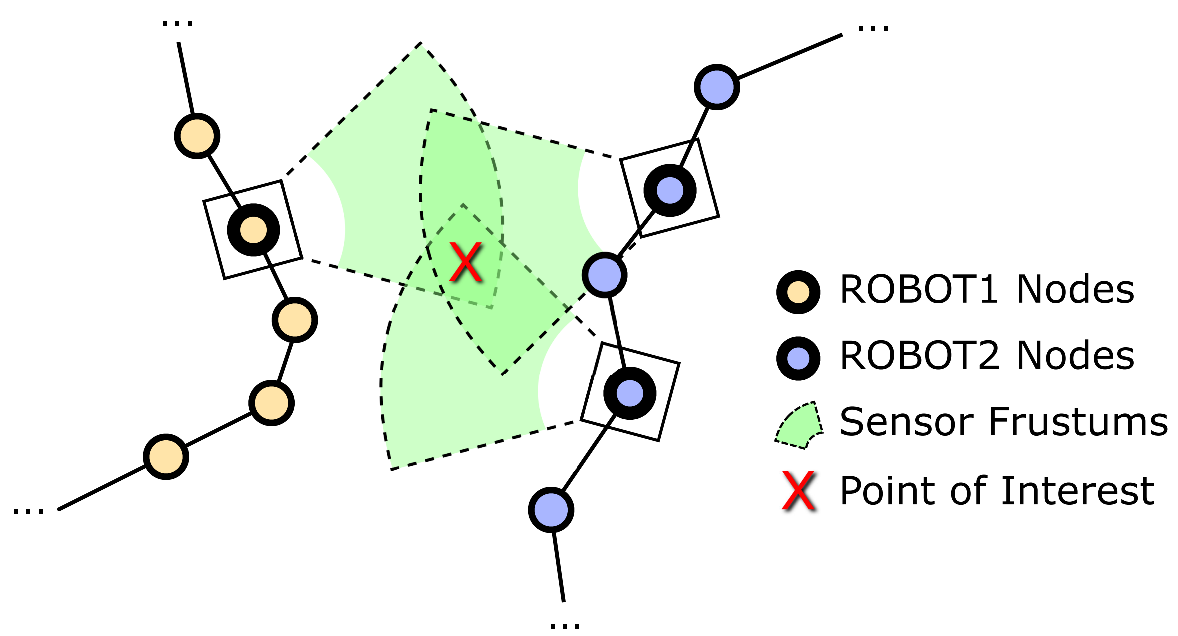
- **SLAM as a supervisory signal**
  - Spatially and Semantically-aware Robot DBs

Spatially and Semantically-Aware Robot DBs
(Where have I seen artwork before?)

FUTURE DIRECTIONS

- **SLAM as a supervisory signal**
  - Spatially and Semantically-aware Robot DBs
  - Expressive Language for Robot Data Querying

Spatially and Semantically-Aware Robot DBs
(Where have I seen artwork before?)



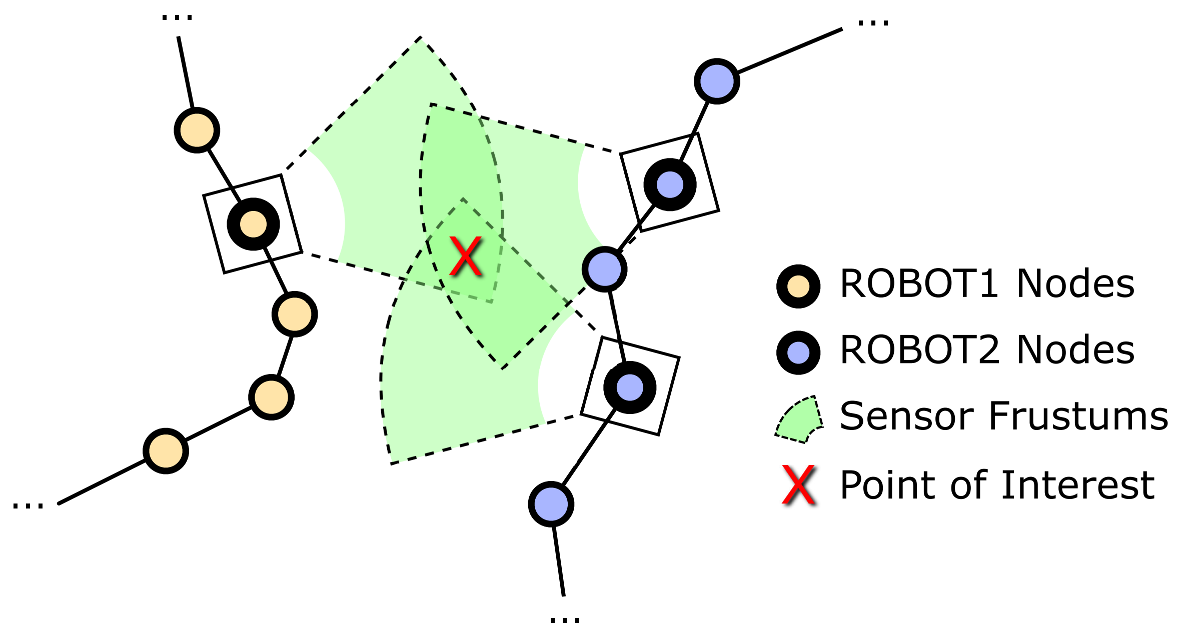
Expressive Language for Robot Data Querying
(Show me X in all robot views, across multiple sessions)

SLAMinDB: Centralized Graph Databases for Mobile Robotics [Fourie et. al 2017]

FUTURE DIRECTIONS

- **SLAM as a supervisory signal**
  - Spatially and Semantically-aware Robot DBs
  - Expressive Language for Robot Data Querying
  - Self-Supervised Cross-Modal Learning in Robots

Spatially and Semantically-Aware Robot DBs
(Where have I seen artwork before?)



Expressive Language for Robot Data Querying
(Show me X in all robot views, across multiple sessions)

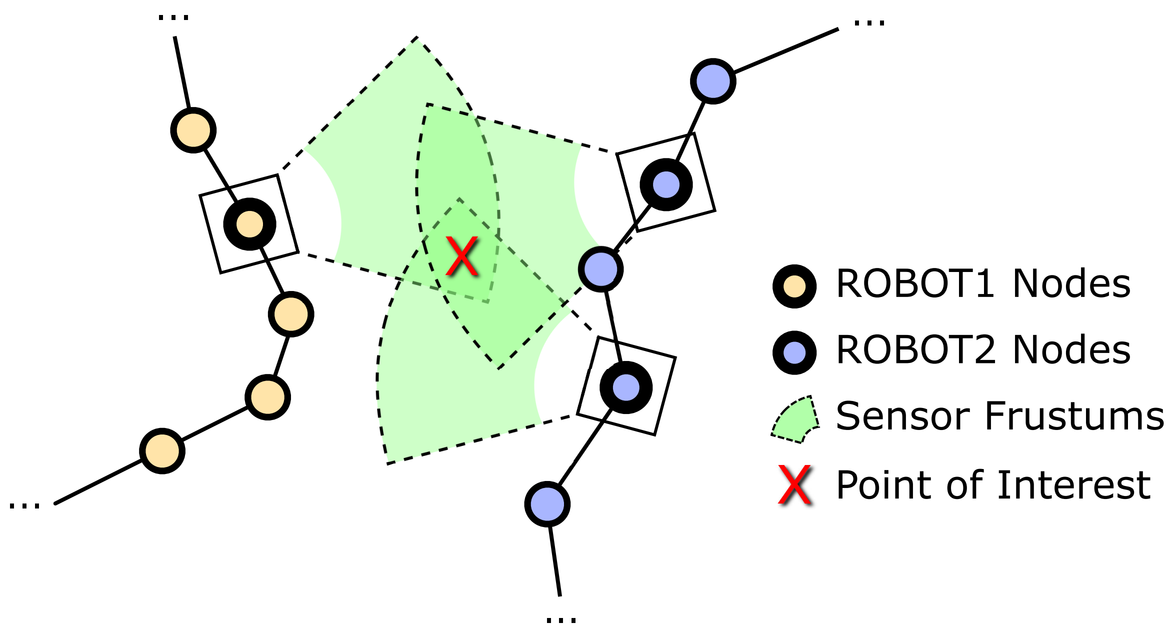
SLAMinDB: Centralized Graph Databases for Mobile Robotics **[Fourie et. al 2017]**

FUTURE DIRECTIONS

- **SLAM as a supervisory signal**
  - Spatially and Semantically-aware Robot DBs
  - Expressive Language for Robot Data Querying
  - Self-Supervised Cross-Modal Learning in Robots

Transferring LiDAR information for camera-based scene reconstruction

Spatially and Semantically-Aware Robot DBs
(Where have I seen artwork before?)

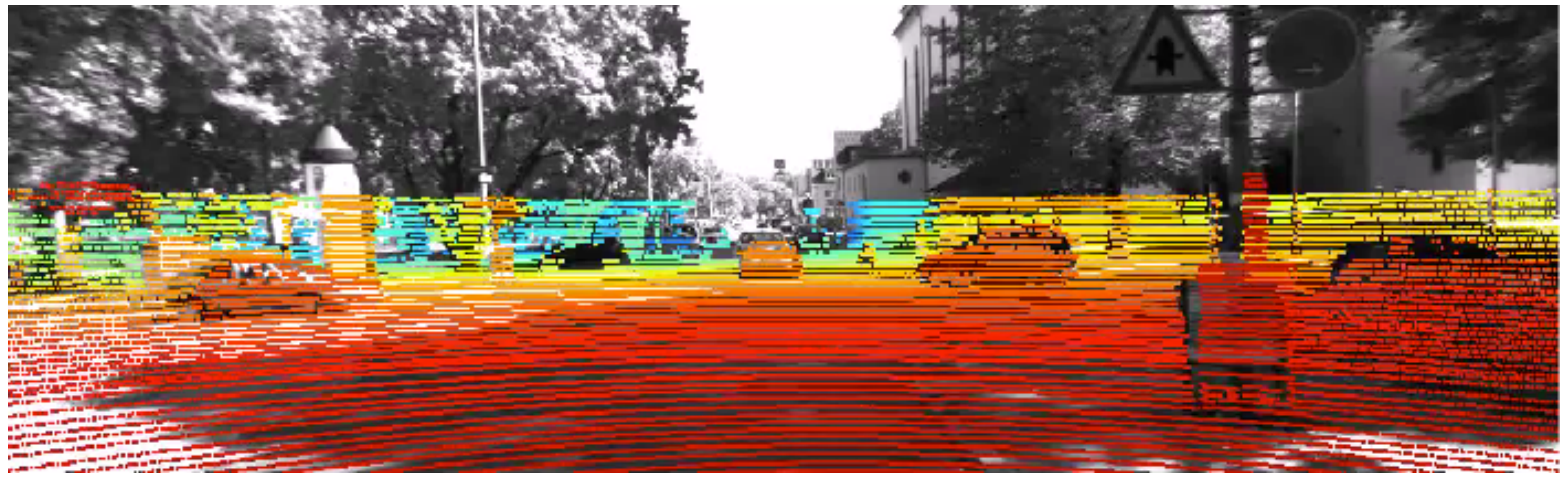


Expressive Language for Robot Data Querying
(Show me X in all robot views, across multiple sessions)

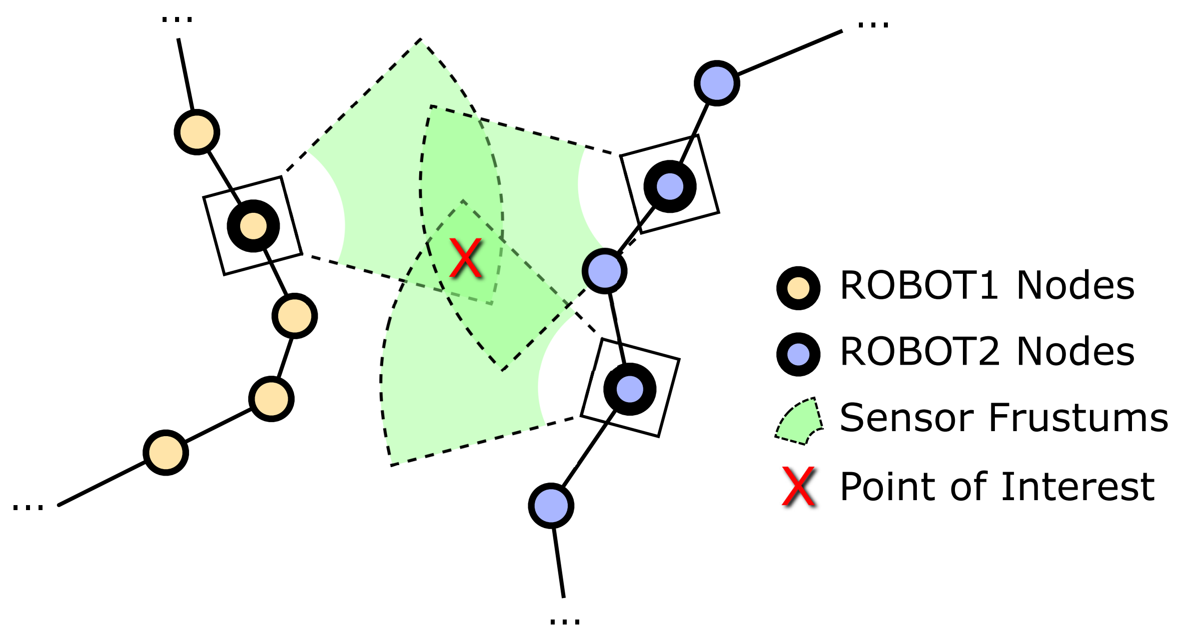
SLAMinDB: Centralized Graph Databases for Mobile Robotics [Fourie et. al 2017]

FUTURE DIRECTIONS

- **SLAM as a supervisory signal**
  - Spatially and Semantically-aware Robot DBs
  - Expressive Language for Robot Data Querying
  - Self-Supervised Cross-Modal Learning in Robots



Spatially and Semantically-Aware Robot DBs
(Where have I seen artwork before?)

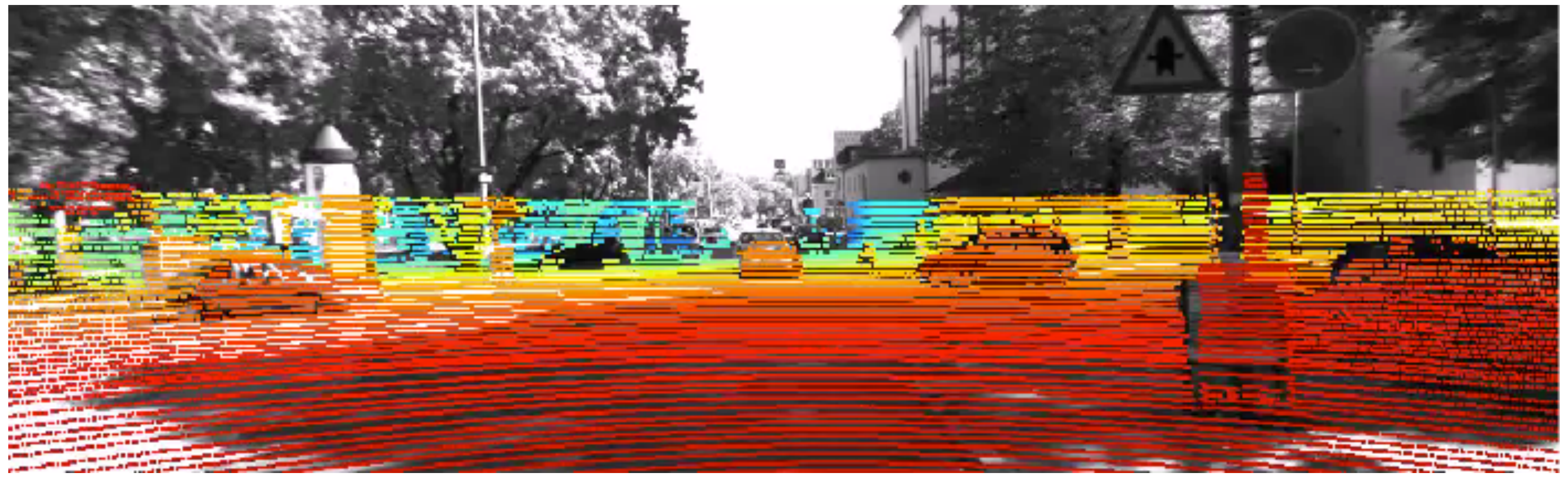


Expressive Language for Robot Data Querying
(Show me X in all robot views, across multiple sessions)

SLAMinDB: Centralized Graph Databases for Mobile Robotics **[Fourie et. al 2017]**

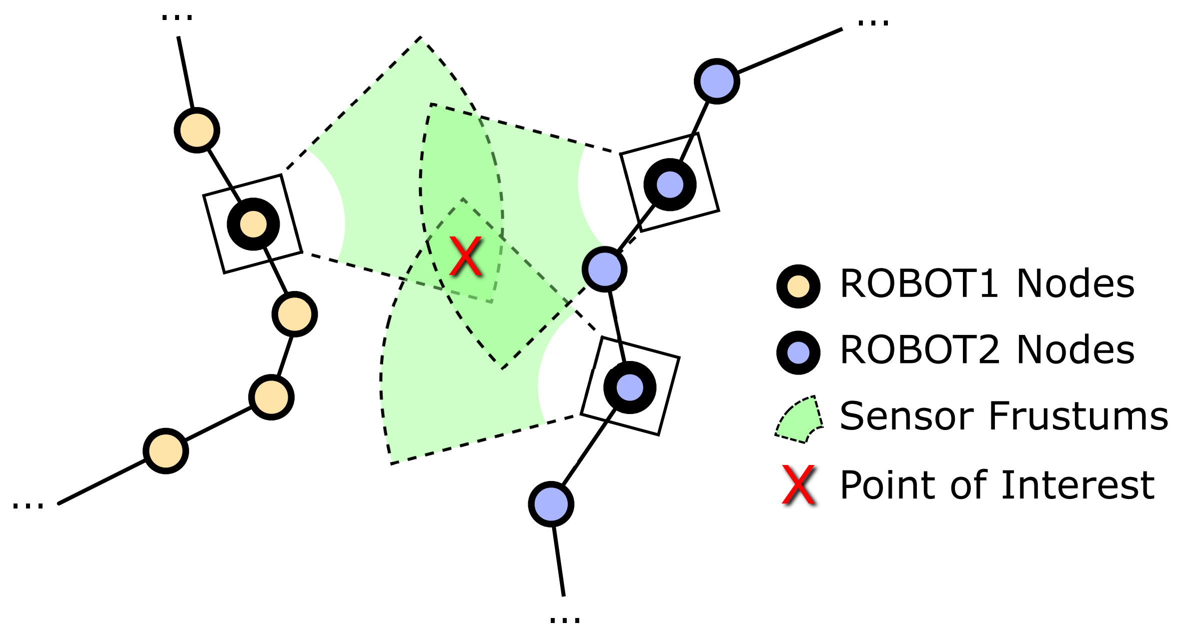
FUTURE DIRECTIONS

- **SLAM as a supervisory signal**
  - Spatially and Semantically-aware Robot DBs
  - Expressive Language for Robot Data Querying
  - Self-Supervised Cross-Modal Learning in Robots



Transferring hindsight
experience for lane-estimation

Spatially and Semantically-Aware Robot DBs
(Where have I seen artwork before?)

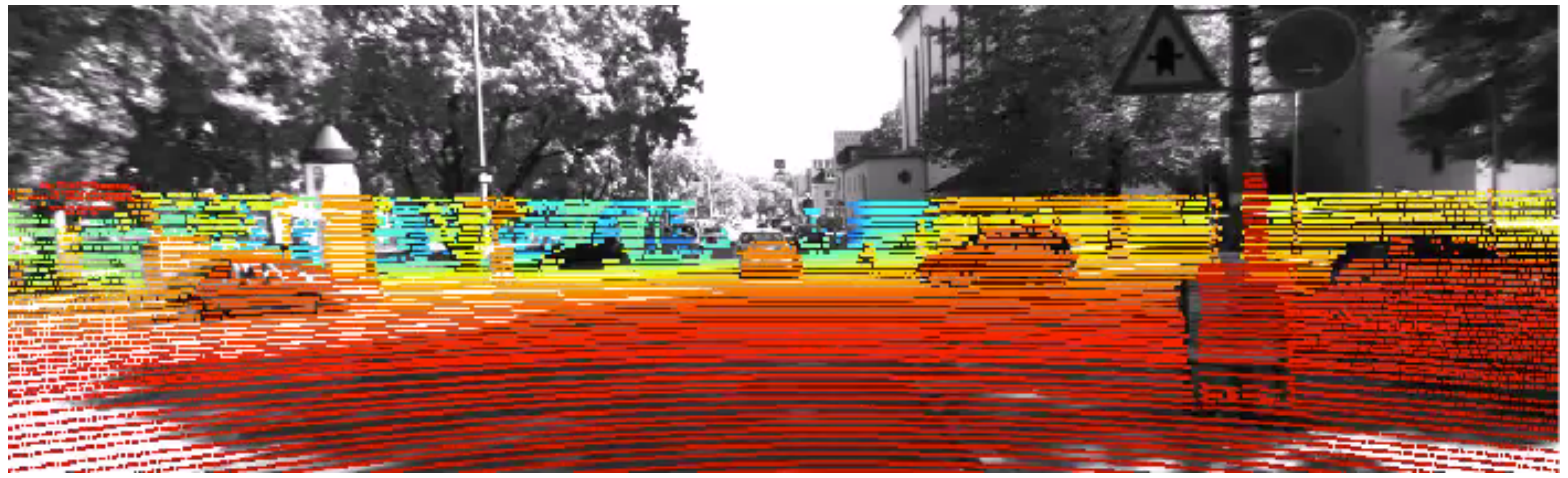


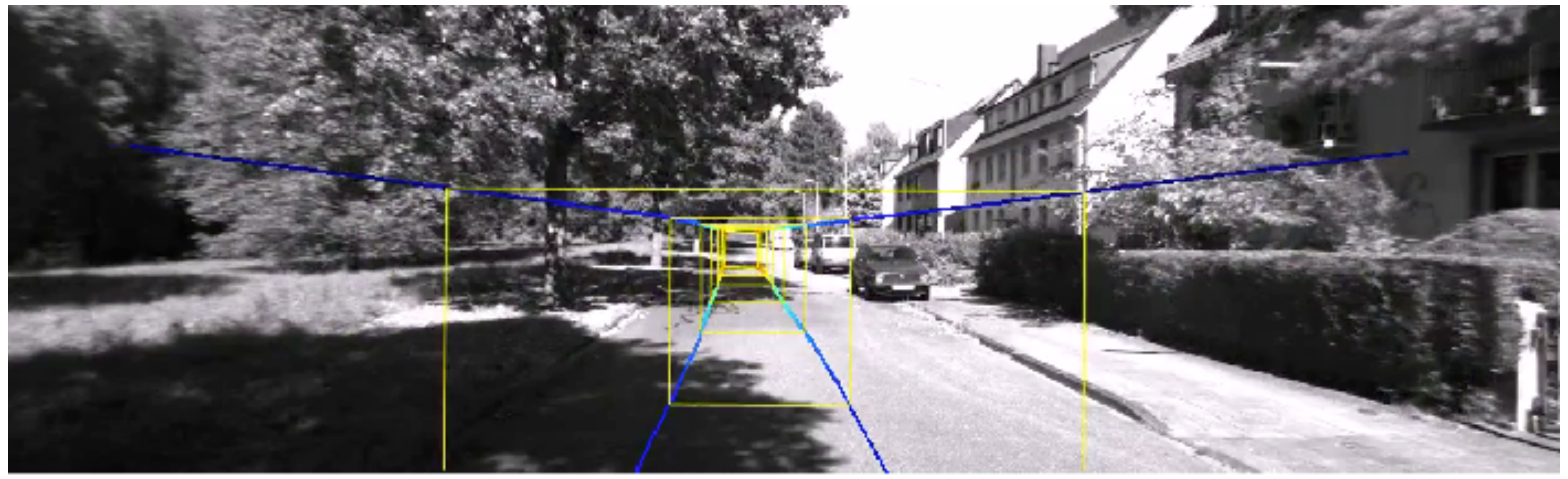
Expressive Language for Robot Data Querying
(Show me X in all robot views, across multiple sessions)

SLAMinDB: Centralized Graph Databases for Mobile Robotics [Fourie et. al 2017]

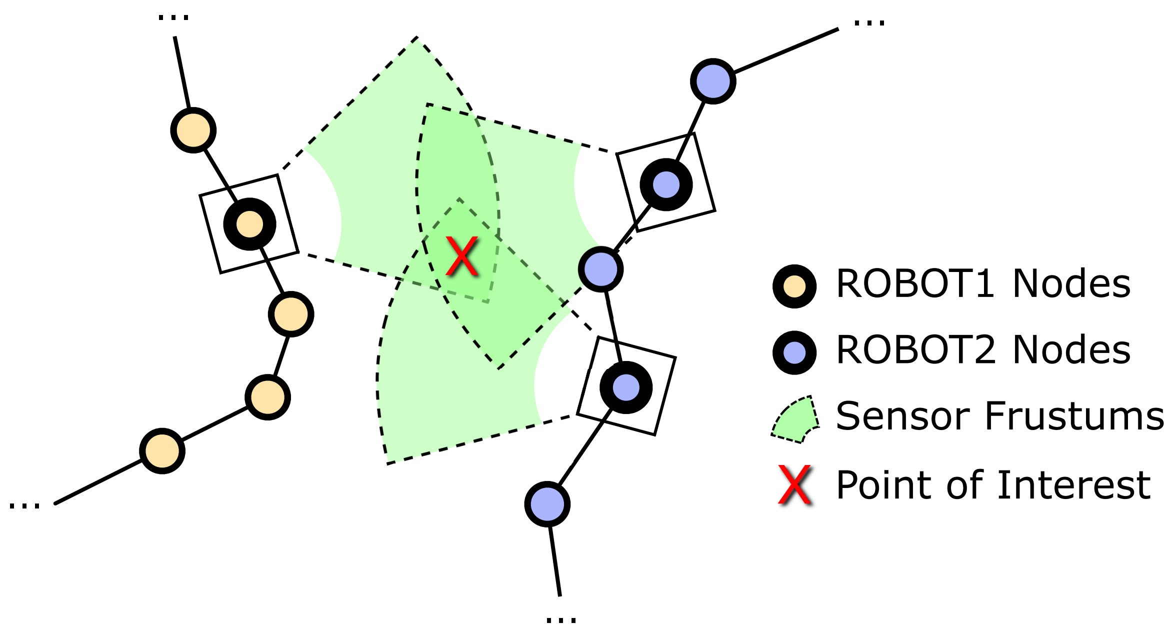
FUTURE DIRECTIONS

- **SLAM as a supervisory signal**
  - Spatially and Semantically-aware Robot DBs
  - Expressive Language for Robot Data Querying
  - Self-Supervised Cross-Modal Learning in Robots





Spatially and Semantically-Aware Robot DBs
(Where have I seen artwork before?)

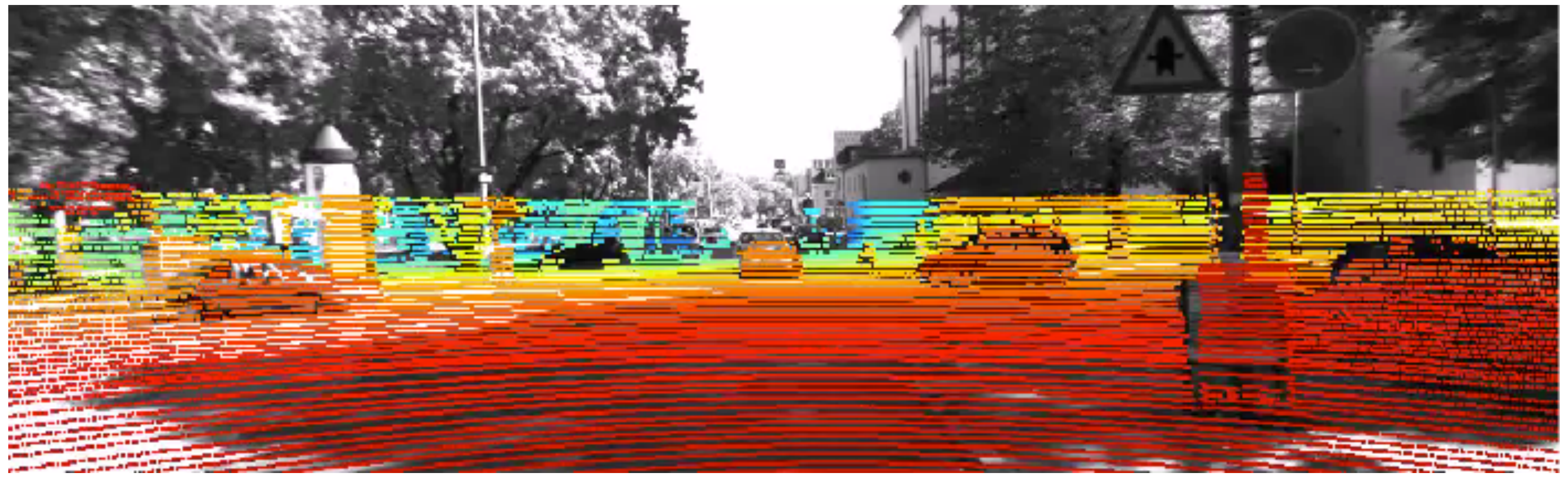


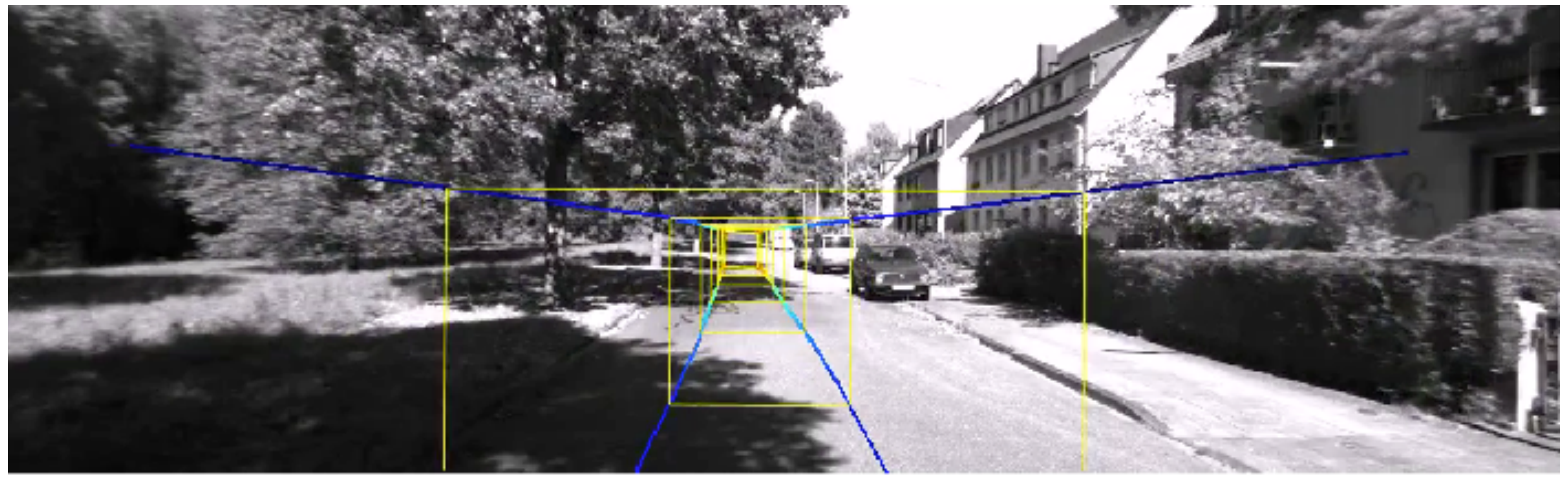
Expressive Language for Robot Data Querying
(Show me X in all robot views, across multiple sessions)

SLAMinDB: Centralized Graph Databases for Mobile Robotics **[Fourie et. al 2017]**

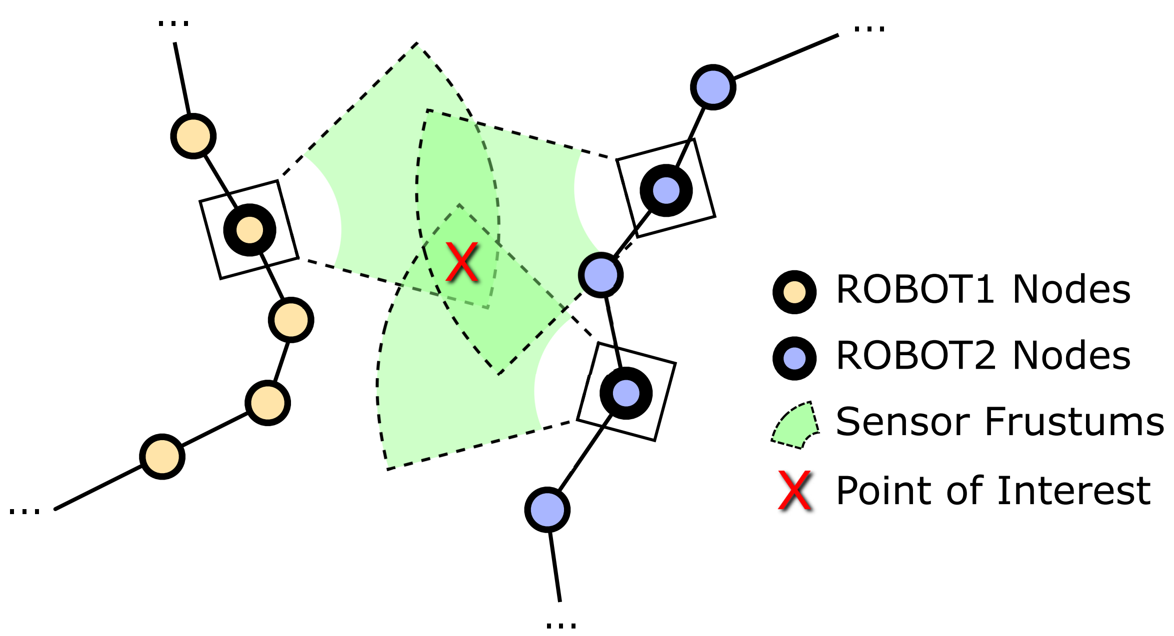
FUTURE DIRECTIONS

- **SLAM as a supervisory signal**
  - Spatially and Semantically-aware Robot DBs
  - Expressive Language for Robot Data Querying
  - Self-Supervised Cross-Modal Learning in Robots
  - Life-long Learning with Simulation





Spatially and Semantically-Aware Robot DBs
(Where have I seen artwork before?)



Expressive Language for Robot Data Querying
(Show me X in all robot views, across multiple sessions)

SLAMinDB: Centralized Graph Databases for Mobile Robotics **[Fourie et. al 2017]**

ACKNOWLEDGEMENTS

John Leonard

Nicholas Roy

Leslie Kaelbling

Antonio Torralba

Thesis Committee

ACKNOWLEDGEMENTS

Seth

Marine Robotics Group

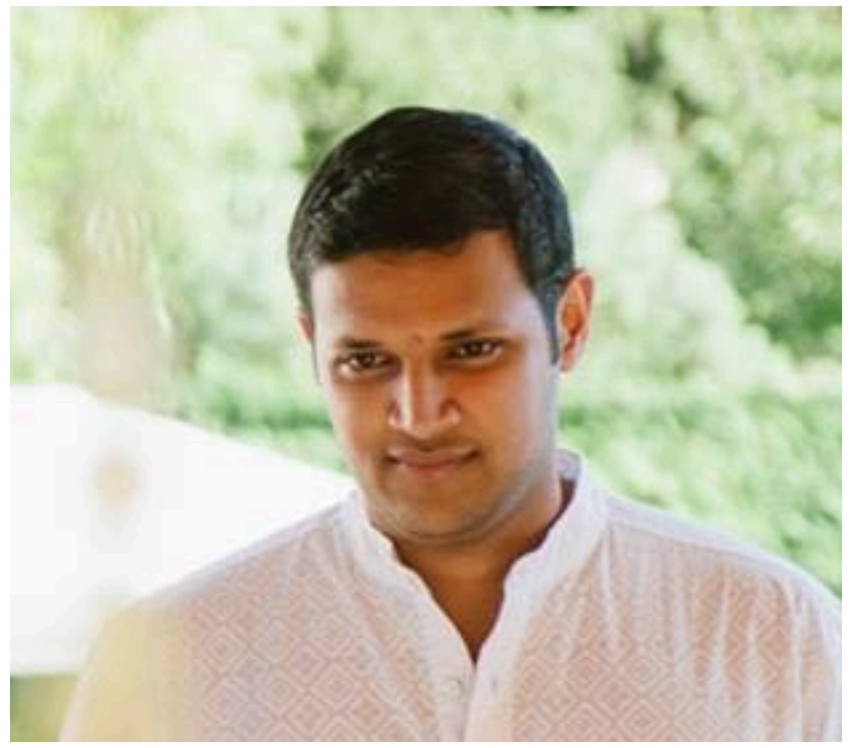
Robotics, Vision and Sensor Networks Group

CSAIL and EECS

MIT Community

ACKNOWLEDGEMENTS

Friends



and many others …

ACKNOWLEDGEMENTS

ACKNOWLEDGEMENTS

Parents

ACKNOWLEDGEMENTS

Sruthi

SLAM-AWARE, SELF-SUPERVISED PERCEPTION IN MOBILE ROBOTS

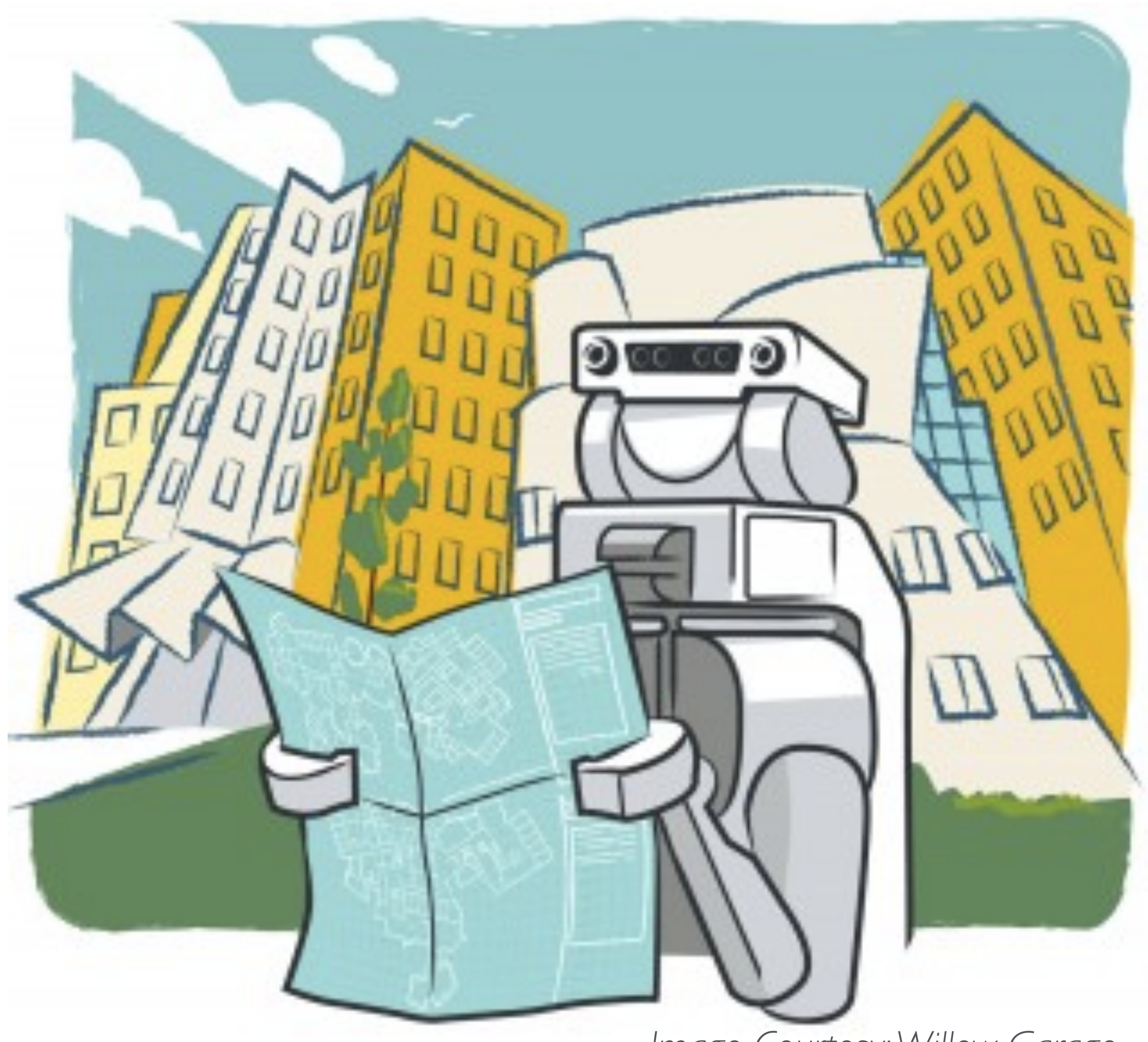


Image Courtesy: Willow Garage

Questions!