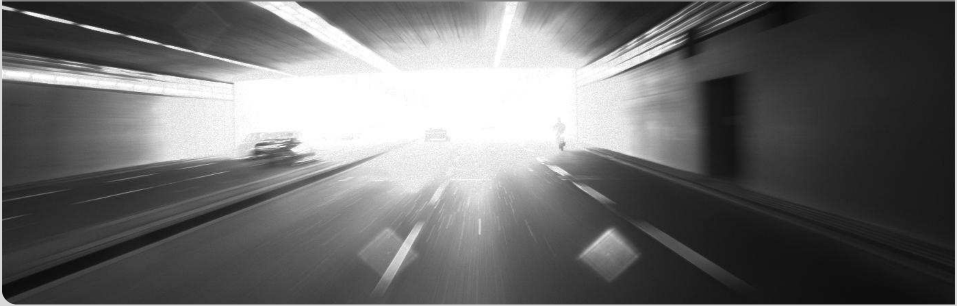
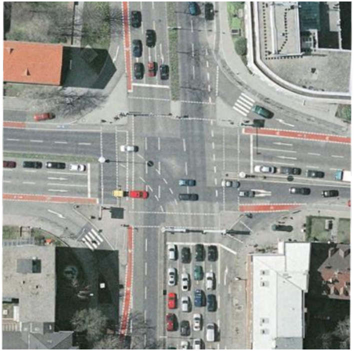
A Generative Model for 3D Urban Scene Understanding from Movable Platforms

Andreas Geiger*, Martin Lauer*, Raquel Urtasun**

*Karlsruhe Institute of Technology **Toyota Technological Institute at Chicago



3D Urban Scene Understanding



Related Work

Urban Scene Segmentation

- Semantic categories: *road*, *vehicles*, *building*, *sky*
- Joint detection and segmentation [Wojek 2008]
- Appearance combined with SfM [Sturgess 2009]

Geometric Methods

- 3D from single images [Hoiem 2007, Saxena 2009]
- Incorporating laws of physics [Gupta 2010]

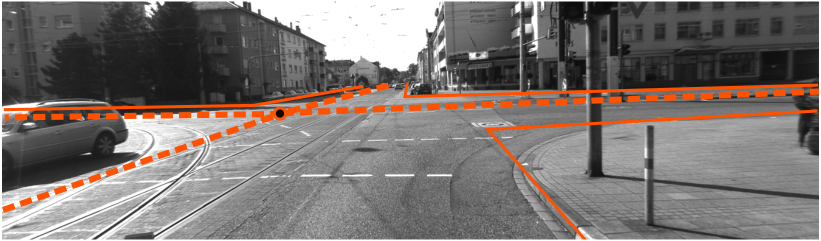
Activity Recognition

- Extraction of typical activity patterns in 2D optical flow
- Non-parametric clustering [Wang 2009, Kuettel 2010]

Our Approach

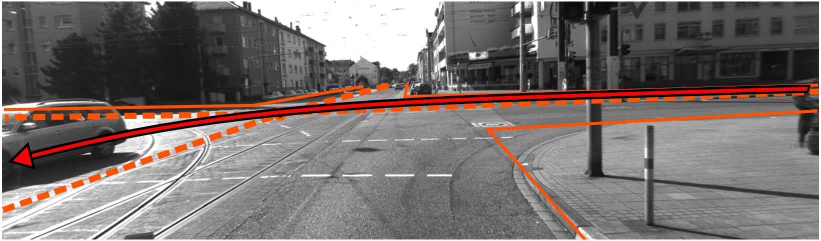
- **Goal**: Infer from short video sequences (moving observer)
  - **Topology** and **geometry** of the scene
  - **Semantic information** (e.g., traffic situation)
- **Probabilistic generative model** of 3D urban scenes
- **Static features:** Building facades
- **Dynamic features:** Moving vehicles

Our Approach



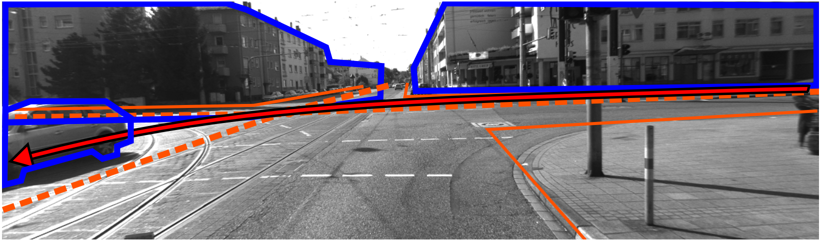
- **Goal**: Infer from short video sequences (moving observer)
  - **Topology** and **geometry** of the scene
  - **Semantic information** (e.g., traffic situation)
- **Probabilistic generative model** of 3D urban scenes
- **Static features:** Building facades
- **Dynamic features:** Moving vehicles

Our Approach



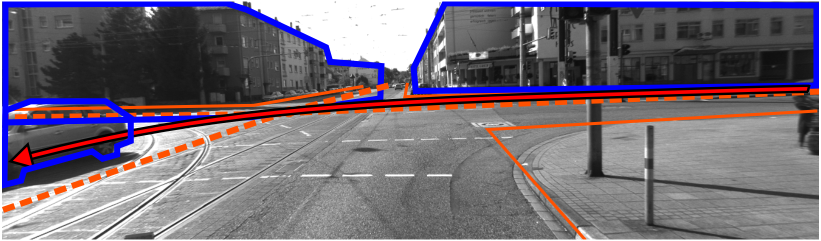
- **Goal**: Infer from short video sequences (moving observer)
  - **Topology** and **geometry** of the scene
  - **Semantic information** (e.g., traffic situation)
- **Probabilistic generative model** of 3D urban scenes
- **Static features:** Building facades
- **Dynamic features:** Moving vehicles

Our Approach



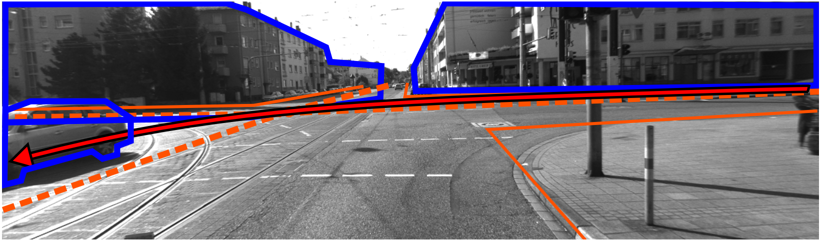
- **Goal**: Infer from short video sequences (moving observer)
  - **Topology** and **geometry** of the scene
  - **Semantic information** (e.g., traffic situation)
- **Probabilistic generative model** of 3D urban scenes
- **Static features:** Building facades
- **Dynamic features:** Moving vehicles

Our Approach



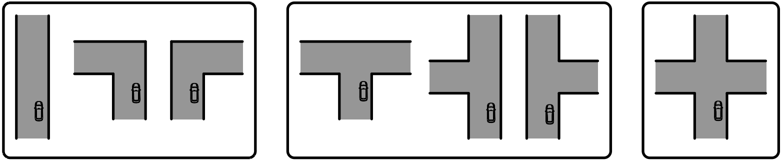
- **Goal**: Infer from short video sequences (moving observer)
  - **Topology** and **geometry** of the scene
  - **Semantic information** (e.g., traffic situation)
- **Probabilistic generative model** of 3D urban scenes
- **Static features:** Building facades
- **Dynamic features:** Moving vehicles

Our Approach



- **Goal**: Infer from short video sequences (moving observer)
  - **Topology** and **geometry** of the scene
  - **Semantic information** (e.g., traffic situation)
- **Probabilistic generative model** of 3D urban scenes
- **Static features:** Building facades
- **Dynamic features:** Moving vehicles

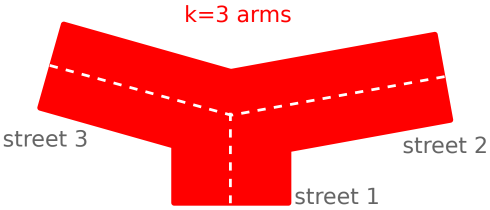
Topology Model



Topology A ($k = 2$) Topology B ($k = 3$) Topology C ($k = 4$)

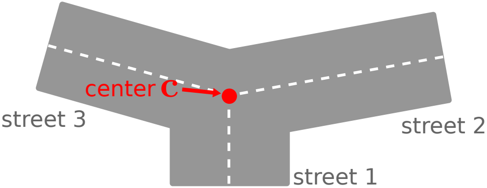
- Reasoning in **bird's eye perspective**

Geometry Model



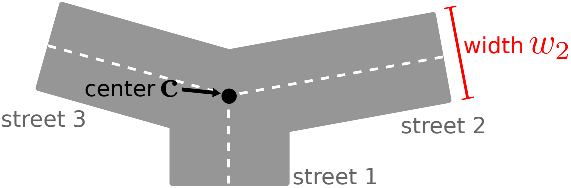
- *k* ... number of intersection arms
- **c** ... center of intersection
- **w** ... street width
- *r* ... global rotation
- **o** ... relative street orientation

Geometry Model



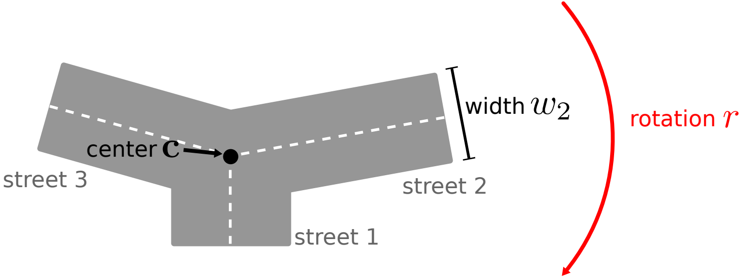
- k ... number of intersection arms
- **c** ... center of intersection
- **w** ... street width
- r ... global rotation
- **o** ... relative street orientation

Geometry Model



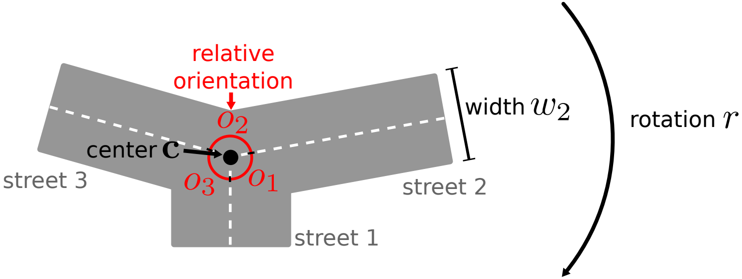
- k ... number of intersection arms
- **c** ... center of intersection
- **w** ... street width
- r ... global rotation
- **o** ... relative street orientation

Geometry Model



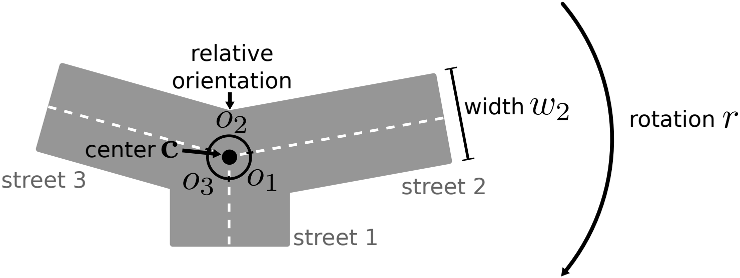
- k ... number of intersection arms
- $\mathbf{c}$ ... center of intersection
- $\mathbf{w}$ ... street width
- r ... global rotation
- $\mathbf{o}$ ... relative street orientation

Geometry Model



- k ... number of intersection arms
- **c** ... center of intersection
- **w** ... street width
- r ... global rotation
- **o** ... relative street orientation

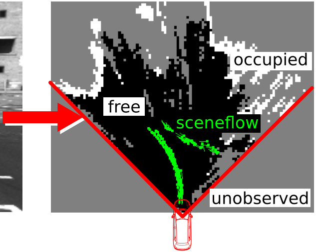
Geometry Model



$$\left.\begin{array}{l}
\blacksquare\ k \text{ ... number of intersection arms} \\
\blacksquare\ \mathbf{c} \text{ ... center of intersection} \\
\blacksquare\ \mathbf{w} \text{ ... street width} \\
\blacksquare\ r \text{ ... global rotation} \\
\blacksquare\ \mathbf{o} \text{ ... relative street orientation}
\end{array}\right\} \mathcal{R}$$

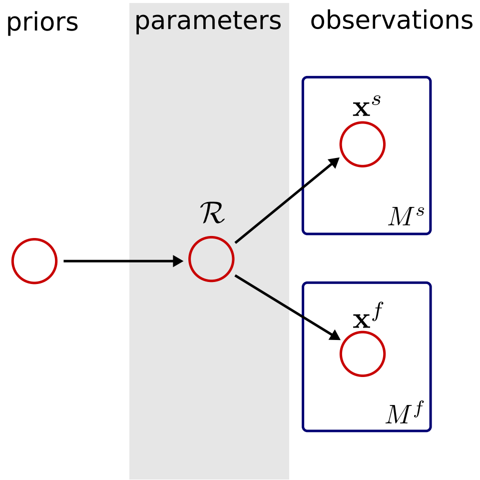
Observations

Karlsruhe Institute of Technology

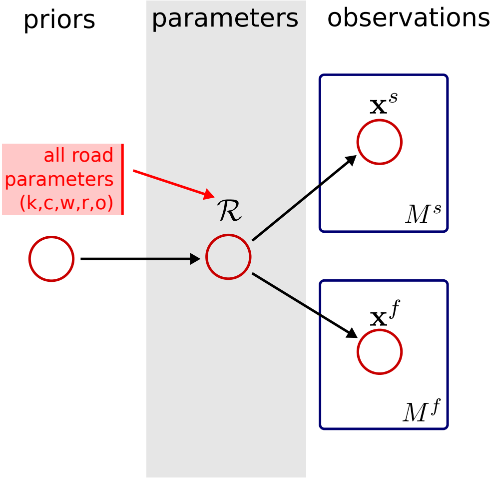


- **Static features:** Occupancy grids. For each cell, a variable indicates if it is occupied (+1), unobserved (0) or free (-1).
- **Dynamic features:** Sparse 3D scene flow
- Registration using **stereo visual odometry**

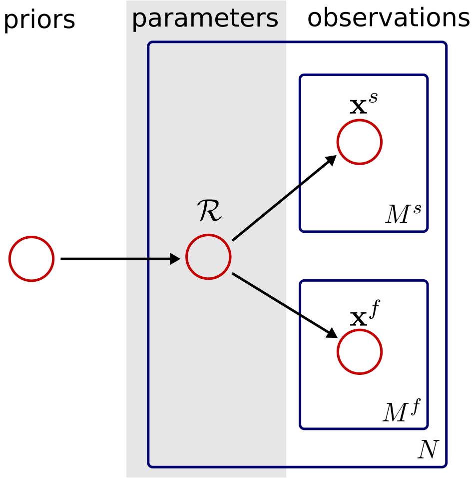
Probabilistic Model



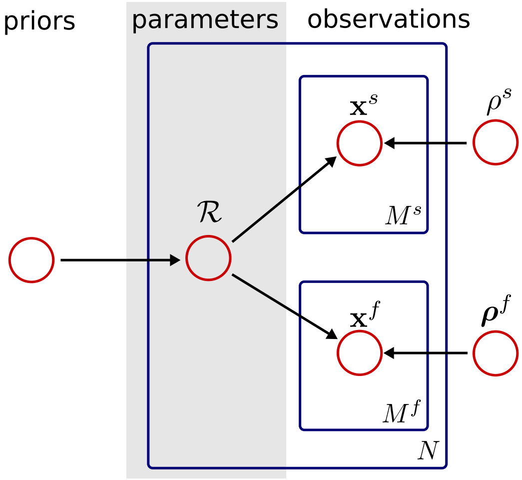
Probabilistic Model



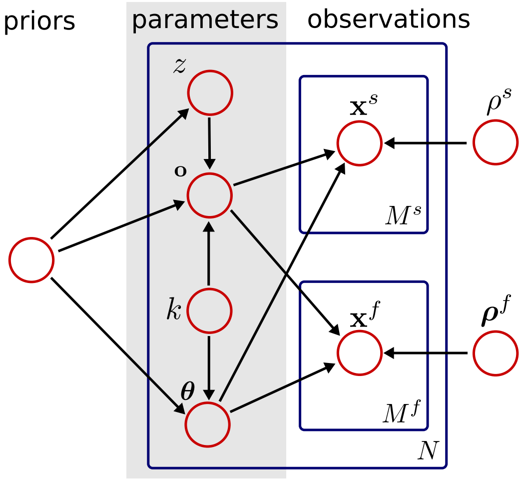
Probabilistic Model



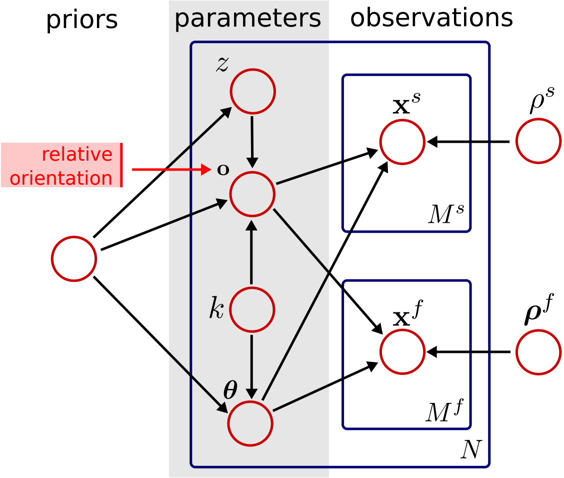
Probabilistic Model



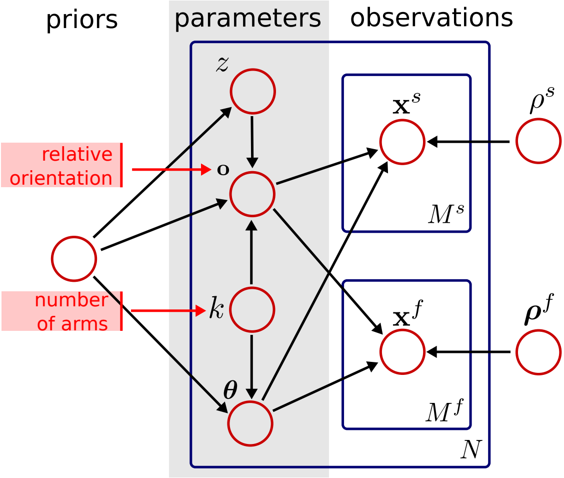
Probabilistic Model



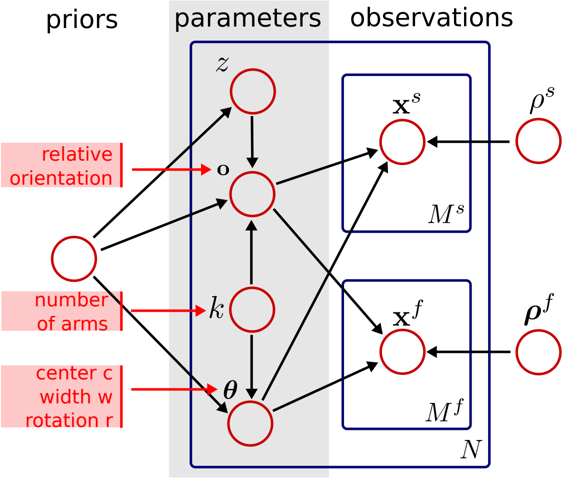
Probabilistic Model



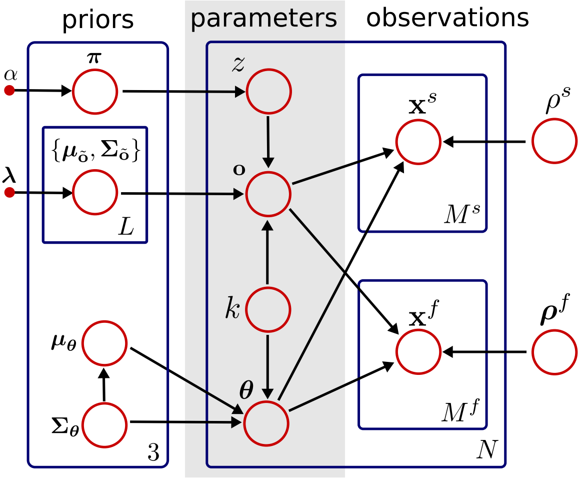
Probabilistic Model



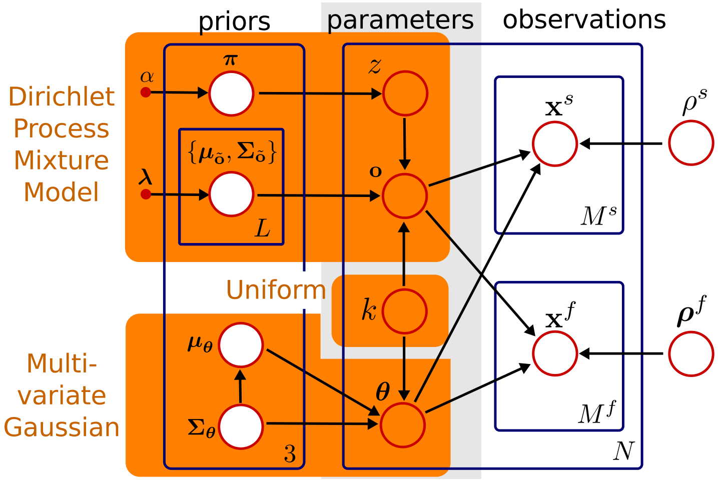
Probabilistic Model



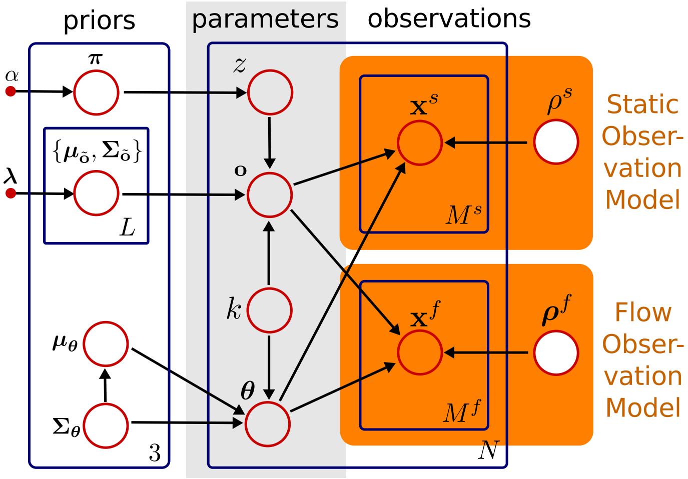
Probabilistic Model



Probabilistic Model



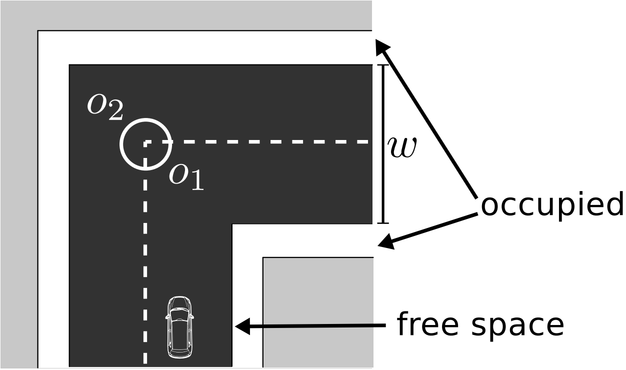
Probabilistic Model



Likelihood: Static Features

- **Static observation** likelihood

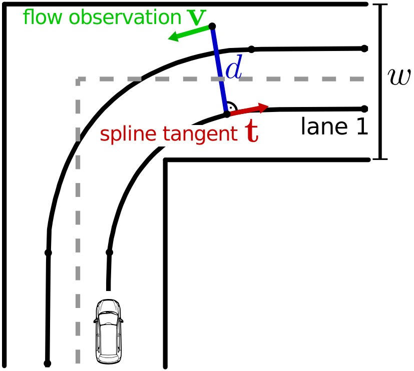
$$p(\mathbf{x}^s|\mathbf{o},\boldsymbol{\theta},\rho^s) \propto \exp\{\beta f(\mathbf{x}^s,\mathbf{o},\boldsymbol{\theta},\rho^s)\}$$



Likelihood: Dynamic Features

- **Flow observation** likelihood

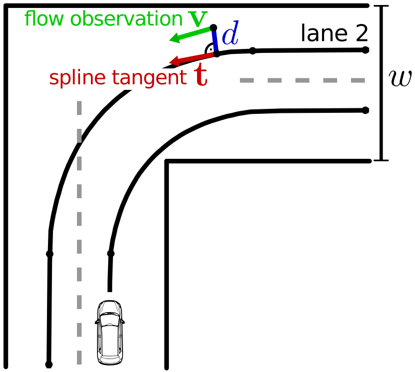
$$p(\mathbf{x}^f|\mathbf{o},\theta,\rho^f) \propto \max_i \exp\left\{-\frac{{d_i}^2}{2\rho_{f1}^2} - \frac{\|\mathbf{v}-\mathbf{t}_i\|_2^2}{2\rho_{f2}^2}\right\}$$



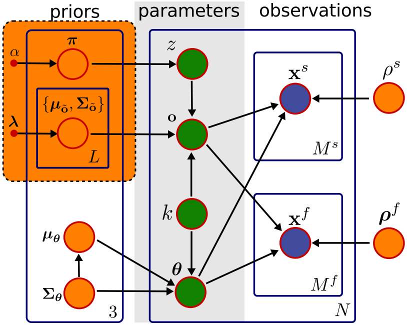
Likelihood: Dynamic Features

- **Flow observation** likelihood

$$p(\mathbf{x}^f|\mathbf{o},\theta,\rho^f) \propto \max_i \ \exp\left\{-\frac{{d_i}^2}{2\rho_{f1}^2} - \frac{\|\mathbf{v}-\mathbf{t}_i\|_2^2}{2\rho_{f2}^2}\right\}$$



Learning & Inference



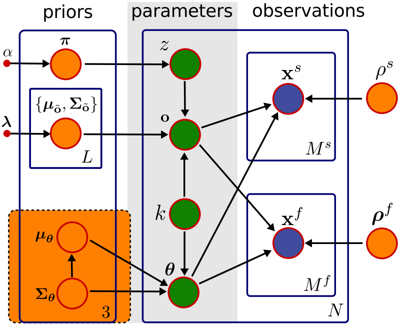
Learning:

- **Orientation:**
  Gibbs sampling with Dirichlet Process Mixture Model (MAP)
- **Center/rotation/width:**
  Maximum Likelihood
- **Observation model:**
  MH sampling (MAP)

Inference:

- **Reversible Jump MCMC:**
- Local MH moves
- Global MH moves
- Reversible jumps

Learning & Inference



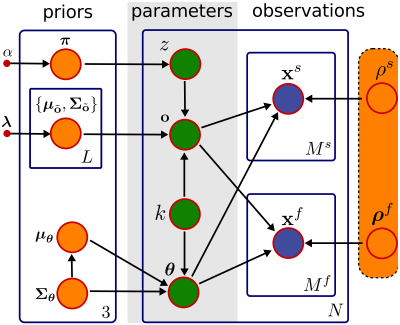
Learning:

- **Orientation:**
  Gibbs sampling with Dirichlet Process Mixture Model (MAP)
- **Center/rotation/width:**
  Maximum Likelihood
- **Observation model:**
  MH sampling (MAP)

Inference:

- **Reversible Jump MCMC:**
- Local MH moves
- Global MH moves
- Reversible jumps

Learning & Inference



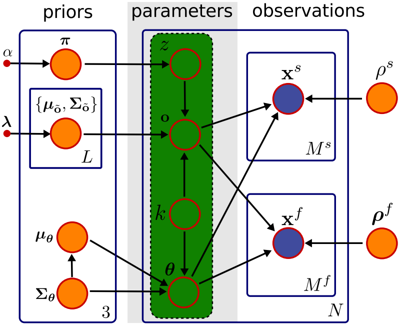
Learning:

- **Orientation:**
  Gibbs sampling with Dirichlet Process Mixture Model (MAP)
- **Center/rotation/width:**
  Maximum Likelihood
- **Observation model:**
  MH sampling (MAP)

Inference:

- **Reversible Jump MCMC:**
- Local MH moves
- Global MH moves
- Reversible jumps

Learning & Inference



Learning:

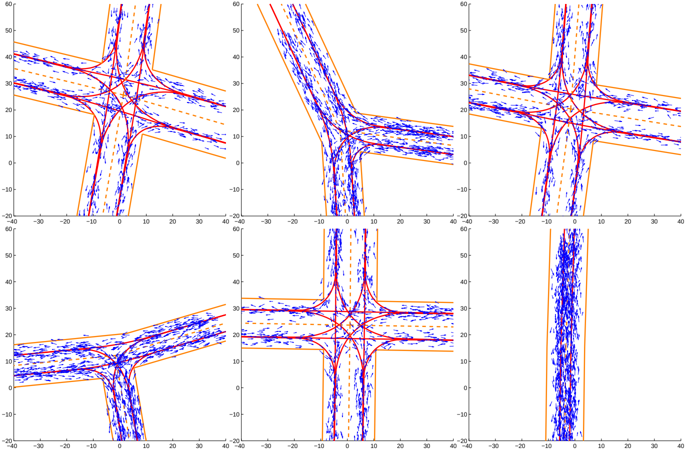
- **Orientation:**
  Gibbs sampling with Dirichlet Process Mixture Model (MAP)
- **Center/rotation/width:**
  Maximum Likelihood
- **Observation model:**
  MH sampling (MAP)

Inference:

- **Reversible Jump MCMC:**
- Local MH moves
- Global MH moves
- Reversible jumps

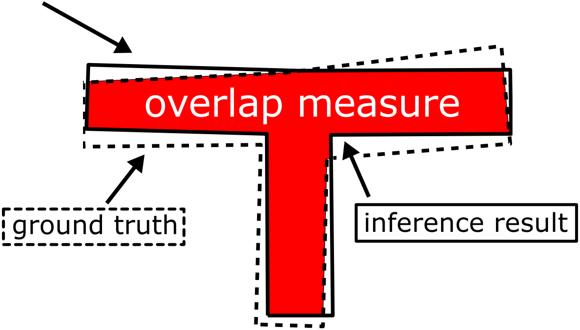
Sampling the Model

- **Samples** from the prior:



Estimating Geometry and Topology

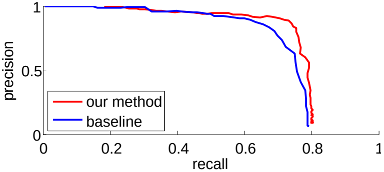
	GP Regression	Ours
Accuracy in estimating *k*	61.1 %	92.9 %
Location error	5.4 m	4.4 m
Orientation error	14.1 deg	6.6 deg
Overlap measure	49.3 %	62.7 %



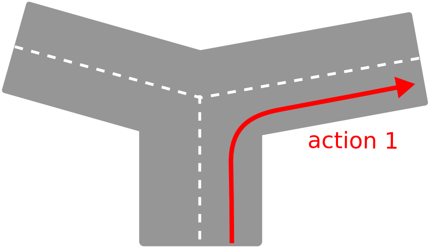
Improving Object Detection

- **Re-weight scores** of [Felzenszwalb08] using spatial prior

- Increase in **average precision** from 71.3 % to 74.9 %



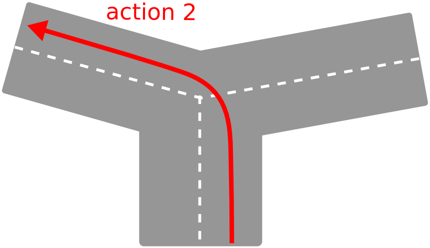
Extracting Semantic Activities



- **Activity:** $k(k-1)$ dimensional binary vector $\mathbf{a}$
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	Ours
Hamming distance	0.16	0.08

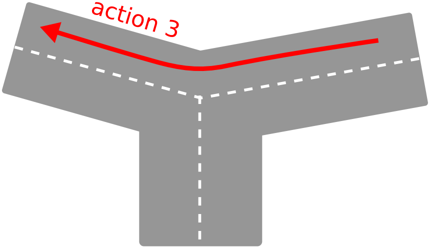
Extracting Semantic Activities



- **Activity:** $k(k-1)$ dimensional binary vector $\mathbf{a}$
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	**Ours**
Hamming distance	0.16	0.08

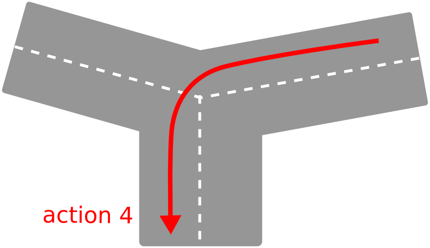
Extracting Semantic Activities



- **Activity:** $k(k-1)$ dimensional binary vector $\mathbf{a}$
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	Ours
Hamming distance	0.16	0.08

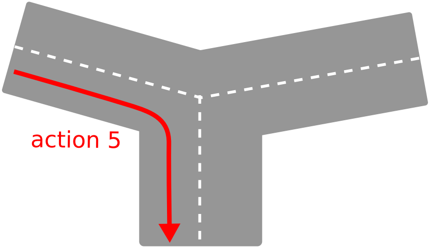
Extracting Semantic Activities



- **Activity:** $k(k-1)$ dimensional binary vector **a**
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	**Ours**
Hamming distance	0.16	0.08

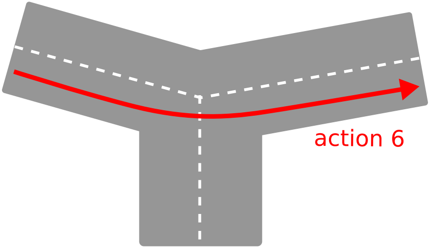
Extracting Semantic Activities



- **Activity:** $k(k-1)$ dimensional binary vector $\mathbf{a}$
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	**Ours**
Hamming distance	0.16	0.08

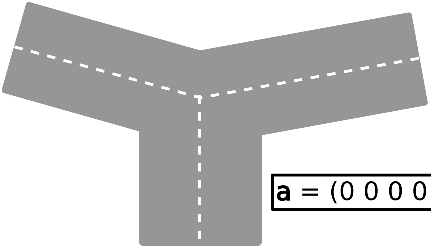
Extracting Semantic Activities



- **Activity:** $k(k-1)$ dimensional binary vector $\mathbf{a}$
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	**Ours**
Hamming distance	0.16	0.08

Extracting Semantic Activities

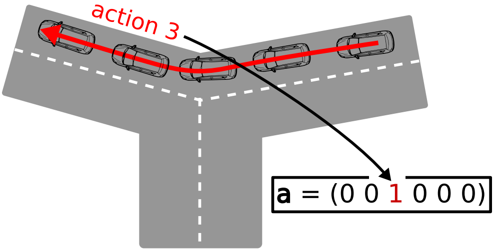


$\boxed{\mathbf{a} = (0\ 0\ 0\ 0\ 0\ 0)}$

- **Activity:** $k(k-1)$ dimensional binary vector $\mathbf{a}$
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	Ours
Hamming distance	0.16	0.08

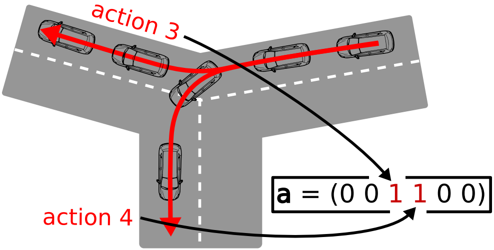
Extracting Semantic Activities



- **Activity:** $k(k-1)$ dimensional binary vector **a**
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	**Ours**
Hamming distance	0.16	0.08

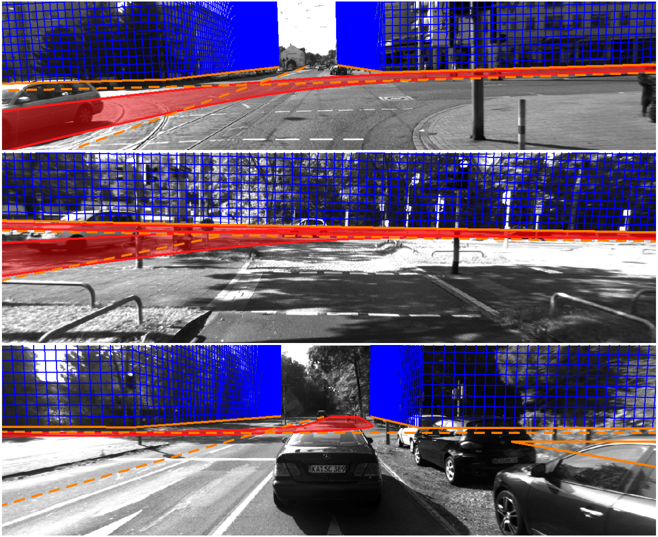
Extracting Semantic Activities



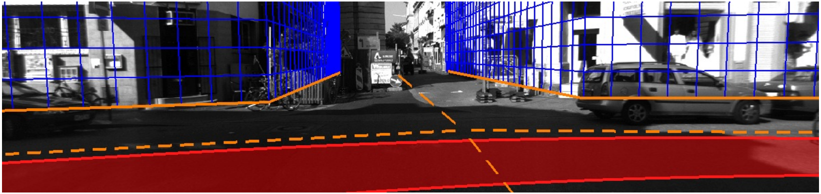
- **Activity:** $k(k-1)$ dimensional binary vector **a**
- Error measure: **Normalized Hamming distance**
- Results:

	GP regression	**Ours**
Hamming distance	0.16	0.08

Inference Results



Summary and Future Work



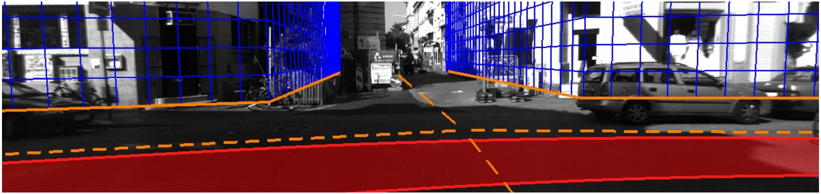
Conclusion:

- **Generative model** of **3D urban scenes**
- **Static** and **dynamic** features
- Improved **object detection** & **activity recognition**

Future work:

- **Features**: Vanishing points, scene labels, ...
- **Joint object and scene layout** reasoning
- Scene understanding with a **single camera**

Summary and Future Work



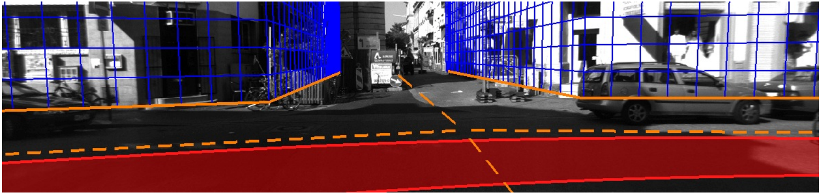
Conclusion:

- **Generative model** of **3D urban scenes**
- **Static** and **dynamic** features
- Improved **object detection** & **activity recognition**

Future work:

- **Features**: Vanishing points, scene labels, **...**
- **Joint object and scene layout** reasoning
- Scene understanding with a **single camera**

Summary and Future Work



Conclusion:

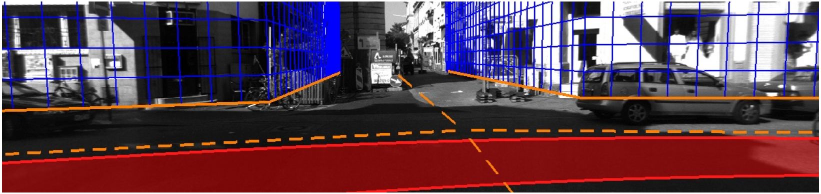
- **Generative model** of **3D urban scenes**
- **Static** and **dynamic** features
- Improved **object detection** & **activity recognition**

Future work:

- **Features**: Vanishing points, scene labels, **...**
- **Joint object and scene layout** reasoning
- Scene understanding with a **single camera**

Summary and Future Work

Karlsruhe Institute of Technology



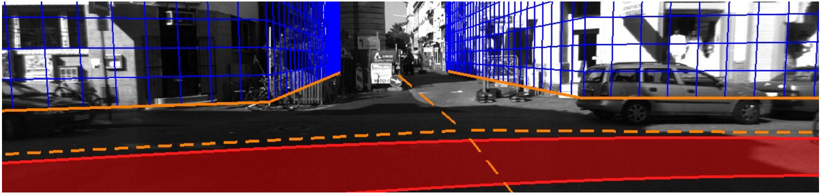
Conclusion:

- **Generative model** of **3D urban scenes**
- **Static** and **dynamic** features
- Improved **object detection** & **activity recognition**

Future work:

- **Features**: Vanishing points, scene labels, **...**
- **Joint object and scene layout** reasoning
- Scene understanding with a **single camera**

Summary and Future Work



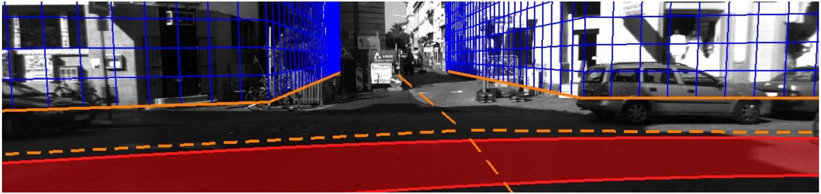
Conclusion:

- **Generative model** of **3D urban scenes**
- **Static** and **dynamic** features
- Improved **object detection** & **activity recognition**

Future work:

- **Features**: Vanishing points, scene labels, ...
- **Joint object and scene layout** reasoning
- Scene understanding with a **single camera**

Summary and Future Work



Conclusion:

- **Generative model** of **3D urban scenes**
- **Static** and **dynamic** features
- Improved **object detection** & **activity recognition**

Future work:

- **Features**: Vanishing points, scene labels, **...**
- **Joint object and scene layout** reasoning
- Scene understanding with a **single camera**

Geometry Model

Unique orientation vector **o**, constrained to the Δ^{k-1} simplex

$$\sum_{i=1}^{k} o_i = 1 \qquad o_i \geq 0 \qquad \alpha_i = r + 2\pi \sum_{j=1}^{i-1} o_j$$