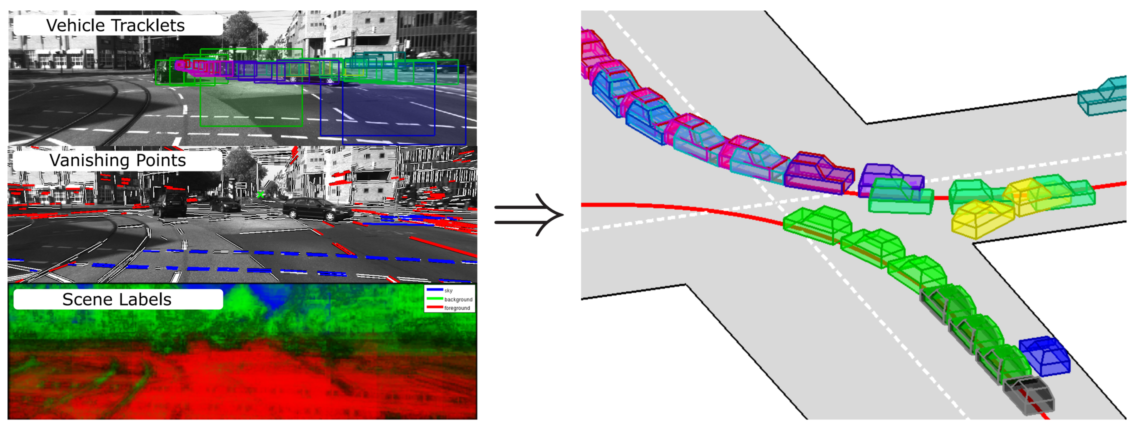
Joint 3D Estimation of Objects and Scene Layout

Andreas Geiger
KIT

Christian Wojek
MPI Saarbrücken

Raquel Urtasun
TTI Chicago

Overview



We propose a novel **generative model** that is able to reason jointly about the **3D scene layout** and the location / orientation of objects.

Input:

- Urban video sequences (∼10 s)
- Monocular, Grayscale @ 10 Hz

Features:

- Vehicle tracklets (tracking-by-detection)
- Semantic scene labels (joint boosting)
- Vanishing points (long lines)

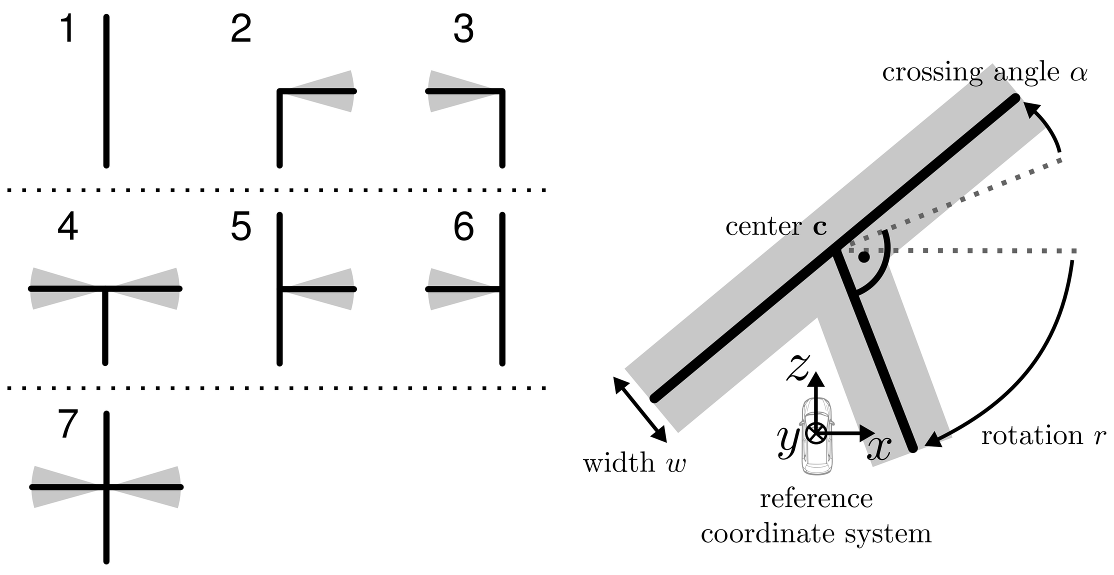
Inference:

- Metropolis-Hastings sampling
- Dynamic programming

Output:

- Street layout (road size, orientation)
- Object locations and orientations
- Intersection crossing activities

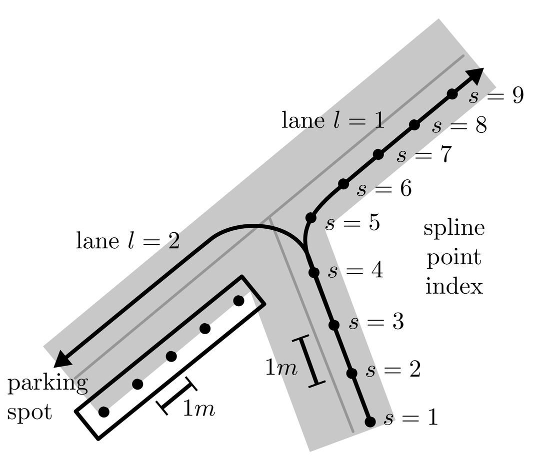
Topology and Geometry



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

- Center of intersection $\mathbf{c}$
- Street width w
- Scene rotation r
- Crossing street angle α

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



We have:

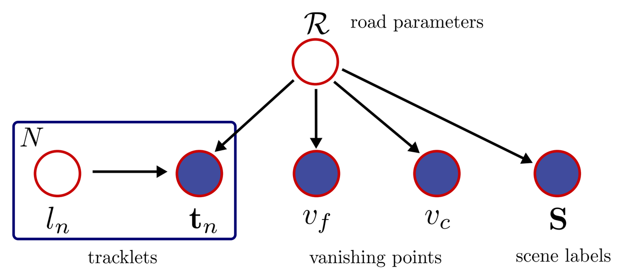
- K streets
- $K(K-1)$ lanes
- $2K$ parking spots

Future Work

- Find more discriminative cues
- Extend inference (e.g., pedestrians)
- Holistic scene interpretation incorporating buildings, infrastructure, vegetation and other types of vehicles

Probabilistic Model

Our goal is to estimate the most likely configuration $\mathcal{R} = (\theta, \mathbf{c}, w, r, \alpha)$ given the image evidence $\mathcal{E} = \{\mathbf{T}, \mathbf{V}, \mathbf{S}\}$, which comprises *vehicle tracklets* $\mathbf{T} = \{\mathbf{t}_1, .., \mathbf{t}_N\}$, *vanishing points* $\mathbf{V} = \{v_f, v_c\}$ and *semantic labels* $\mathbf{S}$. Given $\mathcal{R}$ we assume all observations to be independent:



$$p(\mathcal{E}, \mathcal{R}) = \underbrace{p(\mathcal{R})}_{\text{Prior}} \underbrace{\left[\prod_{n=1}^{N} \sum_{l_n} p(\mathbf{t}_n, l_n | \mathcal{R})\right]}_{\text{Vehicle Tracklets}} \underbrace{p(v_f|\mathcal{R}) \; p(v_c|\mathcal{R})}_{\text{Vanishing Points}} \underbrace{p(\mathbf{S}|\mathcal{R})}_{\text{Semantic Labels}}$$

Prior

$$\begin{aligned} p(\mathcal{R}) &= p(\theta)p(\mathbf{c}, w)p(r)p(\alpha) \\ \theta &\sim \delta(\theta_{MAP}) \\ (\mathbf{c}, \log w)^T &\sim \mathcal{N}(\boldsymbol{\mu}_{cw}, \boldsymbol{\Sigma}_{cw}) \\ r &\sim \mathcal{N}(\mu_r, \sigma_r) \\ \alpha &\sim p(\alpha) \end{aligned}$$

- Road parameters: $\mathcal{R} = (\theta, \mathbf{c}, w, r, \alpha)$
- Scene layout: $\theta \in \{1, .., 7\}$
- Intersection center: $\mathbf{c} \in \mathbb{R}^2$
- Street width: $w \in \mathbb{R}_+$
- Scene rotation: $r \in [-\frac{\pi}{4}, \frac{\pi}{4}]$
- Crossing street angle: $\alpha \in [-\frac{\pi}{4}, \frac{\pi}{4}]$

Vehicle Tracklets

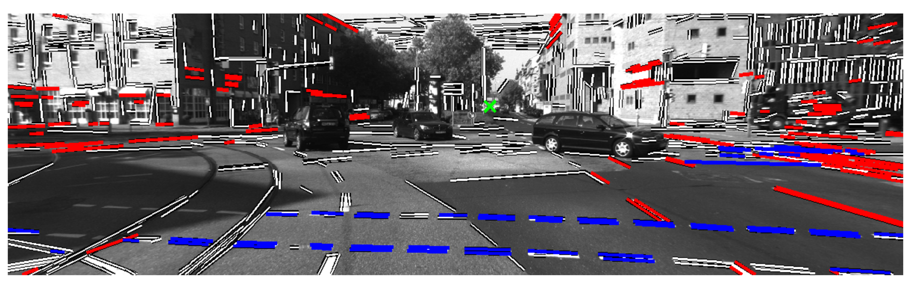
$$\begin{aligned} p(\mathbf{t}|l, \mathcal{R}) &= \sum_{s_1,...,s_M} p(s_1)p(\mathbf{d}_1|s_1, l, \mathcal{R}) \\ &\quad \prod_{m=2}^{M} p(s_m|s_{m-1})p(\mathbf{d}_m|s_m, l, \mathcal{R}) \\ p(\mathbf{d}|s, l, \mathcal{R}) &= p(\mathbf{o}|s, l, \mathcal{R})p(\mathbf{b}|s, l, \mathcal{R}) \\ p(\mathbf{o}|s, l, \mathcal{R}) &= \text{Mult}(\mathbf{o}|\boldsymbol{\pi}_o) \\ p(\mathbf{b}|s, l, \mathcal{R}) &\propto \mathcal{N}(\boldsymbol{\pi}(\mathbf{b})|\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) + \text{const} \end{aligned}$$

- Tracklet: $\mathbf{t} = \{\mathbf{d}_1, .., \mathbf{d}_M\}$
- Lane: $l \in \{1, .., K(K-1) + 2K\}$
- Location on lane: $s \in \mathbb{N}$
- Detection: $\mathbf{d} = (\mathbf{o} \in \{1, .., 8\}, \mathbf{b} \in \mathbb{R}^4)$

Vanishing Points

$$\begin{aligned} p(v_f|\mathcal{R}) &\propto \delta_f \, \mathcal{N}(v_f|\mu_f, \sigma_f) + \text{const} \\ p(v_c|\mathcal{R}) &\propto \delta_c \, \mathcal{N}(v_c|\mu_c, \sigma_c) + \text{const} \end{aligned}$$

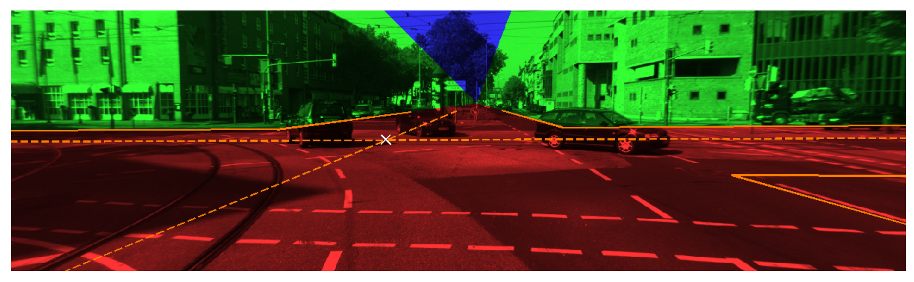
- Forward vanishing point: $v_f \in \mathbb{R}$
- Crossing vanishing point: $v_c \in [-\frac{\pi}{4}, \frac{\pi}{4}]$
- Existance variables: $\delta_f, \delta_c \in \{0, 1\}$



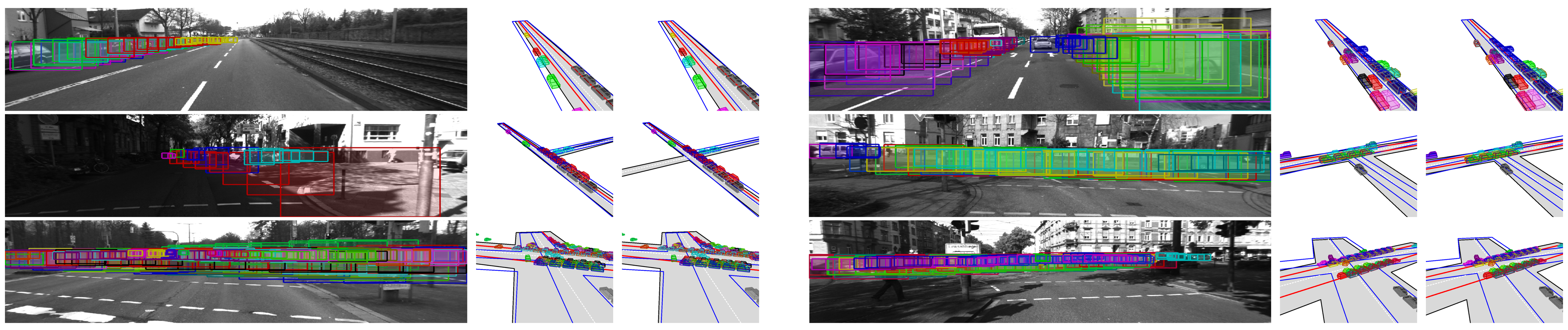
Semantic Labels

$$p(\mathbf{S}|\mathcal{R}) \propto \exp(\gamma \sum_{c} \sum_{(u,v)\in\mathcal{S}_c} S_{u,v}^{(c)})$$

- Class: $c \in \{road, building, sky\}$
- Pixels of reprojected model: $\mathcal{S}_c$
- Scene label probability: $S_{u,v}$



Results



Automatically inferred scene descriptions: Tracklets from all frames superimposed (left). Inference result with θ known (middle) and θ unknown (right). The inferred intersection layout is shown in gray, ground truth labels are given in blue and detected activities are depicted in red.

		Topology	Location	Orientation	Overlap	Activity
Topology known	GP MKL	–	6.0 m	9.6 deg	44.9 %	18.4 %
	Ours	–	**5.8 m**	**5.9 deg**	**53.0 %**	**11.5 %**
Topology unknown	GP MKL	27.4 %	**6.2 m**	21.7 deg	39.3 %	28.1 %
	Ours	**70.8 %**	6.6 m	**7.2 deg**	**48.1 %**	**16.6 %**
Oracle	Stereo	92.9 %	4.4 m	6.6 deg	62.7 %	8.0 %

Topology and Geometry Estimation: Errors / Accuracies with respect to GP-Baseline and Stereo

	Error
Felzenszwalb et al. (raw)	32.6 °
Felzenszwalb et al. (smoothed)	31.2 °
Our method (θ unknown)	**15.7 °**
Our method (θ known)	**13.7 °**

Orientation Estimation: Compared to state-of-the-art object detection we decrease angular orientation errors of moving objects by using context from our model. Measured in bird's eye view.