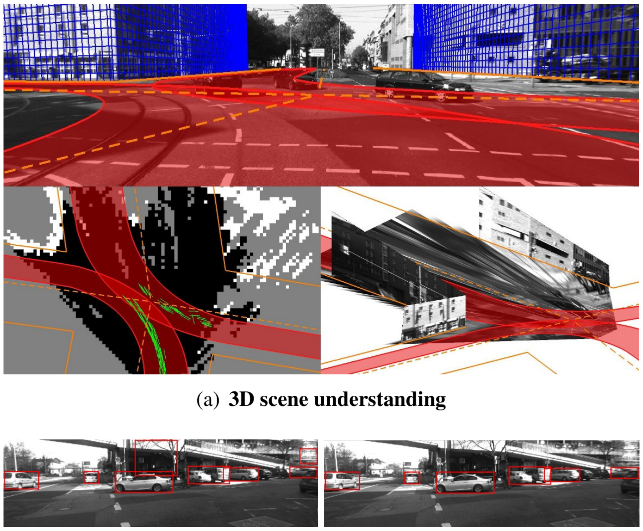
A Generative Model for 3D Urban Scene Understanding from Movable Platforms

Andreas Geiger and Martin Lauer
Department of Measurement and Control
Karlsruhe Institute of Technology
{geiger,martin.lauer}@kit.edu

Raquel Urtasun
Toyota Technological Institute at Chicago
rurtasun@ttic.edu

Abstract

3D scene understanding is key for the success of applications such as autonomous driving and robot navigation. However, existing approaches either produce a mild level of understanding, e.g., segmentation, object detection, or are not accurate enough for these applications, e.g., 3D pop-ups. In this paper we propose a principled generative model of 3D urban scenes that takes into account dependencies between static and dynamic features. We derive a reversible jump MCMC scheme that is able to infer the geometric (e.g., street orientation) and topological (e.g., number of intersecting streets) properties of the scene layout, as well as the semantic activities occurring in the scene, e.g., traffic situations at an intersection. Furthermore, we show that this global level of understanding provides the context necessary to disambiguate current state-of-the-art detectors. We demonstrate the effectiveness of our approach on a dataset composed of short stereo video sequences of 113 different scenes captured by a car driving around a mid-size city.



(a) **3D scene understanding**

(b) **Result of [3] (left) vs. using scene context from our model (right)**

Figure 1. **Inferring topology, geometry and semantics**. **Top:** Reprojection into the original image. The inferred streets are depicted in dash orange, buildings in blue, and activity patterns in red. **Middle:** Bird's eye perspective. The flow observations are depicted in green and the occupancy grid in (white, gray, black) for (occupied, unobserved, free) spaces. The right side depicts a pop-up of the scene from a different camera angle. **Bottom:** By using inferred context for hypothesis reweighting, our algorithm is able to reduce false positives of state-of-the-art detectors [3].

1. Introduction

In recent years much effort has been devoted to the problem of 3D scene understanding, as solving this task is key for applications such as autonomous driving, robot navigation or vision-guided mobile navigation systems. One of the crucial components for the success of such applications is to be able to robustly estimate the 3D layout of the scene from movable platforms. This is extremely challenging, particularly in scenarios such as dynamic urban scenes with a large degree of clutter arising for example when a car traverses a congested city. Despite the large body of work, existing approaches do not offer the level of accuracy required in such real-world applications.

Most of the existing work has been focused on detecting objects of interest [3, 5, 21, 1] or on creating segmentations of the scene into semantic labels (e.g., road, building) [22, 15, 2]. However, the level of actual 3D scene understanding provided by these methods is rather limited. In order to further push-forward this understanding, more geometric approaches have been developed. These approaches estimate rough 3D layouts (pop-ups) from monocular images [16, 9, 8].

Non-parametric models have been proposed to perform activity recognition from *static* platforms capturing road intersections from a bird's eye perspective [13, 20]. However, these models are not transferable to new scenes and thus only work for a fixed scene layout. To be able to apply these models to movable platforms, one would need to learn a different model for every type of scene. This is non-practical since it will require knowing a priori all scene types as well as the use of a very large training set.

In this paper we propose a principled generative model that is able to estimate the varying road topology as well as the 3D content of the scene within a single inference step. Our model employs non-parametric distributions, deals with varying topologies and can be learned from a few examples while avoiding overfitting. We derive a reversible jump MCMC scheme that is able to infer the geometric (e.g., street orientation) and topological (e.g., number of intersecting streets) properties of the scene layout, as well as the semantic activities occurring in the scene, e.g., traffic situations at an intersection.

We demonstrate the effectiveness of our approach in very challenging urban scenarios of real-world intersections with various topologies and a high degree of clutter. Our input sequences are captured from a mobile observer while traversing a mid-size city. We show that simple dynamic and static cues are sufficient for reliable estimation. In particular, we show that our approach can extract geometrical and topological scene knowledge, create pop-up representations as well as infer semantic activities such as the traffic situation at an intersection (see Fig. 1(a) for an illustration).

This global level of scene understanding is not only important for mobile navigation, but also for tasks such as object recognition since it provides the context necessary to disambiguate difficult detections. We demonstrate this in the context of state-of-the-art part-based detectors [3], where our model can be used to improve reliability (see Fig. 1(b) for an illustration).

In the following, we first discuss related work. Next, we introduce our generative scene model and show its performance in challenging urban scenarios. Finally, we discuss extensions of our model and conclude.

2. Related work

3D scene understanding is one of the main goals in computer vision and robotics. As such there exists a wide body of work in both domains. In robotics, 3D data (e.g., lidar, stereo) is widely used to create occupancy grids that represent the navigable space [19]. In combination with detailed maps and accurate GPS, first autonomous systems for urban driving have been presented during the DARPA Urban Challenge [10, 14]. In computer vision, efforts have mainly focused on generating efficient and accurate 3D maps (e.g., stereo algorithms) [17], creating segmentations of the scene into semantic labels (e.g., road, building) [2, 15, 22], detecting objects of interest [1, 3, 5, 21], estimating rough 3D from monocular images [8, 9, 16] and performing activity recognition of dynamic scenes from optical flow [13, 20].

In recent years the accuracy of object detection has increased considerably. One of the main reasons for this has been the PASCAL challenge, exposing object detection to increasing levels of difficulty. While there exist many generic object detectors [3], only a few approaches focus on urban scenes. Ess et al. [1] and Gavrila et al. [5] show impressive results on estimating 3D bounding boxes around pedestrians in cluttered scenes when using stereo. Wojek et al. [21] proposed a generative model to robustly track objects from a moving observer using a single camera. Similarly to us, their model relies on reversible jump MCMC for estimation.

Multiple approaches have investigated the creation of scene segmentations into semantic categories, e.g., road, sky. Wojek et al. [22] perform joint object detection and image segmentation in urban environments. Sturgess et al [15] combine appearance and structure-from-motion features for image segmentation of road scenes. However, the level of understanding generated by object detection [1, 5, 21] and image segmentation [2, 15, 22] without higher-level reasoning tells relatively little about the underlying scene structure.

More geometric approaches have been developed to push forward this understanding by learning how to estimate rough 3D (i.e., pop-ups) from monocular images [9, 16]. These techniques, however, use only weak constraints of global consistency, work only on relatively clean scenarios and can result in infeasible solutions. More recently, Gupta et al. [8] extend this idea to incorporate global consistency constraints such as those provided by laws of physics based on mass and volume of the blocks. They were able to automatically generate qualitative 3D image parses. Unfortunately, these qualitative parses are too simplistic and inaccurate to help robot navigation or autonomous driving. In contrast, in this paper we propose a generative model of dynamic 3D urban scenes which is able to produce accurate estimates of complex 3D topologies, even in the presence of large clutter and image variation.

Activity recognition of dynamical scenes is another focus of research which is related to our approach. Ess et al. [2] propose a segmentation-based approach to recognize traffic scenes in front of a moving vehicle. In particular, they estimate the presence of objects of interests, e.g., car, pedestrian, as well as type of road topology using classification of scene features. Unlike our approach, this only results in labels and does not provide any 3D information which is key for applications in navigation. The closest approach to ours is the works of Kuettel et al. [13] and Wang et al. [20], where non-parametric models were used to perform activity recognition from static observers capturing intersections from a bird's eye perspective. However, their models are not transferable to new scenes and require a static camera.

Unlike all of the aforementioned approaches, we are able to robustly estimate the 3D layout of the road, the location of the buildings, detect objects of interest as well as dynamic traffic activities in the scene.

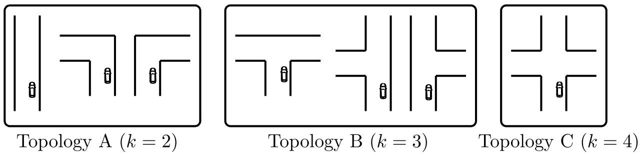


Figure 2. **Topologies.** Typical real world topologies.

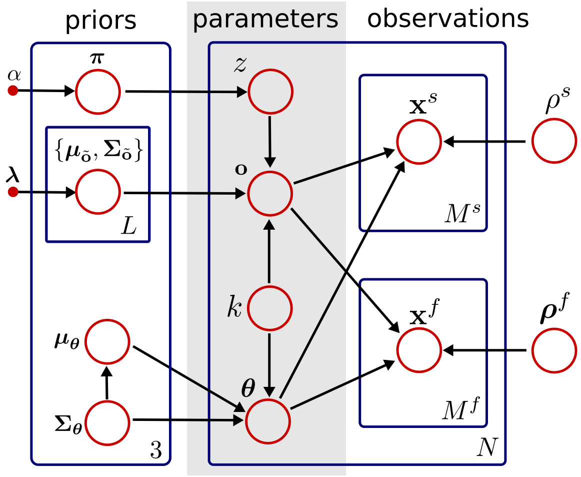
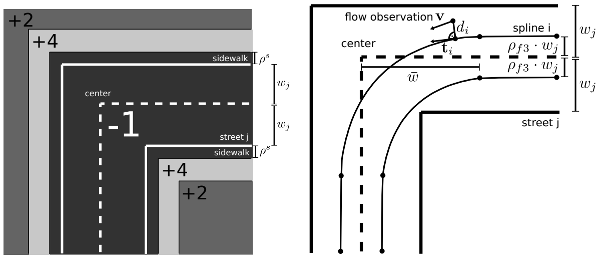


Figure 3. **Graphical model.** See section 3 for details.

3. Urban Scene Understanding

In this section we introduce our probabilistic generative model of 3D urban scenes. We consider a moving observer (e.g., robot, mobile platform, vehicle) navigating non-open spaces, such as road networks in cities or corridors/floors within buildings. Our goal is to infer from a small set of frames ($\approx$ 10 seconds) the observer's environment, which consists of the scene topology (see Fig. 2), geometry (e.g., orientation of the roads) as well as semantic information (e.g., traffic situation). Urban scenes are typically composed of a static environment as well as dynamic objects, e.g., cars driving on roads which are located between buildings. In this paper we propose a probabilistic generative model that captures the dependencies between both static and dynamic components of the scene.

We use 2D occupancy grids [19] computed from *ELAS*[1] disparity maps [6] as static features and 3D scene flow [12] as dynamical features. Using accurate non-linear visual odometry [12], we represent all dynamic and static features in the bird's eye perspective of the last frame's camera coordinate system. Our 2D occupancy grid is a voxelized 2D mapping of the environment where for each voxel a variable indicates if the voxel is occupied ($+1$), unobserved (0) or free (-1). For each scene, n, we represent the m-th voxel of the occupancy grid with a variable $x^s_{nm} \in \mathbb{R}$, the occupancy label for that voxel. Similarly, let $\mathbf{x}^f_{nm} \in \mathbb{R}^4$ be the m-th flow observation that is composed of its 2D location and velocity.



(a) Static observations
(b) Dynamic observations

Figure 4. **Observation model for static and dynamic features.**

For a given scene, the observations depend on the topology of the scene as well as its geometry. We parameterize these dependencies with an orientation variable $\mathbf{o} \in \mathbb{R}^k$ that contains the angles between neighboring streets present in the scene, as well as a variable $\boldsymbol{\theta} \in \mathbb{R}^{3+k}$ which contains the location $\mathbf{c} \in \mathbb{R}^2$, global rotation $r \in \mathbb{R}$ and street widths $\mathbf{w} \in \mathbb{R}^k$ of a road junction. To ensure positivity, we parameterize the street widths in the log domain, i.e., $\boldsymbol{\theta} = \{\mathbf{c}, r, \log \mathbf{w}\}$. Note that here we have dropped the dependency on the scene to simplify notation.

The joint distribution is defined by (compare Fig. 3)

$$p(\mathbf{X}^s, \mathbf{X}^f, \mathbf{O}, \boldsymbol{\Theta}, \mathbf{k}) = \prod_{n=1}^{N} p(k_n) p(\mathbf{o}_n|k_n) p(\boldsymbol{\theta}_n|k_n)$$

$$\times \prod_{n=1}^{N} \prod_{m=1}^{M^s_n} p(x^s_{nm}|\mathbf{o}_n, \boldsymbol{\theta}_n, \rho^s) \prod_{n=1}^{N} \prod_{m=1}^{M^f_n} p(\mathbf{x}^f_{nm}|\mathbf{o}_n, \boldsymbol{\theta}_n, \rho^f)$$

where N is the number of scenes, M^s_n is the number of static features in the n-th scene, M^f_n is the number of dynamical features in that scene and $\mathbf{X}^s = \{x^s_{nm}\}$, $\mathbf{X}^f = \{\mathbf{x}^f_{nm}\}$ are the sets of all static and dynamic features. $\mathbf{O}$, $\boldsymbol{\Theta}$ and $\mathbf{k}$ contain the orientations $\{\mathbf{o}_n\}$, the model parameters $\{\boldsymbol{\theta}_n\}$ and the number of streets for all scenes. Note that the model has different dimensionality for different topologies as the dimensionality of $\boldsymbol{\theta}_n$ and $\mathbf{o}_n$ depends on the number of streets.

We model the static observations as a Gibbs distribution

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

where f denotes the correlation between the occupancy grid and a model-dependent geometric prior that expresses our prior belief on the location of the free space, e.g., road, buildings alongside the road. The prior depends on the road width and orientation specified by the model parameters. An example of such prior is illustrated in Fig. 4(a). Here, ρ^s specifies the expected distance of buildings from the curbside and β is a design parameter to control the overall reliability of static observations.

The scene flow observations are modeled using a B-spline representation of the lanes, where for each pair of

[1] www.cvlibs.net

roads two splines are computed, one for each traffic direction, as illustrated in Fig. 4(b). The probability of a flow vector hereby depends on its distance to the B-spline as well as on how well its velocity vector aligns with the tangent vector of the respective B-spline. In particular, we model these dependencies as a hard mixture, where each observation is member of one lane (i.e., B-spline)

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

Here, d_i is the distance between an observed flow vector and the i-th spline, $\mathbf{t}_i$ denotes the tangent unit vector of the spline at the corresponding foot point, and $\mathbf{v}$ is the unit velocity vector of the flow observation. The observation model parameters $\boldsymbol{\rho}^f = (\rho_{f1}, \rho_{f2}, \rho_{f3})^T$ are the standard deviations of d_i, the velocity, as well as the distance from the spline to the street centerline, measured relative to the road width as illustrated in Fig. 4(b).

As we do not wish to prefer any topology a-priori, we assume a uniform prior on $p(k_n)$. We further model the prior probability on $\boldsymbol{\theta}$ with a multivariate Gaussian to capture the correlations between its components, hence imposing a log-normal distribution on the street widths. The relative orientation of the streets is modeled as a separate variable $\mathbf{o}_n$. In order to enforce a unique ordering of the streets, we define $\mathbf{o}_n$ on the $k_n - 1$ simplex Δ^{k_n-1} as an element-wise representation of the relative angular distance between two consecutive streets in counter-clockwise orientation. Hence $\sum_{i=1}^{k_n} o_{ni} = 1$ and $o_{ni} \geq 0$. The absolute orientation of the i-th street, α_{ni}, can be obtained from the global rotation r_n and the orientations of the streets up to the i-th one as

$$\alpha_{ni} = r_n + 2\pi \sum_{j=1}^{i-1} o_{nj}$$

Since we expect a multimodal distribution with unknown number of modes, we model $p(\mathbf{o}_n|k_n)$ with a non-parametric distribution. Unfortunately, employing an infinite mixture of Dirichlet distributions involves a conjugate prior of non-standard form. We use a standard trick in the statistics community to represent infinite mixtures of Dirichlet distributions with infinite mixtures of Gaussians by using a variable transformation based on the softmax function of a substitute variable $\tilde{\mathbf{o}}_n$, on which a Dirichlet process mixture with normally distributed observations and a conjugate Normal-Inverse-Wishart base measure can be easily employed. This restricts $\mathbf{o}_n$ to the $k_n - 1$ simplex. To remove the remaining degree of freedom, we fix $\exp(\tilde{o}_k)$ to 1, hence implying the unique bijection between $\mathbf{o}$ and $\tilde{\mathbf{o}}$. The full prior is given by

$$\begin{array}{lclclcl} k & \sim & \mathcal{U}(0,1) & & \boldsymbol{\theta} & \sim & \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \\ \boldsymbol{\pi} & \sim & \mathrm{GEM}(\alpha) & & z & \sim & \boldsymbol{\pi} \end{array}$$

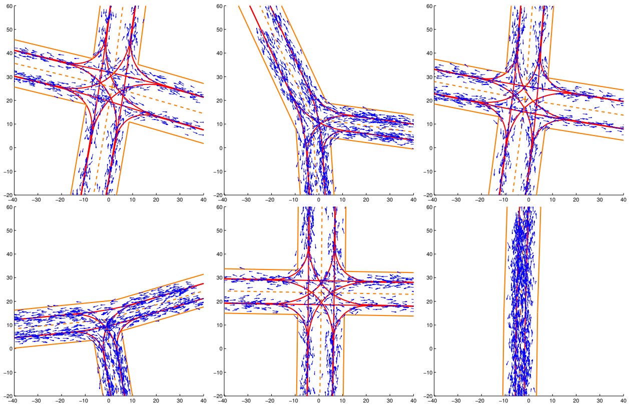


Figure 5. **Samples from the model.** For clarity of illustration, we do not show the dynamic and static observations that can also be sampled from our model.

$$\begin{array}{lcl} \{\boldsymbol{\mu}_{\tilde{\mathbf{o}}}^{(l)}, \boldsymbol{\Sigma}_{\tilde{\mathbf{o}}}^{(l)}\} & \sim & \mathcal{N}\text{-Inv-Wishart}(\boldsymbol{\lambda}) \\ \tilde{\mathbf{o}} & \sim & \mathcal{N}(\boldsymbol{\mu}_{\tilde{\mathbf{o}}}^{(z)}, \boldsymbol{\Sigma}_{\tilde{\mathbf{o}}}^{(z)}) \\ o_i & = & \frac{\exp(\tilde{o}_i)}{\sum_{j=1}^{k} \exp(\tilde{o}_j)} \end{array}$$

with base measure hyperparameters $\boldsymbol{\lambda} = (\kappa, \nu, \boldsymbol{\phi}, \boldsymbol{\Delta})$. Note that the parameters of $\mathbf{o}$ and $\boldsymbol{\theta}$ are duplicated three times, corresponding to the three types of topology we model. We now describe learning and inference in this model.

4. Learning and Inference

Our training set is composed of short sequences, where for each sequence n, k_n, $\mathbf{o}_n$ and $\boldsymbol{\theta}_n$ are labeled with the help of GoogleMaps images. During learning, k_n, $\mathbf{o}_n$, $\boldsymbol{\theta}_n$, x^s_{nm}, $\mathbf{x}^f_{nm}$ are observed, and $\boldsymbol{\pi}$, $\boldsymbol{\mu}_{\tilde{\mathbf{o}}}$, $\boldsymbol{\Sigma}_{\tilde{\mathbf{o}}}$, $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, ρ^s, $\boldsymbol{\rho}^f$, x^s_{nm}, $\mathbf{x}^f_{nm}$ are observed during inference. The rest of the parameters have to be inferred accordingly. In the following we show how learning and inference is performed.

Learning: Since k_n, $\mathbf{o}_n$ and $\boldsymbol{\theta}_n$ are observed, ρ^s, $\boldsymbol{\rho}^f$, the prior over $\mathbf{o}$ and the prior over $\boldsymbol{\theta}$ can be learned independently from the rest of the model. Assuming a non-informative prior over $(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, we obtain $(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ from $\{\boldsymbol{\theta}_n\}$ using maximum likelihood estimation. The prior over $\mathbf{o}$ is computed by applying the inverse softmax function to all $\mathbf{o}_n$ (yielding $\{\tilde{\mathbf{o}}_n\}$) and learning a distribution over a multimodal Gaussian mixture using a Rao-Blackwellized Gibbs sampler with broad prior distributions on the hyperparameters $\{\boldsymbol{\mu}_{\tilde{\mathbf{o}}}, \boldsymbol{\Sigma}_{\tilde{\mathbf{o}}}\}$. For more details, we refer the reader to [18]. During inference, we keep $\boldsymbol{\pi}$ and $\{\boldsymbol{\mu}_{\tilde{\mathbf{o}}}, \boldsymbol{\Sigma}_{\tilde{\mathbf{o}}}\}$ fixed to their MAP values. We further employ non-informative priors for ρ^s and $\boldsymbol{\rho}^f$. Since their posterior can not be computed analytically, we use a MH sampler with symmetric proposal distributions in order to obtain a representative set of samples. During inference, we fix both variables at their sample mean after the burn-in period.

Inference: We perform inference for each scene independently. Note that $\mathbf{x}_m^f$ and x_m^s are observed while the number of adjacent roads k, their orientation $\mathbf{o}$, width $\mathbf{w}$, the center of the crossroads $\mathbf{c}$ and the global orientation of the scene with respect to the ego position r are unknown. Since the priors over $\mathbf{o}$ and $\boldsymbol{\theta}$ are learned, we perform inference by sampling from the full posterior distribution $p(k, \boldsymbol{\theta}, \mathbf{o}|\{x_m^s\}, \{\mathbf{x}_m^f\})$[2]. Once samples are computed from the posterior distribution, we are able to compute expectations over k, $\boldsymbol{\theta}$ and $\mathbf{o}$.

Since, depending on k, the model has different number of parameters, we employ reversible jump MCMC (RJ-MCMC) [7] sampling for inference. We now briefly describe RJ-MCMC. We refer the reader to the supplementary material for a more detailed introduction. Green [7] proposed an extension of Metropolis-Hastings, named RJ-MCMC, that allows transdimensional jumps between models of different size. For simplicity, lets consider an example where we are interested in two different model topologies with states X_1 and X_2 respectively. Let p_1 and p_2 be the posterior distribution of interest, and let π_1, $\pi_2 = 1 - \pi_1$ be priors to control the amount of samples from each topology. Unfortunately a direct comparison of the densities is misleading since the probability measures on X_1 and X_2 might be different. To overcome this problem, reversible jumps introduce additional states U_1 and U_2 to define a bijection between the augmented state spaces $\tau : X_1 \times U_1 \rightarrow X_2 \times U_2$. Using a proposal distribution of the form $q_1(u_1|x_1)$, we can create a vector (x_1, u_1), transform it into (x_2, u_2) by applying the mapping τ and neglect u_2 to obtain a candidate for x_2. The acceptance probability of a jump from x_1 to x_2 becomes $\mathcal{A}(x_1, x_2) = \min\{1, \frac{\pi_2 \cdot p_2(x_2) \cdot q_2(u_2|x_2) \cdot |det(\mathcal{J}_\tau(x_1,u_1))|}{\pi_1 \cdot p_1(x_1) \cdot q_1(u_1|x_1)}\}$ where $det(\mathcal{J}_\tau(x_1, u_1))$ denotes the determinant of the Jacobian of τ. Analogously, we can switch from X_2 to X_1 using a proposal distribution $q_2(u_2|x_2)$ and compute the corresponding acceptance probability. Note that it is important to implement forward and backward steps together to meet the *detailed balance condition*, which ensures that the resulting sampler converges to the true underlying posterior distribution.

To switch between different sizes of the model we implemented moves that add and remove adjacent roads. Due to the complex structure of our model Gibbs-sampling is infeasible, hence we rely on Metropolis-Hastings moves. We combine *local moves* which vary a subset of the model given the last sample parameters and *global moves* which sample from the prior. Local moves are designed to sample from areas of high posterior probability to explore likely areas of the parameter space, while global moves are designed to jump between different areas of high probability to overcome the problem of poor initialization. All moves are selected randomly with the same probability. Our 14 transition kernels are shown in table 1.

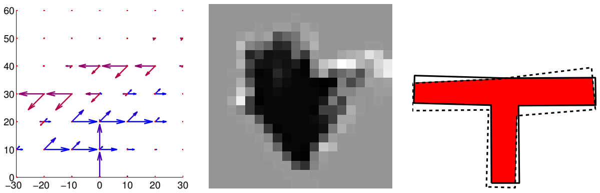
local Metropolis-Hastings moves
1. vary center of crossroads $\mathbf{c}$ slightly
2. vary overall orientation r slightly
3. vary width of all roads $\mathbf{w}$ slightly
4. select one road randomly and vary its width w_i
5. vary center of crossroads $\mathbf{c}$ slightly and adapt width of roads $\mathbf{w}$ to keep some curbsides unchanged
6. vary orientation of all roads $\mathbf{o}$ slightly
7. vary overall orientation r slightly and adapt $\mathbf{o}$ to keep the direction of the adjacent roads
8. select one adjacent road randomly and vary its direction slightly keeping the direction of the other adjacent roads
global Metropolis-Hastings moves
9. sample all parameters $\boldsymbol{\theta}$ and $\mathbf{o}$ from prior
10. sample center of crossroads $\mathbf{c}$ from prior
11. sample width of all roads $\mathbf{w}$ from prior
12. sample orientation of all roads $\mathbf{o}$ from prior
reversible jumps
13. add an adjacent road
14. remove an adjacent road

Table 1. **Survey of MCMC kernels for inference.**

The local moves 1-4 and 6 are design by sampling from a Gaussian with small variance centered around the current parameter value. The acceptance ratio of such a move is just the ratio between the posterior probability of the new parameter set compared to the current one. Since there exist interdependencies between the parameters, e.g., a change in the overall orientation of the scene r does not only vary the direction of the ego road but also of adjacent roads, we also sample several parameters together in moves 5-8. To increase the acceptance ratio when changing r (move 7), we correct $\mathbf{o}$ in order to keep the directions of the adjacent roads unchanged. We proceed analogously for moves 5 and 8.

The reversible jumps are designed as follows: To add an adjacent road to a k-armed crossroad we randomly select one of the angles defined in $\mathbf{o}$ to be split, say o_k. We then sample a split ratio u from a Beta distribution and the width of the new road v from $p(v|\mathbf{c}, r, \mathbf{w})$. The parameters of the Beta distribution are set so that its probability mass is concentrated in the $(0.3, 0.7)$ interval. The bijection τ maps $(\boldsymbol{\theta}, \mathbf{o}, u, v)$ onto $(\boldsymbol{\theta}', \mathbf{o}')$ by extending the orientation and the width vectors to include o'_{k+1} and w'_{k+1} respectively, and setting $o'_k = u \cdot o_k$, $o'_{k+1} = (1 - u) \cdot o_k$ and $w'_{k+1} = v$. Hence, the Jacobian of τ becomes $-o_k$. The inverse move randomly selects a road i to be removed, collapses the two orientation parameters o_i and o_{i+1}, and removes the width w_i from the width vector. Assuming that the add-move is selected with probability $s_{add,k}$ from the set of possible moves and the inverse move is selected with

[2] For notational simplicity we omit the dependence on the priors here



(a) Flow histogram (b) Static obstacle grid (c) Overlap measure

Figure 6. **Baseline features and overlap measure.** For the baseline we build fixed-size feature vectors from both, dynamic flow observations ($D = 392$) and static occupancy grids ($D = 400$). The overlap is measured by the intersection over the union score, as it is also employed in the PASCAL challenge.

probability $s_{remove,k+1}$ we obtain the acceptance probability for the add move as

$$\min\{1, \frac{s_{remove,k+1} \cdot p(\boldsymbol{\theta}', \mathbf{o}'|\{x^s_m\}, \{\mathbf{x}^f_m\}, k) \cdot o_k}{s_{add,k} \cdot p(\boldsymbol{\theta}, \mathbf{o}|\{x^s_m\}, \{\mathbf{x}^f_m\}, k+1) \cdot q(u, v|\boldsymbol{\theta})}\}$$

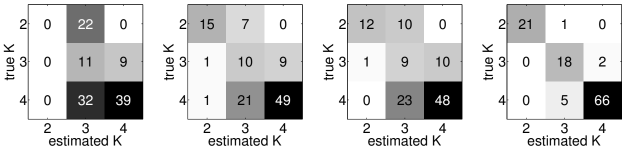
where $q(u, v|\boldsymbol{\theta})$ denotes the proposal probability for u and v described above, and we have used a uniform prior on k.

In practice it might happen that the RJ-MCMC sampler gets stuck in low probability areas of the posterior distribution if initialized poorly. To overcome this problem we adopt the idea of *simulated annealing* [11] and replace the posterior distribution p by an annealed probability $p^{\frac{1}{\alpha}}$ with $\alpha \geq 1$. In doing so, the acceptance ratio increases and the sampler is able to escape more easily. After burn-in we set $\alpha = 1$ to guarantee samples from the posterior of interest.

5. Experimental Evaluation

In this section we evaluate our approach with respect to extracting geometric and topological scene knowledge, improving object recognition by hypothesis re-weighting as well as inferring semantic activities, e.g., turn-out maneuvers. We built a database composed of 113 greyscale stereo sequences of length 10 seconds captured by randomly driving through a mid-size city. The database comprises 22 sequences of topology A, 20 sequences of topology B and 71 sequences of topology C (see Fig. 2). To obtain ground truth, we labeled the geometries and topologies by using GoogleMaps images, aligning them to the camera coordinate system of the last frame in each sequence using a high-accuracy ($<$ 20 cm) GPS+IMU system.

For all experiments we set the concentration parameter α to 2 in order to encourage a small number of components, and the hyperparameters $\kappa = 3$, $\nu = 5$, $\boldsymbol{\phi} = \mathbf{0}$ and $\boldsymbol{\Delta} = (0.01 \cdot \mathbf{I})$ in order to form a broad prior distribution. For the observation model we set $\beta = 300$, which accounts for the maximal number of 300 flow observations to which we limit our dynamic features.



(a) GP (flow) (b) GP (static) (c) GP (all) (d) Our method

	GP (flow)	GP (static)	GP (all)	Ours
Accuracy k	44.2 %	65.5 %	61.1 %	**92.9 %**
Location	6.1 m	5.6 m	5.4 m	**4.4 m**
Orientation	18.8 deg	11.5 deg	14.1 deg	**6.6 deg**
Overlap	42.2 %	51.7 %	49.3 %	**62.7 %**

(e) Parameter accuracy

Figure 7. **Inference of topology and geometry.** (a–d) Confusion matrices of the inferred number of streets k for the baselines and our method. (e) Location and orientation errors as well as the PASCAL score over the road segmentations.

Sampling: Since our model is generative, we can sample from it. As shown in Fig. 5, typical samples exhibit different topologies as well as geometries. Static and dynamic observations can also be sampled from our model.

Estimating geometry and topology: We compare our approach to Gaussian process (GP) regression. Since regression requires inputs of fixed dimensionality, we transform our dynamic and static features into vectors of fixed size (see Fig. 6). We first project both types of features into the camera coordinate system of the last frame. For the flow features we compute orientation histograms with 8 canoncial directions placed at 10 m distance in a 60 × 60 m grid in front of the observer, yielding a 392-dimensional feature vector. To overcome the problem of sparsity and quantization, we implemented a voting scheme, where each flow vector votes for its orientation in a small neighborhood. Similarly, we discretize the static features from the occupancy grid into 5 m × 5 m bins in a 100 m × 100 m grid, resulting in 400-dimensional features. In total, this gives a 792-dimensional feature vector which we normalize to range $[0, 1]$. We use an RBF with constant noise as kernel and learn the hyperparameters via maximum likelihood.

Fig. 7 (a)-(d) depicts the confusion matrix for estimating the topology. Here, $k = 2$ denotes scenes of type A, $k = 3$ for type B, and $k = 4$ for type C. We compare our approach to GP regression on single feature types and on a combination of static and dynamic features. Note that our method has an accuracy of 93 % while the best baseline, which relies only on static features, achieves only 65.5 % accuracy.

We further evaluated the error in predicting the location (intersection center) and orientation of individual streets. For the baselines, given the estimated k, we regress to the full set of parameters. Since an error in estimating k would badly affect the error in orientation, we assign each street to its closest ground truth neighbor and evaluated only assigned streets. Note that this measure clearly favors the baselines. We also evaluate the precision of the road estimation using the intersection over the union, i.e., PASCAL score. To this end, we compute this score on all roads taking into account a street length of three times the mean street width, as illustrated in Fig. 6 (c). Note that this is a segmentation task. As shown in Fig. 7 (e), our approach clearly outperforms the baselines, and excels particularly in estimating the street orientation.

Improving Object Recognition: The knowledge that our model extracts about the scene layout provides important contextual information that can be used to improve object detectors. For example, it can express which objects are more likely to be at a certain location, e.g., cars on the road. We evaluate this improvement on the state-of-the-art object detector [4] trained on car instances of the PASCAL VOC 2008 challenge. For evaluation, we hand-labeled for each sequence all car instances in the last frame. This results in 355 labeled car instances.

Our approach re-weights the scores returned by [4] by employing geometric knowledge about typical car locations. Towards this end, we first compute the training set mean and standard deviations of the object widths, heights and positions. In particular, we re-score the detections by adding the following term to the scores of [4]

$$\lambda \cdot \left[\max_i \ \exp\left(-\frac{d_i^2}{2\rho_{f1}^2}\right) + \sum_{i=1}^{3} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) \right]$$

Here d_i is the distance of a car detection to the i-th spline, ρ_{f1} is the lane parameter from our observation model, and $\{\mu_i, \sigma_i\}$ are mean and standard deviation of the object width, height and position, respectively. In our experiments, we set $\lambda = 0.5$, hence a value between 0 and 2 will be added to the detector score, which itself ranges $[-1, +1]$, depending on the size and location of the detected object.

Fig. 8 depicts precision-recall curves for the baseline [4] and our approach. As evidenced by the figure, our geometrical and topological constraints increase detection performance. For example, at a recall rate of 0.7, precision is increased from 0.7 to 0.9. The average precision is increased from 71.3 % to 74.9 %. This is also illustrated in Fig. 1(b), where in order to include the rightmost car into the detection result, the threshold of the baseline has to be lowered to a value which produces two false positives. In contrast, our re-scored ranking is able to handle this case.

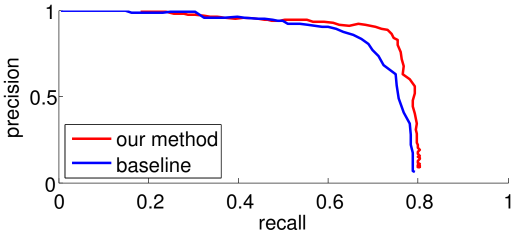


Figure 8. **Improving object recognition.** Re-scoring object detection hypotheses using geometric constraints established by our method yields better detection rates.

	GP (flow)	GP (static)	GP (all)	Ours
Hamming	0.18	0.23	0.16	**0.08**

Figure 9. **Activity recognition.** Normalized hamming distance between the estimated binary activity vector and the ground truth.

Extracting Semantic Activities: We now show our method's ability to infer higher-order knowledge, in particular traffic situations. We define an activity as a crossing action over an intersection. Given an intersection with k arms, there exist $k(k-1)$ types of crossing activities. Hence, we represent an activity as a binary $k(k-1)$ dimensional vector $\mathbf{a}$, where $a_i = 1$ denotes that the i-th type of crossing occurs in the sequence. Note that more than one coordinate can be active for a particular a. A natural distance measure for binary vectors is the normalized Hamming distance. It can be interpreted as the ratio of correctly determined crossing maneuvers. Again, we use GP regression as the baseline. For our method, we count the number of flow vectors which uniquely contribute to the respective spline and set $a_i = 1$ if this number exceeds 10 observations. The results are depicted by Fig. 9. Note that our method dramatically improves performance over the baselines. Inferred activities are highlighted in red in Fig. 10. Note that all flow observations (green) are nicely explained by the active lanes (red).

6. Conclusion and Future Work

We have proposed a generative model of 3D urban scenes that reasons about static and dynamic objects in the environment by capturing their dependencies. We have further derived a reversible jump MCMC scheme that is able to infer the geometric (e.g., street orientation) and topological (e.g., number of intersecting streets) properties of the scene layout as well as the semantic activities occurring in the scene e.g., traffic situations at an intersection. Furthermore, we have shown how our 3D reasoning can be used to improve the accuracy of current state-of-the-art object detectors. In the future, we plan to extend our generative model to jointly perform object detection, tracking (e.g., pedestrians and vehicles) and segmentation of the scene. We believe that such an extension will be able to greatly improve each of these tasks, particularly in the presence of small objects which require context-based interpretation.

References

[1] A. Ess, B. Leibe, K. Schindler, and L. V. Gool. Moving obstacle detection in highly dynamic scenes. In *ICRA*, 2009.

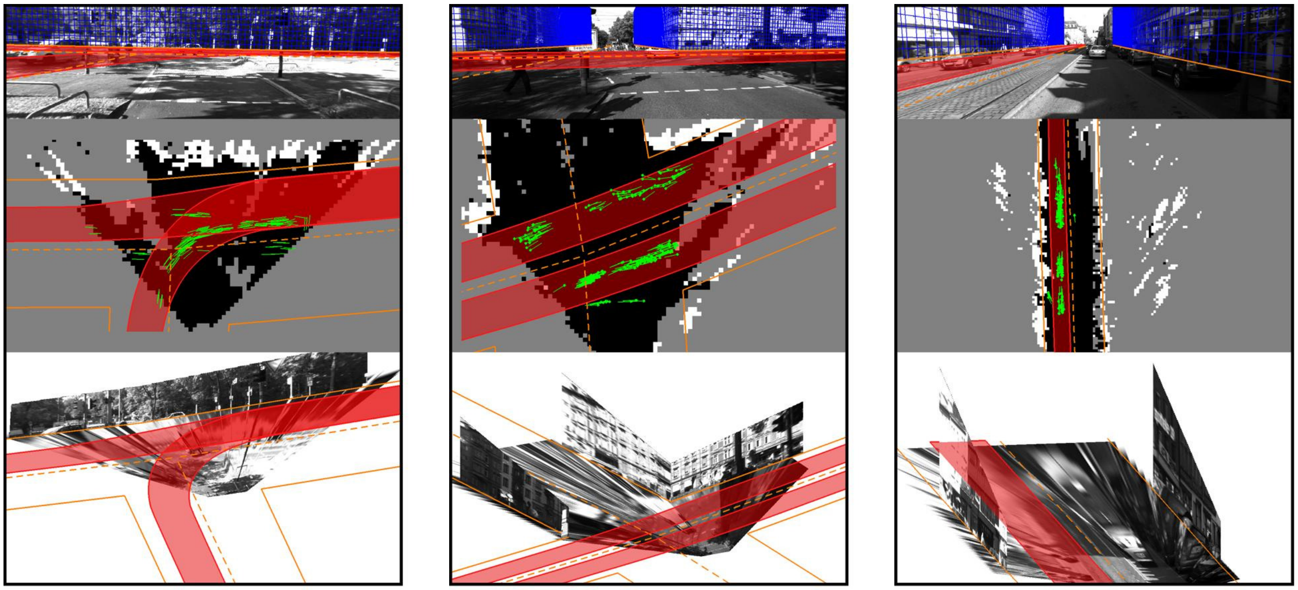


Figure 10. **Inference of topology and geometry.** (Top) Reprojection into the original image. The inferred streets are depicted in dash orange and the buildings in blue. (Middle) Bird's eye perspective. The flow observations are depicted in green and the occupancy grid in (white, gray, black) for (occupied, unobserved, free) spaces. (Bottom) 3D pop-ups of the scene from a different camera angle. The inferred active road segments are highlighted in red.

1945, 1946

[2] A. Ess, T. Mueller, H. Grabner, and L. van Gool. Segmentation-based urban traffic scene understanding. In *BMVC*, 2009. 1945, 1946

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010. 1945, 1946

[4] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 1951

[5] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 2007. 1945, 1946

[6] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *ACCV*, 2010. 1947

[7] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995. 1949

[8] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. *CVPR*, 2010. 1945, 1946

[9] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1), October 2007. 1945, 1946

[10] S. Kammel, J. Ziegler, and C. Stiller. Team annieway's autonomous system for the 2007 darpa urban challenge. *J. Field Robot.*, 25:615–639, September 2008. 1946

[11] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671680, 1983. 1950

[12] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In *IV*, 2010. 1947

[13] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. What's going on?: Discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010. in press. 1945, 1946

[14] M. Montemerlo, J. Becker, and S. Thrun. Junior: The stanford entry in the urban challenge. *J. Field Robot.*, 25:569–597, September 2008. 1946

[15] L. L. P. Sturgess, K. Alahari and P. Torr. Combining appearance and structure from motion features for road scene understanding. *BMVC*, 2009. 1945, 1946

[16] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009. 1945, 1946

[17] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. Journal of Computer Vision*, 47:7–42, 2002. 1946

[18] E. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, MIT, 2006. 1948

[19] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, September 2005. 1946, 1947

[20] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *PAMI*, 2009. 1945, 1946

[21] C. Wojek, S. Roth, K. Schindler, and B. Schiele. Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes. ECCV, 2010. 1945, 1946

[22] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, pages 733–747, 2008. 1945, 1946