Understanding High-Level Semantics by Modeling Traffic Patterns

Hongyi Zhang
Peking University
hongyiz@mit.edu

Andreas Geiger
MPI Tübingen
andreas.geiger@tue.mpg.de

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

Abstract

In this paper, we are interested in understanding the semantics of outdoor scenes in the context of autonomous driving. Towards this goal, we propose a generative model of 3D urban scenes which is able to reason not only about the geometry and objects present in the scene, but also about the high-level semantics in the form of traffic patterns. We found that a small number of patterns is sufficient to model the vast majority of traffic scenes and show how these patterns can be learned. As evidenced by our experiments, this high-level reasoning significantly improves the overall scene estimation as well as the vehicle-to-lane association when compared to state-of-the-art approaches [10].

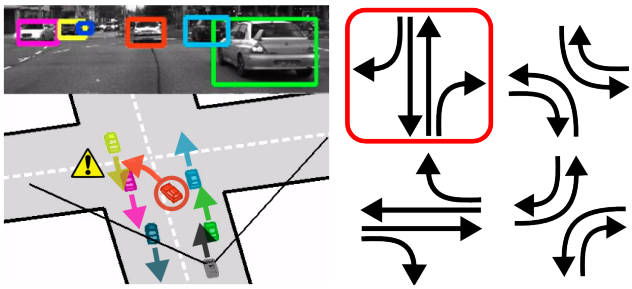


Figure 1. **Inference failure** when ignoring high-order dependencies: In [10] high-order dependencies between objects are ignored, leading to physically implausible inference results with colliding vehicles (left). We propose to explicitly account for traffic patterns (right, correct situation marked in red), thereby substantially improving scene layout and activity estimation results.

1. Introduction

Several decades after Roberts first attempts in 1965 [24], the problem of 3D scene understanding has witnessed great advances thanks to developments in object detection, semantic segmentation and image classification, amongst others. For indoor scenarios, we are able to robustly compute the layout of rooms [14, 23, 19, 30], and even achieve real-time performance [26]. Typical outdoor scenes are more complex as they often violate the Manhattan world assumption. First attempts in this setting focused on computing 3D pop-ups [15, 25, 12]. However, their geometry estimates are rather qualitative, not providing the level of understanding required for tasks such as mobile navigation.

In this paper, we are interested in understanding the semantics of outdoor scenes captured from a movable platform for the task of autonomous driving. We believe that even in the presence of cluttered real-world scenes, accurate knowledge can be inferred by exploiting strong contextual models. Existing approaches in this domain have focused mainly on semantic segmentation [3, 27], 3D object detection [22, 5], or very simple scene models [29]. Kuettel et al. [17] focus on surveillance scenarios. Unfortunately, such approaches are hard to transfer to autonomous systems due to the assumption of a static observer and the fact that the scene must be observed for a relatively long periord of time. A notable exception is the work of Geiger et al. [9, 10], who infer the 3D geometry of intersections as well as the location and orientation of objects in 3D from short video sequences. Unfortunately their approach does not capture high-order dependencies, and as a consequence, interactions between objects are not properly captured, leading to illegitimate traffic situations in the presence of ambiguous observations, as illustrated in Fig. 1 (left). As humans, however, we can easily infer the correct situation as we are aware of the existence of traffic signals and passing rules at signalized intersections, which can be summarized by typical traffic flow patterns, e.g., Fig. 1 (right). Moreover, we actively allocate our beliefs by weighing detection evidence against the strength of our prior knowledge.

In this paper, we take our inspiration from humans and propose a generative model of 3D urban scenes, which is able to reason not only about the geometry and objects present in the scene, but also about the high-level semantics in the form of traffic patterns. As shown in our experiments, not only does this help in associating objects with the correct lanes but also improves the overall scene estimation. We learn the traffic patterns from real scenarios and propose a novel object likelihood which, by integrating detection evidence over the full lane width, represents

lane surfaces very accurately. Our experiments reveal that a small number of traffic patterns is sufficient to cover the majority of traffic scenarios at signalized intersections. Taken together, the proposed model significantly outperforms the state-of-the-art [10] in estimating the geometry and tracklet associations. Furthermore, it provides high-level knowledge about the scene such as the current traffic light phase, without detecting and recognizing traffic lights explicitly, a task that is extremely difficult without the use of hand-annotated maps.

2. Related Work

A wide variety of approaches have been proposed to recover the 3D layout of **static** indoor scenes from a single image [13, 28, 19, 26]. These methods mainly build on edges and image segments as features, and make use of the 'Manhattan world' assumption. Several approaches have tried to explain the room clutter as 3D cuboids [14, 23, 19, 26, 30]. Recently, depth data has been used towards the goal of estimating support relationships between objects [20]. For static outdoor scenes, impressive results have been demonstrated in the case of pop-ups inferred from monocular imagery [15, 25]. Models incorporating physical relations between simple building blocks were introduced in [12] to produce physically realistic parsings of the scene. More recently, [11] investigated the problem of labeling occluded regions and [6] extracted additional constraints about the scene by recovering human poses from single images. Unfortunately, these approaches are mainly qualitative, do not model object dynamics and lack the level of accuracy necessary for real-world applications such as autonomous driving or robot navigation. In contrast, here we propose a method that is able to extract accurate geometric information by reasoning jointly about static and dynamic scene elements as well as their complex inter-play.

For a long time, **dynamic** objects have been mainly considered in isolation [7] or using simple motion models [16]. Only very recently, social interaction between individuals has been taken into account [18, 31, 1]. Choi et al. [1] introduce a hierarchy of activies, modeling the behaviour of groups. Using long-term scene observations, [17] proposes a method for unsupervised activity recognition and abnormality detection that is able to recover spatio-temporal dependencies and traffic light states from a static camera that is mounted on top of the roof of a building. While showing promising results, they neglect the interplay of objects with their environment and focus on surveillence scenarios with a fixed camera viewpoint. This limits their applicability as the learned scene models can not be transferred to new scenes, which is required when performing scene understanding from movable platforms. A notable exception is [21], which explicitly models collision avoidance. In contrast, here we are interested in inferring semantics at a higher level, such as multi-object traffic patterns at intersections, in order to improve the layout and object estimation process. Importantly, we do inference over intersections that we have never seen before. This is not possible with approaches such as [17].

Prior work on **3D traffic scene analysis** from movable platforms is mostly limited to ground plane estimation [2], classification [3, 8] or very simple planar scene models [29]. Only recently, Geiger et al. [9, 10] tackle the problem of urban traffic scene understanding by considering static infrastructure (e.g., buildings) and dynamic objects (e.g., vehicles). However, their limiting assumption of independent tracklets given the road layout can lead to implausible inference results such as vehicles on collision course, as illustrated in Fig. 1. In this work, we aim at alleviating these problems by incorporating a latent variable model that captures learned traffic patterns which jointly determine vehicle velocities and traffic light states. This not only enables us to infer high-level knowledge about the scene, but also improves tracklet-to-lane associations significantly, as shown in our experiments. Moreover, we show how a small set of plausible traffic patterns can be learned from annotated data. Additionally, we propose a novel object location likelihood, that, by marginalizing over the lateral position on the lane, models lanes much more accurately than [10] and improves the estimation of parameters such as the street orientations.

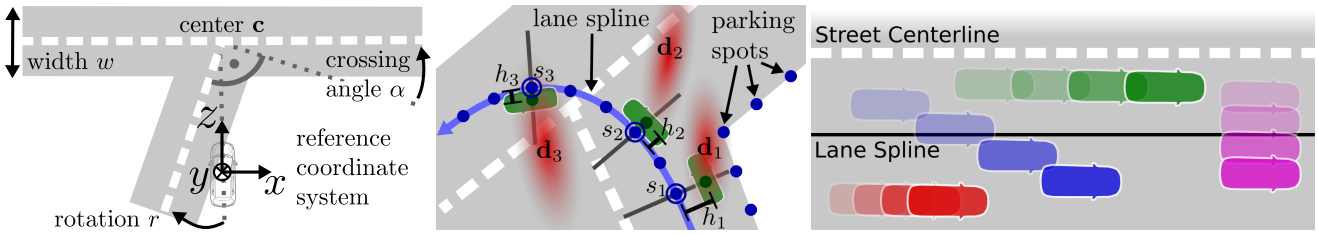
3. Modeling Traffic Patterns

In this section we present our generative model of scenes, which reasons jointly about high-level semantics in the form of traffic patterns as well as the 3D layout and objects present in the scene. We restrict our domain to 3-arm and 4-arm intersections, which are frequent intersections that exhibit interesting traffic patterns. We first learn a subset of predominant patterns from training data. At inference we recover these patterns from short video sequences and jointly associate vehicles to the corresponding lanes.

Following [10], we represent the geometry of the intersection in bird's eye coordinates and denote $\mathbf{c} \in \mathbb{R}^2$ the center of the intersection, r the rotation of our own car with respect to the intersection, w, the street width and α the crossing angle. We refer the reader to Fig. 2(a) for an illustration. Note that we assume the same width for all streets and force alternate arms to be collinear. This results in a very expressive model with only a few random variables. We utilize semantic segmentation, 3D tracklets and vanishing points as observations to estimate the traffic patterns, the layout and the vehicle-to-lane associations. We write our generative model as

$$p(\mathcal{E}, \mathcal{R}) = p(\mathcal{R})p(\mathbf{T}|\mathcal{R})p(\mathbf{V}|\mathcal{R})p(\mathbf{S}|\mathcal{R}) \quad (1)$$

where the image evidence $\mathcal{E} = \{\mathbf{T}, \mathbf{V}, \mathbf{S}\}$ comprises vehicle tracklets $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$, vanishing points $\mathbf{V} =$



(a) Geometry model (b) Tracklet model (c) Lane Model

Figure 2. **Geometry, Tracklet and Lane Models:** (a) road model parameters $\mathcal{R} = \{\mathbf{c}, r, w, \alpha\}$ which are infered by our approach, (b) tracklet with 3 detections and MAP estimates of the hidden variables s_i, h_i. Red: Uncertain object detections $\mathbf{d}_i$ in 3D. Green: True location of the vehicle along the normal of the lane spline at distance h_i. Blue: Lane spline and parking spots with associated s_i's (blue circles). In (c), despite less probable, the blue and purple tracklet get assigned a higher likelihood than the red and green ones in the model of [10]. Instead, our model correctly considers the red and the green tracklets more likely.

$\{v_f, v_c\}$, and semantic labels $\mathbf{S}$. We assume that the camera is calibrated and the camera parameters are known.

We employ simple priors to model the geometry $\mathcal{R}$, in the form of a Gaussian distribution for the global rotation r and the center of the intersection c, as well as a log-normal distribution for the width w. The latter is used to enforce a positive width w. We learn all distributions using maximum likelihood, and employ the same likelihoods as [10] for semantic segmentation and vanishing points, and develop novel algorithms to estimate the dynamic components of the scene, i.e., traffic patterns and car-to-lane associations.

3.1. A Traffic-Aware Tracklet Model

We aim at estimating the lane each vehicle is driving on, or in which road the vehicle is parked. We note that a "lane" in the paper corresponds to a driving direction. Towards this goal, drivable locations are represented with splines, which connect incoming and outgoing lanes of the intersection. Additionally, we allow cars to be parked on the side of the road, see Fig. 2 for an illustration. Let l be a latent variable indexing the lane or parking position associated with a tracklet. For a K-armed intersection, we have $l \in \{1, .., K(K-1)+2K\}$ possible states, with $K(K-1)$ the number of lanes and $2K$ the number of parking areas. Given a tracklet-to-lane association, we also model the (binary) stop-or-go state and longitudinal position of the tracklet, as well as its lateral position. Note that this is in contrast to [10] which assumes that the cars drive in the middle of the road with uniform prior over velocity and acceleration. This difference is very important in practice. As depicted by Fig. 2(c), the blue and purple tracklets will have higher likelihood than the green and red tracklets in the model of [10]. However, in practice those tracklets are very unlikely. Thus, we include a latent variable h that models the lateral location of a vehicle and integrate it with uniform prior probability over the lane width (i.e., lateral to the spline). This leads to higher likelihoods for the green and red tracklets, reflecting true vehicle dynamics more naturally. In addition, [10] does not model dependencies in vehicle dynamics, whereas we encode the intuition that vehicles switch between "stop" and "go" states rather rarely. We do this by learning a penalty on the state transition probability. This penalty helps to further reduce detection noise and to improve the tracklet estimation results.

In order to accurately estimate traffic patterns and lane associations, high-quality tracklet observations are crucial. We compute tracklets using a two-stage approach: In the first stage we form short contiguous tracklets by associating detections using the hungarian method while predicting bounding boxes over time using a Kalman filter. The cost matrix is computed from the normalized bounding box overlap (intersection over union) and the normalized cross-correlation score, where we additionally account for the detection uncertainty by taking the max over a small region. The second stage overcomes occlusions by joining tracklets which are up to 20 frames apart. Again, we make use of the hungerian method (but this time on tracklets) and compute the distance matrix from the cross-correlation score as well as the difference in the predicted bounding box and the true bounding box locations. Including the appearance term into our association led to much more stable tracklet associations, especially in the case of heavy occlusions. Finally, we follow [10] and project the bounding boxes into 3D using error propagation.

In the following, we use $s \in \{1, .., S\}$ to denote an object's anchor point at the spline curve (i.e., its longitudinal position), and $h \in \mathbb{R}$ to denote its true location along the normal direction of the spline (i.e., its lateral position), respectively. Together, they define a lane spline coordinate system such that every point on the 'y=0' plane in bird's eye perspective can be represented. In addition, $b \in \{\text{stop}, \text{go}\}$ is used to denote the binary stop-or-go status of a tracklet. To simplify notation, we will use g_i to represent the pair $g_i = (s_i, b_i)$. Subscripts will be used to refer to different frames/detections, e.g., s_i refers to the $i-$th frame/detection

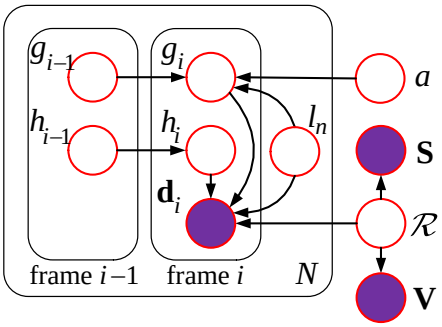


Figure 3. **Graphical model** showing two frames in plate notation.

of a tracklet. We define a 3D tracklet as a set of object detections $\mathbf{t} = \{\mathbf{d}_i, .., \mathbf{d}_M\}$, where each object detection $\mathbf{d}_i = (f_i, \mathbf{b}_i, \mathbf{o}_i)$ contains the frame index $f_i \in \mathbb{N}$, the object bounding box $\mathbf{b}_i \in \mathbb{R}^4$ defined as 2D position and size, as well as an orientation probability histogram $\mathbf{o}_i \in \mathbb{R}^8$ with 8 bins estimated by the detector [4].

In order to reason about traffic semantics, we introduce an additional latent variable a representing the possible traffic flow patterns. Fig. 4 illustrates the learned traffic patterns we use for 3-armed and 4-armed intersections. Details on the learning procedure can be found in Sec. 3.2. Additionally, we use an outlier pattern which defines a shared speed distribution for all cars on all lanes, yielding 5 states in total. We define the probability distribution over all tracklet observations as

$$p(\mathbf{T}|\mathcal{R}) = \sum_a p(a) \prod_n \sum_{l_n} p(l_n) p(\mathbf{t}_n|l_n, a, \mathcal{R}) \quad (2)$$

where a is the traffic pattern and l_n is the lane index of tracklet n, denoted as $\mathbf{t}_n$. We assume a uniform prior over traffic patterns a and lane assignments l. In the following we will define the likelihood of a single tracklet $p(\mathbf{t}|l, a, \mathcal{R})$, dropping the tracklet index n for clarity.

Our tracklet formulation combines a hidden Markov model (HMM) and a dynamical system with nonlinear constraints. Let $p(\mathbf{d}_i|s_i, h_i, a, l, \mathcal{R})$ be the emission probability of detection $\mathbf{d}_i$ at frame i. Let further $\mathbf{d}^{i-1} = \{\mathbf{d}_1, \ldots, \mathbf{d}_{i-1}\}$ denote the set of detections up to frame i. The tracklet likelihood $p(\mathbf{t}|a, l, \mathcal{R})$ is given as

$$p(\mathbf{t}|a, l, \mathcal{R}) = p(\mathbf{d}_1|a, l, \mathcal{R}) \prod_i p(\mathbf{d}_i|\mathbf{d}^{i-1}, a, l, \mathcal{R}) \quad (3)$$

For the sake of clarity let us drop the dependencies on $a, l, \mathcal{R}$ in the following. The marginal $p(\mathbf{d}_i|\mathbf{d}^{i-1}, a, l, \mathcal{R})$ is then given by

$$p(\mathbf{d}_i|\mathbf{d}^{i-1}) = \sum_{g_i} \int p(\mathbf{d}_i, g_i, h_i|\mathbf{d}^{i-1}) \mathrm{d}h_i \quad (4)$$

Algorithm 1: GENERATIVE PROCESS

(1) Sample the road geometry $\mathcal{R} \sim p(\mathcal{R})$
(2) Sample a traffic pattern $a \sim p(a)$
(3) **foreach** *tracklet* **do**
(4) Sample lane $l \sim p(l)$
(5) Sample the first hidden state $\{g_1, h_1\} \sim p(g_1, h_1|\mathcal{R})$
(6) Sample the first vehicle detection $d_1 \sim p(d_1|g_1, h_1, l, \mathcal{R})$
(7) **foreach** *frame* $i > 1$ **do**
(8) Sample the hidden state $\{g_i, h_i\} \sim p(g_i, h_i|g_{i-1}, h_{i-1}, a, l)$
(9) Sample the vehicle detection $d_i \sim p(d_i|g_i, h_i, l, \mathcal{R})$
(10) **return** $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$

with the joint probability of $\mathbf{d}_i, g_i$ and h_i defined as

$$p(\mathbf{d}_i, g_i, h_i|\mathbf{d}^{i-1}) = p(g_i|\mathbf{d}^{i-1}) p(h_i|\mathbf{d}^{i-1}) p(\mathbf{d}_i|g_i, h_i) \quad (5)$$

Here, $p(\mathbf{d}_i|g_i, h_i) = \mathcal{N}((s_i, h_i), (\xi\Lambda_{\mathbf{d}_i})^{-1}) \times p_{\text{heading}}^{\gamma}$ is modeled as a Gaussian distribution centered at (s_i, h_i) in the lane spline coordinate system, whose precision matrix $\Lambda_{\mathbf{d}_i}$ is obtained by error propagation of the bounding box into 3D, $\xi \in [0, 1]$ is a smoothing scalar to account for the possible correlation of errors within a series of detections and p_{heading} is a multinomial distribution obtained from the soft-max of the object orientation score corresponding to the tangent orientation at the spline point s_i. Note that ξ and γ can be interpreted as the feature weights in log probability. The predictive location probabilities are then

$$p(g_i|\mathbf{d}^{i-1}) = \sum_{g_{i-1}} p(g_{i-1}|\mathbf{d}^{i-1}) p(g_i|g_{i-1}) \quad (6)$$

$$p(h_i|\mathbf{d}^{i-1}) \propto \int_{-\frac{w_l}{2}}^{\frac{w_l}{2}} p(h_{i-1}|\mathbf{d}^{i-1}) p(h_i|h_{i-1}) \mathrm{d}h_{i-1} \quad (7)$$

where we obtain the location distributions by integrating h_i over the lane width w_l and marginalizing g_i

$$p(g_i|\mathbf{d}^i) \propto \int_{-\frac{w_l}{2}}^{\frac{w_l}{2}} p(\mathbf{d}_i, g_i, h_i|\mathbf{d}^{i-1}) \mathrm{d}h_i \quad (8)$$

$$p(h_i|\mathbf{d}^i) = \sum_{g_i} p(g_i|\mathbf{d}^i) p(h_i|g_i, \mathbf{d}^i) \quad (9)$$

with the posterior of h_i given g_i defined as

$$p(h_i|g_i, \mathbf{d}^i) \propto p(h_i|\mathbf{d}^{i-1}) p(\mathbf{d}_i|g_i, h_i) \quad (10)$$

Note that $p(g_i|\cdot)$ is modeled as a discrete distribution and $p(h_i|\cdot)$ follows a Gaussian distribution that is truncated at $\pm\frac{w_l}{2}$, with w_l the lane width. To keep our approach computationally tractable, we approximate the mixture of truncated Gaussian distributions arising in Eq. 9 with a sin-

#patterns	1	2	3	4	5	6
3-arm	64	74	80	**81**	81	81
4-arm	216	304	349	**366**	368	369

Table 1. **Learning Traffic Patterns:** Number of explained tracklets by the learned patterns for different maximum number of total patterns. Note that 4 patterns are sufficient to explain the majority of scenarios in our dataset.

gle truncated Gaussian distribution using moment matching. This can be done efficiently in closed form as described in the supplementary material.

The longitudinal transition probability is defined as

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

where $\pi(s_i - s_{i-1})$ represents a look-up table that depends on the difference $s_i - s_{i-1}$ in consecutive frames (i.e., driving speed).The lateral transition probability $p(h_i|h_{i-1})$ is a constant location model with additive Gaussian noise denoted as $\Delta\sigma_h^2$. As $p(h_i|\cdot)$ is represented via truncated Gaussian distributions, we approximate the posterior of h_{i-1} in the integrand of Eq. 7 by multiplying the non-truncated versions, truncating the result to $[-\frac{w_l}{2}, +\frac{w_l}{2}]$ and normalizing appropriately. For the parking spots, s is assumed to be constant and h is truncated at the end of the parking area.

Fig. 3 depicts our generative model. The generative process works as follows: First the road geometry and the traffic pattern are sampled. Next, we sample the hidden states h and g conditioned on the geometry in order to generate the first frame of the tracklet. This gives us the longitudinal position on the spline as well as the lateral distance to the spline. We then sample the vehicle detection conditioned on all $g_1, h_1, l, \mathcal{R}$. Finally, we generate the remaining observations of the tracklet by first sampling the hidden states using the dynamics and then sampling the vehicle detections conditioned on all other variables. This sampling process is summarized in Algorithm 1.

3.2. Learning

We restrict the set of possible traffic patterns to those that are collision-free. For 4-armed intersections we additionally require symmetry, yielding a total of 19 possible 3-arm and 11 possible 4-arm scenarios (see Fig. 4). Our goal is to recover a small subset of patterns which explains the annotated data well. Note that the number of possibilities is small enough to explore all pattern combinations exhaustively. Each combination is scored according to the total number of tracklets explained by the best pattern in the current set. A tracklet is explained by a pattern if its lane association and stop-or-go state agrees with the pattern. For each topology (3-arm or 4-arm), we successively increase the number of patterns. As illustrated in Table 1, 4 patterns are sufficient to explain the majority of scenarios. The learned patterns are illustrated in red in Fig. 4. We learn the transition probability of the binary hidden states b on active/inactive lanes respectively using the tracklets and corresponding ground truth tracklet-to-lane associations. Table 2 shows the learned state transition probabilities.

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
Uniform	0.450	0.050	0.050	0.450

Table 2. **Tracklet State Transition Probability:** S: stop states, G: go states. Tracklets on lanes of different states exhibit different "stop-go" transition statistics. However, note that all lane states have low switching probabilities.

3.3. Inference

In this section we describe how to infer the road parameters, the traffic patterns, the lane associations and the hidden states in our model. Since Eq. 1 cannot be computed in closed form, we approximate it using Metropolis-Hastings sampling. We accept a proposal $\mathcal{R}'$ given the current road parameters $\mathcal{R}$ with probability

$$\mathcal{A} = \min\left\{1, \frac{p(\mathcal{E}, \mathcal{R}')q(\mathcal{R}|\mathcal{R}')}{p(\mathcal{E}, \mathcal{R})q(\mathcal{R}'|\mathcal{R})}\right\} \tag{11}$$

where $p(\mathcal{E}, \mathcal{R})$ is given by Eq. 1 and $q(\mathcal{R}'|\mathcal{R})$ is the proposal distribution. The transition kernel is a combination of local moves, that modify $\mathcal{R}$ slightly using symmetric Gaussian mixture proposals, and global moves that sample $\mathcal{R}$ directly from $p(\mathcal{R})$. All move types are selected at random with equal probability. Given the traffic pattern a and road parameters $\mathcal{R}$, the lane association of tracklet $\mathbf{t}_n$ is given by the maximum of

$$p(l_n|a, \mathbf{t}_n, \mathcal{R}) \propto p(\mathbf{t}_n|a, l_n, \mathcal{R}). \tag{12}$$

where we assume a uniform prior $p(l_n)$ on the lanes. Similarly, we infer the maximum-a-posteriori traffic pattern for a particular sequence by taking the product over all tracklets $\mathbf{t}_n$ and marginalizing the lane associations l_n

$$p(a|\mathbf{T}, \mathcal{R}) \propto \prod_{n=1}^{N} \sum_{l_n} p(\mathbf{t}_n|a, l_n, \mathcal{R}). \tag{13}$$

Here, we assume a uniform prior on traffic patterns $p(a)$. Given the MAP estimate of the traffic pattern a and the lane association l_n, the MAP assignment of hidden states $\{g_1, \ldots, g_M\}$ is obtained by marginalizing out the hidden states $\{h_1, \ldots, h_M\}$, and running the Viterbi algorithm on the resulting HMM.

4. Experimental Evaluation

In this section, we show that our model can significantly improve the inference of tracklet-to-lane associations and

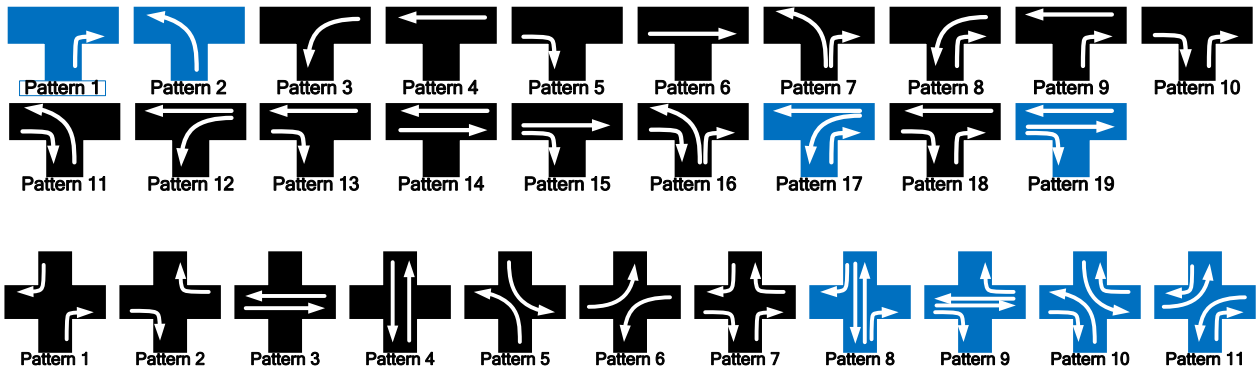


Figure 4. **Learning Traffic Patterns:** This figure shows all possible patterns with the learned ones in blue (top: 3-arm, bottom: 4-arm).

overall scene configuration over the state-of-the-art. For a fair comparison of tracklet-to-lane associations and road parameter inference, the method of [10] was run using our improved tracklets. Note that a comparison with [17] is not possible as their method requires a static observer and long term observations. In contrast, in our setting we have observations over short periods of time captured from a moving observer. Our dataset consists of all video sequences of signalized intersections from the dataset of [9], summing up to 11 three-armed and 49 four-armed intersection sequences. The last frame of each sequence is captured when the ego-car is entering the intersection as this is the point where an autonomous system would have to make a decision. Note that observing only parts of the intersection traversal renders the inference problem very challenging. As our primary goal is to reason about the vehicle dynamics and the traffic patterns we do not infer the topology. However, we note that the topology can be inferred analogously to [9, 10]. Throughout all of our experiments, we fix the hyperparameters to $\gamma = 0.05, \Delta\sigma_h^2 = 0.25$ and $\xi = 0.7$, unless otherwise stated. For $\pi(\cdot)$ we choose a uniform prior on the range $[2, 20]$m/s. We perform leave-one-out cross-validation and report the average performance.

Road Parameter Estimation: Following the error metric employed in [9, 10], we first evaluate the performance of our model in inferring the location (center of intersections), the orientation of its arms as well as the overlap of the inferred road area with ground truth. The results are shown in Table 3. Although our model uses the same features as [10], it gains extra benefits from a more realistic scene model.

Tracklet-to-Lane Associations: For each sequence we manually labeled all identifiable tracklet-to-lane associations, which serve as ground truth in our evaluation. Note that in contrast to the activities proposed in [9, 10] this measure is much more restrictive as it not only considers which lanes are given the green light, but instead requires each tracklet to be associated to the correct lane. This is a difficult task, especially given the fact that some tracklets are so short (in time or space) that they can only be disambiguated using high-level knowledge. As shown in Table 3 by modeling traffic patterns and object location and dynamics more accurately, we achieve a significant reduction in terms of tracklet-to-lane association (T-L) error wrt. [10].

Traffic Patterns: We labeled the traffic pattern for each of the sequences in our dataset, which is summarized in Table 4. In our dataset, 4 videos are dominated by pattern transitions and another 9 videos contain unidentifiable patterns which do not correspond to any of the patterns in our model. We evaluate the performance of our model on these videos with the exception of the traffic pattern inference task. As shown in Table 3 (right column), our model can infer traffic patterns with high accuracy while only having access to short monocular video sequences.

Qualitative Results: Fig. 5 depicts inference results of our model for some of the sequences. Note that the scenario shown in Fig. 1 is correctly inferred by the proposed model, illustrated at the bottom-left of Fig. 5. Due to space limitations we refer the reader to the supplementary material for full details of our inference results.

Sensitivity to Parameter Variations: We also analyze the sensitivity of our approach to three hyperparameters, i.e. the logarithm weight γ of the heading probability, the variance of the Gaussian kernel $\Delta\sigma_h^2$ in the transition probability of h, and a scaling constant ξ on the uncertainty of detections. As depicted in Fig. 6, the performance of our method does not suffer dramatically when moving away from the optimal setting, indicating the robustness of our approach.

Running Time: Our parallelized MATLAB implementation requires ~ 1.5 min to infer $\mathcal{R}$ when using 15000 samples to approximate $p(\mathcal{E}, \mathcal{R})$. Estimating the MAP vehicle locations given the road parameters only takes about 1 second for all tracklets of a sequence.

Method	T-L error (all)		T-L error (>10m)		Location		Orientation		Overlap		Pattern error	
	3-arm	4-arm	3-arm	4-arm	3-arm	4-arm	3-arm	4-arm	3-arm	4-arm	3-arm	4-arm
[10]	46.7%	49.9%	17.9%	30.1%	**4.3 m**	5.4 m	3.3 deg	8.0 deg	58.7%	56.0%	–	–
Ours	**15.2%**	**30.1%**	**3.6%**	**14.0%**	5.7 m	**4.9 m**	**2.4 deg**	**4.3 deg**	**61.5%**	**61.3%**	**18.2%**	**19.4%**

Table 3. **Geometry Estimation and Tracklet-to-Lane Association Results:** Results of tracklet-to-lane association (T-L) error, intersection location errors (bird's eye view), street orientation errors and street area overlap (see [10] for a definition).

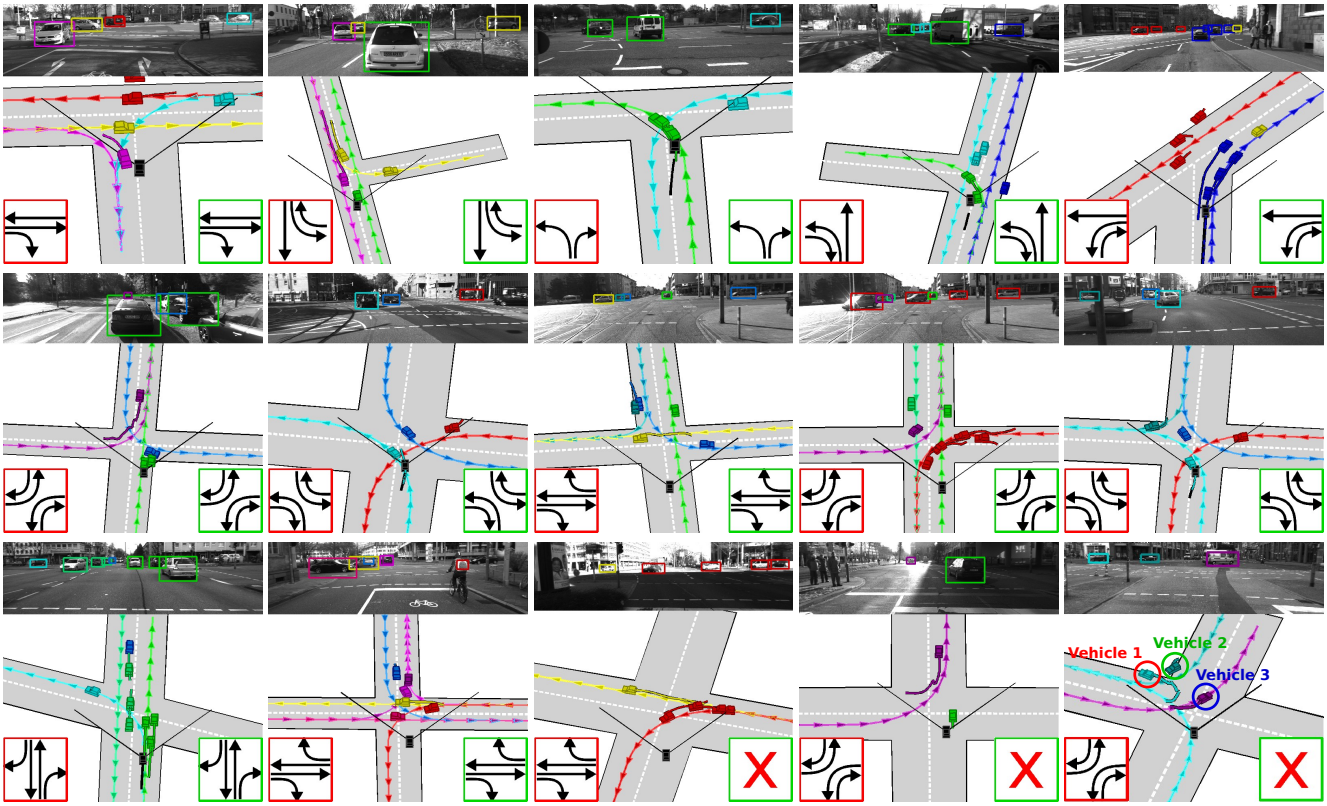


Figure 5. **Qualitative Results:** We show tracklets using the colors of the lanes they have been assigned to by our method. For clarity, only lanes with at least one visible tracklet are shown. The pictogram in the lower-left corner (red) of each image shows the inferred traffic pattern, the symbol in the lower-right corner (green) the ground truth pattern. An 'X' denotes ambiguities in the sequence. **First row:** Correctly inferred 3-arm intersection scenarios. **Second row:** Correctly inferred 4-arm intersection scenarios. **Last row:** The first figure shows successful inference for the sequence from Fig. 1. The second figure displays a backpack that has been wrongly detected as a car. The other cases are ambiguous ones as they contain transitions between two adjacent patterns or irregular driving behaviors such as U-turns (rightmost figure). Additional results on all sequences from our dataset and a case study can be found in the supplementary material.

	#Pat1	#Pat2	#Pat3	#Pat4	#Outliers
3-armed	0	1	6	4	0
4-armed	13	14	6	3	13

Table 4. **Traffic Patterns:** Number of occurrences of each traffic pattern (see Fig. 4) in the data.

5. Conclusions

In this paper, we proposed a generative model of 3D urban scenes which is able to reason jointly about the geometry and objects present in the scene, as well as the high-level semantics in form of traffic patterns. As shown by our experiments, this results in significant improvements in terms of performance over the state-of-the-art in all aspects of the scene estimation and allows to infer the current traffic light situation. In the future, we plan to extend our approach to model transitions between traffic patterns. Another interesting avenue for future work is to incorporate map information as weak prior knowledge. Even though maps are inaccurate and possibly outdated, they might still provide useful cues in the context of robust scene layout inference.

References

[1] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, 2012. 2

[2] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. Robust multi-person tracking from a mobile platform. *PAMI*,

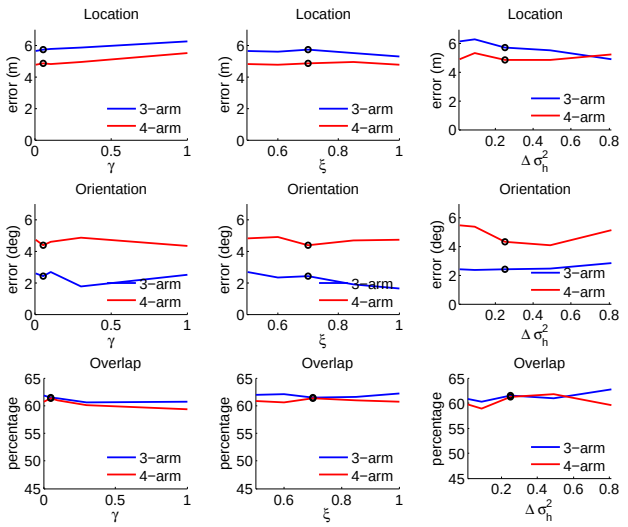


Figure 6. **Robustness against Parameter Variations:** We depict the robustness of our method against varying three parameters: the logarithm weight γ of the heading probability, the variance of the Gaussian kernel $\Delta\sigma_h^2$ in the transition probability of h, and the scaling constant ξ on the uncertainty of detections. The black dot marks the parameter setting used in all of our experiments, as reported in Table 3.

31:1831–1846, 2009. 2

[3] A. Ess, T. Mueller, H. Grabner, and L. van Gool. Segmentation-based urban traffic scene understanding. In *BMVC*, 2009. 1, 2

[4] P. Felzenszwalb, R.Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32:1627–1645, 2010. 4

[5] S. Fidler, S. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *NIPS*, December 2012. 1

[6] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single-view geometry. In *ECCV*, 2012. 2

[7] D. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73:41–59, 2007. 2

[8] A. Geiger and B. Kitt. Objectflow: A descriptor for classifying traffic motion. In *IEEE Intelligent Vehicles Symposium*, San Diego, USA, June 2010. 2

[9] A. Geiger, M. Lauer, and R. Urtasun. A generative model for 3d urban scene understanding from movable platforms. In *CVPR*, 2011. 1, 2, 6

[10] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, Granada, Spain, December 2011. 1, 2, 3, 6, 7

[11] R. Guo and D. Hoiem. Beyond the line of sight: Labeling the underlying surfaces. In *ECCV*, 2012. 2

[12] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010. 1, 2

[13] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009. 2

[14] V. Hedau, D. Hoiem, and D. A. Forsyth. Recovering free space of indoor scenes from a single image. In *CVPR*, 2012. 1, 2

[15] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75:151–172, 2007. 1, 2

[16] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008. 2

[17] 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. 1, 2, 6

[18] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: modeling social and grouping behavior on a linear programming multiple people tracker. *1st Workshop on Modeling, Simulation and Visual Analysis of Large Crowds*, 2011. 2

[19] D. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010. 1, 2

[20] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[21] S. Pellegrini, A. Ess, K. Schindler, and L. J. V. Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 2

[22] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3d2pm - 3d deformable part models. In *ECCV*, Firenze, Italy, 2012. 1

[23] L. D. Pero, J. Bowdish, D. Fried, B. Kermgard, E. Hartley, and K. Barnard. Bayesian geometric modeling of indoor scenes. In *CVPR*, 2012. 1, 2

[24] L. G. Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology. Dept. of Electrical Engineering, 1963. 1

[25] A. Saxena, S. H. Chung, and A. Y. Ng. 3-D depth reconstruction from a single still image. *IJCV*, 76:53–69, 2008. 1, 2

[26] A. Schwing and R. Urtasun. Efficient Exact Inference for 3D Indoor Scene Understanding. In *ECCV*, 2012. 1, 2

[27] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009. 1

[28] H. Wang, S. Gould, and D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. In *ECCV*, 2010. 2

[29] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular visual scene understanding: Understanding multi-object traffic scenes. *PAMI*, 2012. 1, 2

[30] J. Xiao, B. C. Russell, and A. Torralba. Localizing 3d cuboids in single-view images. In *Advances in Neural Information Processing Systems*, December 2012. 1, 2

[31] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *CVPR*, 2011. 2