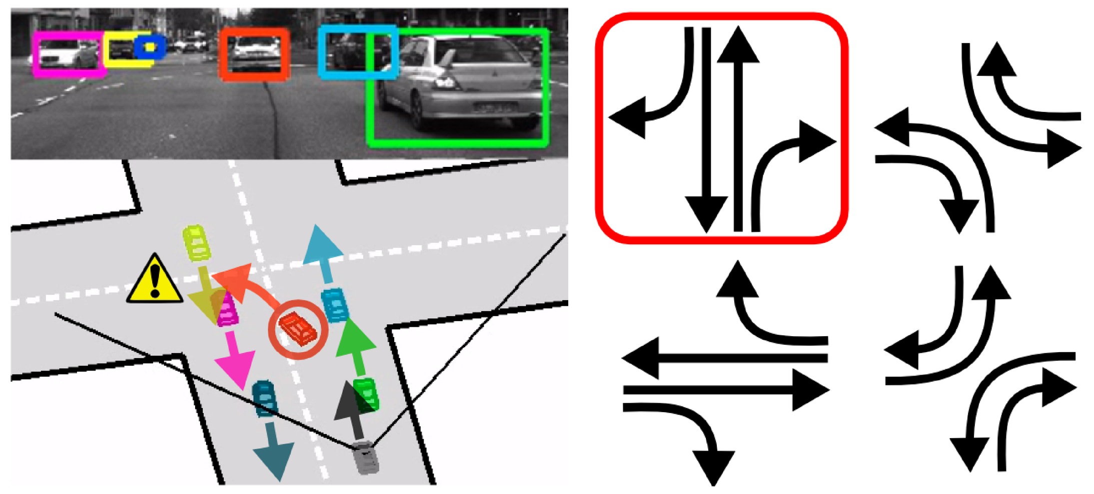
Understanding High-Level Semantics by Modeling Traffic Patterns

{ Hongyi Zhang*, Andreas Geiger† and Raquel Urtasun‡ } *Peking University, †MPI Tübingen, ‡TTI Chicago

Introduction



Given a short monocular video sequence from a movable platform we propose a joint probabilistic model for estimating:

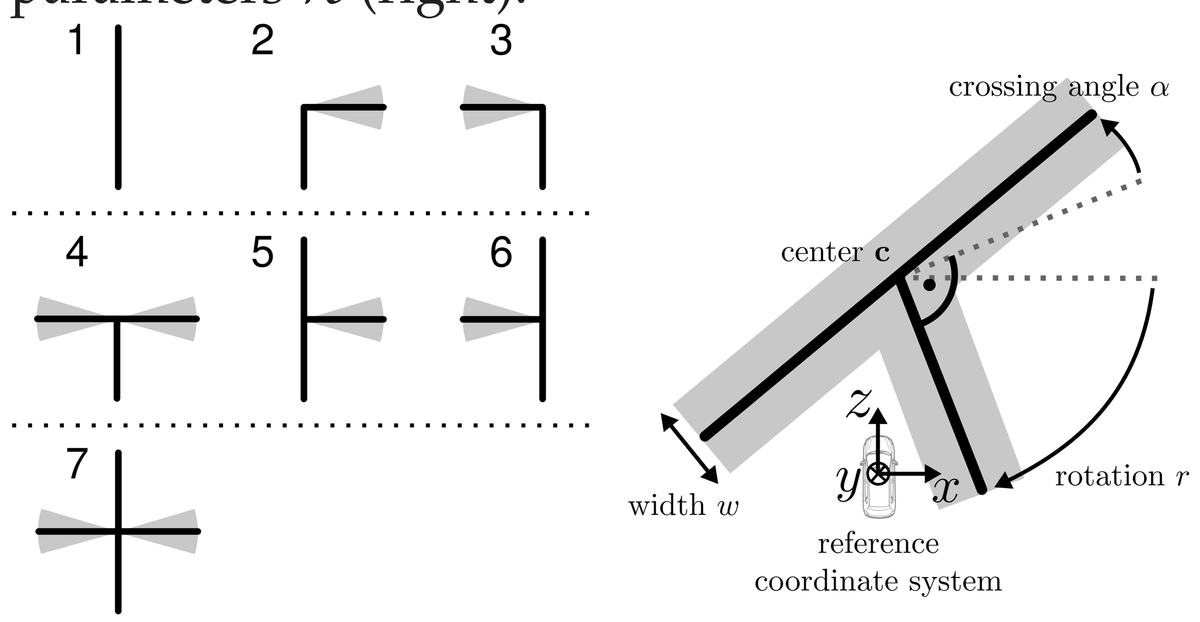
- The 3D urban scene layout
- The objects (e.g., cars) in the scene

Contributions with respect to [1]:

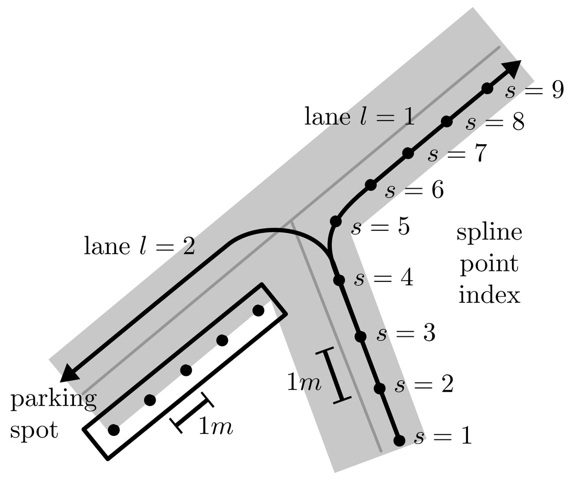
- Model for Traffic patterns
- Interactions between tracklets
- Novel dynamical model

Topology and Geometry

We model street scenes in **bird's eye perspective** using 7 scene layouts θ (left) and the geometry parameters $\mathcal{R}$ (right):



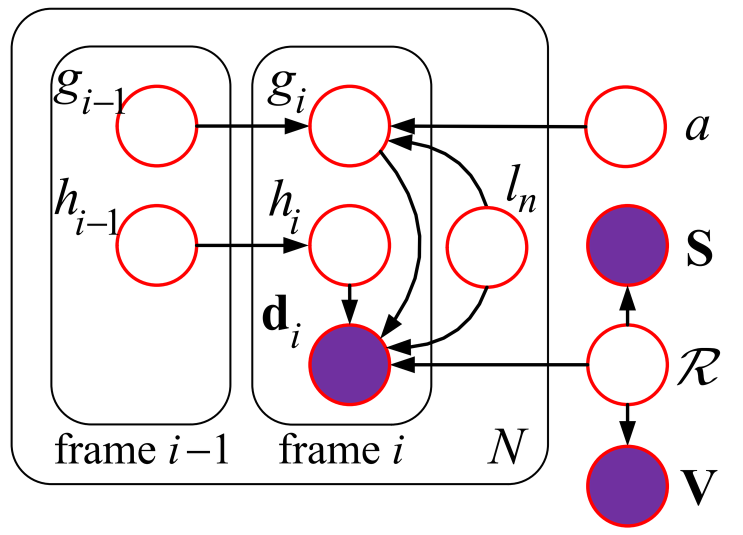
We model the set of possible vehicle locations with lanes connecting the streets and parking spots at the road side:



We have:

- K streets
- $K(K-1)$ lanes
- $2K$ parking spots

Probabilistic Model

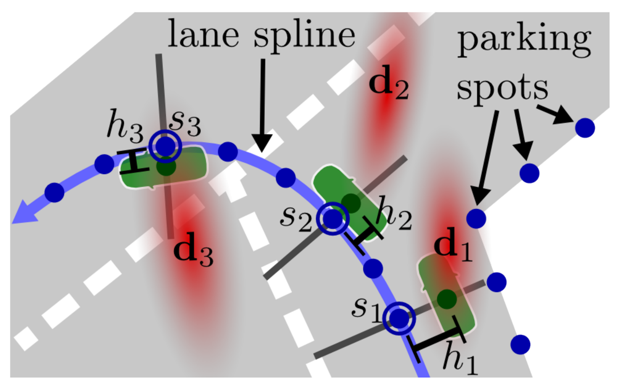


Variables:

- $\mathcal{R}$: Road parameters (width, rotation, etc.)
- a: Traffic patterns (i.e. traffic signal phase)
- l_n: Lane n-th tracklet is driving on
- (g_i, h_i): Vehicle dynamics
- $\mathbf{d}_i$: Tracklet detection (location, heading)
- **S**: Scene label evidence
- **V**: Vanishing point evidence

Tracklets

Detection likelihood:



$$p(\mathbf{d}_i|g_i, h_i) = \mathcal{N}((s_i, h_i), (\xi\Lambda_{\mathbf{d}_i})^{-1}) \times p^{\gamma}_{\text{heading}}$$

- $g = (s, b)$, s: spline point, $b \in \{\text{stop}, \text{go}\}$
- $\Lambda_{\mathbf{d}_i}$: tracklet precision matrix
- p_{heading}: heading probability
- ξ, γ: model parameters

Forward dynamics:

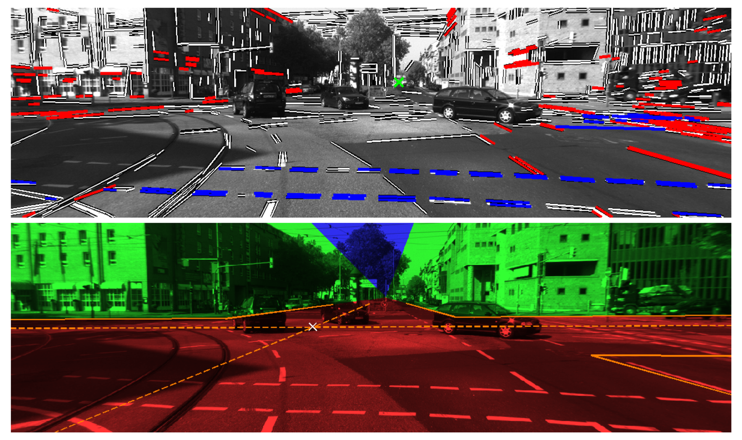
$$p(g_i|g_{i-1}) = \begin{cases} p(b_i|b_{i-1})\pi(\cdot) & \text{if } b_i = \text{go} \\ p(b_i|b_{i-1}) & \text{if } b_i = \text{stop} \wedge s_i = s_{i-1} \\ 0 & \text{if } b_i = \text{stop} \wedge s_i \neq s_{i-1} \end{cases}$$

where the transition probability $p(b_i|\cdot)$ also depends on a, l which decide if the lane is active, and $\pi(\cdot)$ models the driving speed.

Lateral dynamics:

- $h_i = h_{i-1} + \Delta\sigma_h^2$ (Gaussian noise)

Vanish. Points / Sem. Labels



Inference

Inferring road geometry:

- Simulated annealing (MH sampling)
- Mixture of local and global moves

Inferring traffic patterns:

$$p(a|\mathbf{T}, \mathcal{R}) \propto \prod_{n=1}^{N} \sum_{l_n} p(\mathbf{t}_n|a, l_n, \mathcal{R})$$

Inferring car-to-lane associations:

$$p(l_n|a, \mathbf{t}_n, \mathcal{R}) \propto p(\mathbf{t}_n|a, l_n, \mathcal{R}).$$

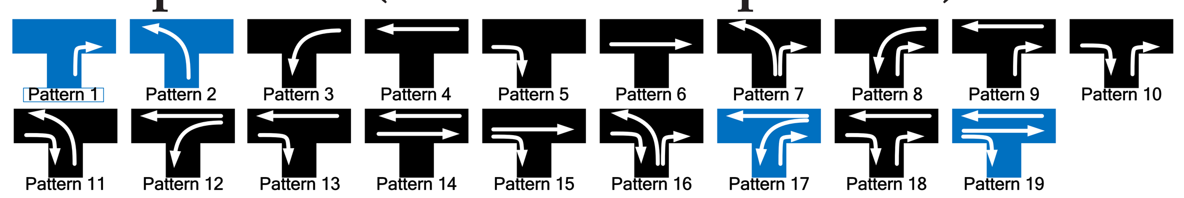
where a tracklet $\mathbf{t} = \{\mathbf{d}_1, \ldots, \mathbf{d}_M\}$ is represented by the set of its detections and $p(\mathbf{t}_n|a, l_n, \mathcal{R})$ can be approximated by Expectation Propagation.

Learning

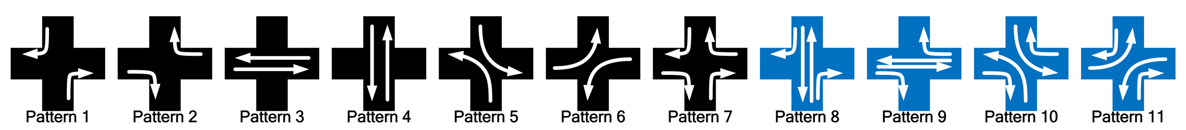
Learning traffic patterns:

- Enumerate all combinations of K patterns
- Score them by number of correct tracklets
- 4 patterns explain most scenarios

3-arm patterns (blue: learned patterns):



4-arm patterns (blue: learned patterns):

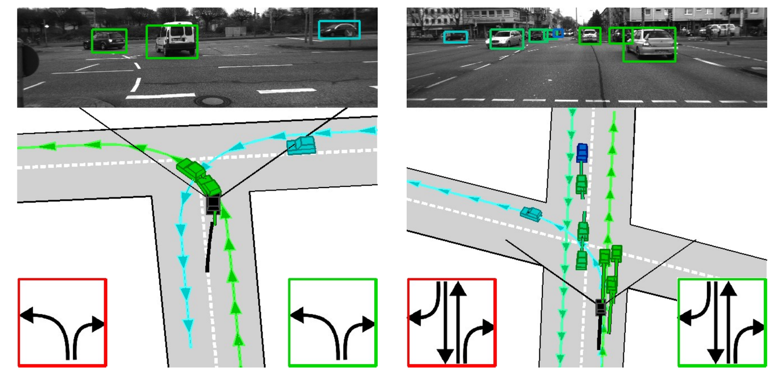


Learning forward dynamics:
Estimate $p(b_{i-1}, b_i)$ on active/inactive lane separately (S: stop states, G: go states)

Lane State	S→S	G→S	S→G	G→G
Inactive	0.888	0.017	0.015	0.080
Active	0.027	0.010	0.005	0.958

Results

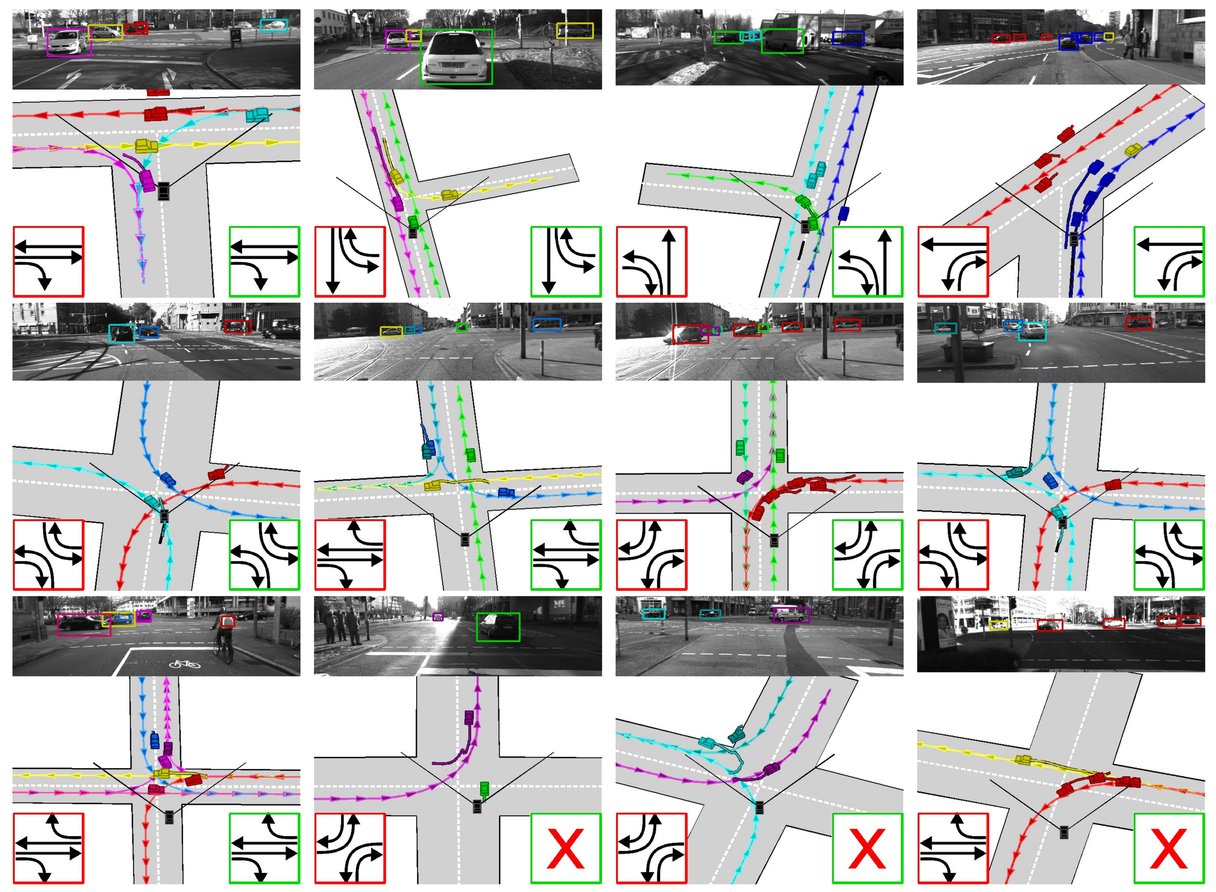
Case study:



Red: Inferred pattern
Green: True pattern

Left: traffic pattern disambiguates lane association of the static car (rightmost).
Right: Correct inference result for scene from the Introduction. [1] infers colliding vehicles.

Qualitative Results:



Pattern and car-to-lane association error:

Method	T-L error (all) 3-arm	T-L error (all) 4-arm	T-L error (>10m) 3-arm	T-L error (>10m) 4-arm	Pattern error 3-arm	Pattern error 4-arm
[1]	46.7%	49.9%	17.9%	30.1%	–	–
Ours	**15.2%**	**30.1%**	**3.6%**	**14.0%**	**18.2%**	**19.4%**

Road geometry estimation:

Method	Location 3-arm	Location 4-arm	Orientation 3-arm	Orientation 4-arm	Overlap 3-arm	Overlap 4-arm
[1]	**4.3 m**	5.4 m	3.3 deg	8.0 deg	58.7%	56.0%
Ours	5.7 m	**4.9 m**	**2.4 deg**	**4.3 deg**	**61.5%**	**61.3%**

References

[1] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, Granada, Spain, December 2011.