Reading between the Lanes: Road Layout Reconstruction from Partially Segmented Scenes

Lars Kunze, Tom Bruls, Tarlan Suleymanov, and Paul Newman

***Abstract*— Autonomous vehicles require an accurate and adequate representation of their environment for decision making and planning in real-world driving scenarios. While deep learning methods have come a long way providing accurate semantic segmentation of scenes, they are still limited to pixel-wise outputs and do not naturally support high-level reasoning and planning methods that are required for complex road manoeuvres. In contrast, we introduce a hierarchical, graph-based representation, called *scene graph*, which is reconstructed from a partial, pixel-wise segmentation of an image, and which can be linked to domain knowledge and AI reasoning techniques.**

In this work, we use an adapted version of the Earley parser and a learnt probabilistic grammar to generate scene graphs from a set of segmented entities. Scene graphs model the structure of the road using an abstract, logical representation which allows us to link them with background knowledge. As a proof-of-concept we demonstrate how parts of a parsed scene can be inferred and classified beyond labelled examples by using domain knowledge specified in the Highway Code. By generating an interpretable representation of road scenes and linking it to background knowledge, we believe that this approach provides a vital step towards explainable and auditable models for planning and decision making in the context of autonomous driving.

I. Introduction

Autonomous vehicles need to perceive their surroundings accurately for safe decision making and navigation in complex urban environments. These highly-structured environments can be described by hierarchical graphs containing semantic and spatial constraints. Such graphical representations can be employed for (cost-based) planning, inferring object classes, or reasoning about missing or occluded parts. More importantly, they provide a way to explain the behaviour and decision making of the vehicle which is paramount for real-world deployment and adoption. In this paper, we introduce such a representation, which is generated from partially segmented scenes and allows us to reason about the environment.

Recently, deep semantic segmentation networks have achieved impressive results for pixel-wise scene understanding of images [1], [2]. However, these methods suffer from interpretation and debugging difficulties and often fail to include prior information or dependencies/constraints (in the output space). More importantly, they do not naturally support high-level reasoning which is required for planning and navigation.

Authors are from the Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK. {lars, tombruls, tarlan, pnewman}@robots.ox.ac.uk

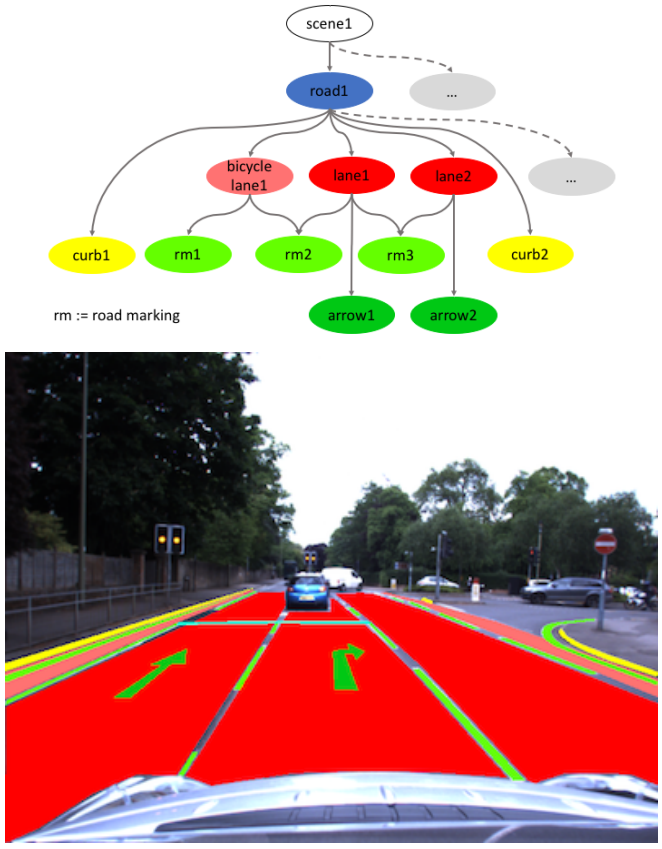


Fig. 1. Hierarchical scene graph representation (top) that was reconstructed from a partially segmented image (bottom). In this work we present a probabilistic scene parser that reconstructs the layout of road scenes from partial segmentations of road markings and curbs.

In contrast, all important (static or dynamic) objects influencing the decision making are detected separately in the mediated approach [3], [4]. This produces a world representation which can be employed more directly for planning and navigation. Interestingly, most approaches focus on detecting a single type of object or perform detection of several types of objects independently. Thereby they neglect that urban traffic scenes are highly structured and that there exist spatial and semantic constraints between objects, since these scenes are built and function according to specified rules.

Therefore, we introduce *scene graphs*, a hierarchical, graph-based representation, to model road layouts (i.e. lane geometries). Fig. 1 shows an example scene graph for a segmented road scene. We focus on the reconstruction of scene graphs from partial, pixel-wise segmentation. In par-

ticular, we consider segmented entities of road markings and curbs to reconstruct the semantic structure of road scenes. The road layout is reconstructed from these entities using both a learnt probabilistic context-free grammar and a learnt spatial, relational model. A road layout is chosen from a set of competing hypotheses by estimating the maximum a posteriori probability (MAP) of each model. Furthermore, we show that scene graphs can be refined by linking them with domain knowledge about the road construction, e.g. from the Highway Code.

In this paper, we make the following contributions:

- we introduce *scene graph*, a formal logic-based description of road scenes using a graph-based representation;
- we present an approach based on dynamic programming for parsing road scenes and reconstructing scene graphs from partial, segmentations and a learnt probabilistic grammar; and
- we demonstrate how scene graphs can be further refined and used for reasoning when linked to domain knowledge.

The remainder of the paper is structured as follows. We first discuss related work in Sec. II. In Sec. III, we provide and overview of the approach and explain how scene graphs are generated from both object segmentations and learnt prior models. In Sec. IV, we explain how scenes are partially segmented using deep networks for road markings and curbs. In Sec. V we explain how we represent a scene, learn both a probabilistic context-free grammar and a spatial relational model to describe scenes, and how scenes are parsed and interpreted using an adapted version of the probabilistic Earley parser. In Sec. VI, we showcase and discuss several examples of scene graphs and explain how they can be further refined. Lastly, we discuss possible application in Sec. VII before we conclude in Sec. VIII.

II. Related Work

In this section, we review different approaches for scene understanding in the context of autonomous vehicles. We mainly focus on graph based methods, since these are closest to the scene graph.

1) Graph-based Approaches: Representing the contents of scenes using graph-based approaches is not novel. In the context of urban traffic scenes, however, there exist only a few papers that take the spatial and semantic constraints into account by introducing graphs.

In [5], different sensor modalities and hierarchical graphs containing relational knowledge are fused to model traffic scenes. The output is still a pixel-wise segmented image not directly employable for automated driving.

Several other papers implement more high-level reasoning to infer the lane geometries. The authors of [6] introduce a theoretical, hierarchical framework including uncertainties to reason about multiple hypotheses for the lane geometry. Similar methods that work on real-world data are introduced in [7], [8]. From linear patches of lane markings a graph is built including their spatial relationship represented by continuous distributions and non-parametric belief propagation is used to infer the different lanes in the scene. However, these methods are not guaranteed to work in urban environments.

In [9], the lane separators are modelled as latent variables without linear constraints so that the framework becomes applicable to more complex scenes. By encoding geometric relationships at different levels (i.e. lane markings, lane separators, lanes, and road), the authors show that they improve inference of the lane geometries even in case of many false detections at the root nodes. This work is similar to our approach as we also represent the geometric relationships of different entities according to the hierarchy.

The driving rules of a traffic scene are given by the type of road markings that often appear in similar configurations. Therefore, [10] connects them as a graph and optimises a CRF with handcrafted spatial features of the road markings to predict their class. Similarly, we learn a distribution of geometric and relation features to predict and evaluate the type and the role of an entity within the hierarchy.

Work by [11] is most similar to our approach. In their work, they learn a probabilistic grammar based on a set of features and use a dynamic programming approach to generate a scene graphs which describe the furniture layout of synthetic indoor scenes. Whereas their approach considers full object knowledge from CAD models, our approach reconstructs scenes from partial observations of real-world environments.

2) Mediated Approaches: Proposed solutions differ widely in terms of the objects that are taken into account, used sensors, required computation time, usage of prior information, and abstraction level of the output. In general, our approach is flexible to consider different kinds of information from various resources. In this work, we consider segments of road markings and curbs as input.

In [12] a coarse road geometry/scene analysis is estimated from the acquired semantic segmentation. This framework is significantly extended in [3] where the precise intersection geometry is inferred from vanishing points, semantic labels, and tracklets of traffic participants. However, these methods cannot be used for navigation directly as they do not map to precise lane geometries and do not include the road rules. The former is solved in [13] by looking more closely into the tracks of the surrounding vehicles. Our also approach models the geometry of high-level concepts based on the low-level image segmentations. Thereby, information about lanes and boundaries can potentially be used for navigation planning. Through advancements in deep learning we have now come to a point where even reasoning of the space behind occluded parts of the images is possible for inferring road geometries [14]. In future work, we also plan to extent our work in this direction.

3) Deep Networks: All of above mentioned methods require handcrafted features/probabilities in some way to optimise the graph. It has been shown by now that deep networks with learned feature maps achieve much better semantic (instance) segmentation [1], [2] and thus understanding of the scene. Besides, they are able to generalise better when auxiliary output tasks are employed [15]. How-

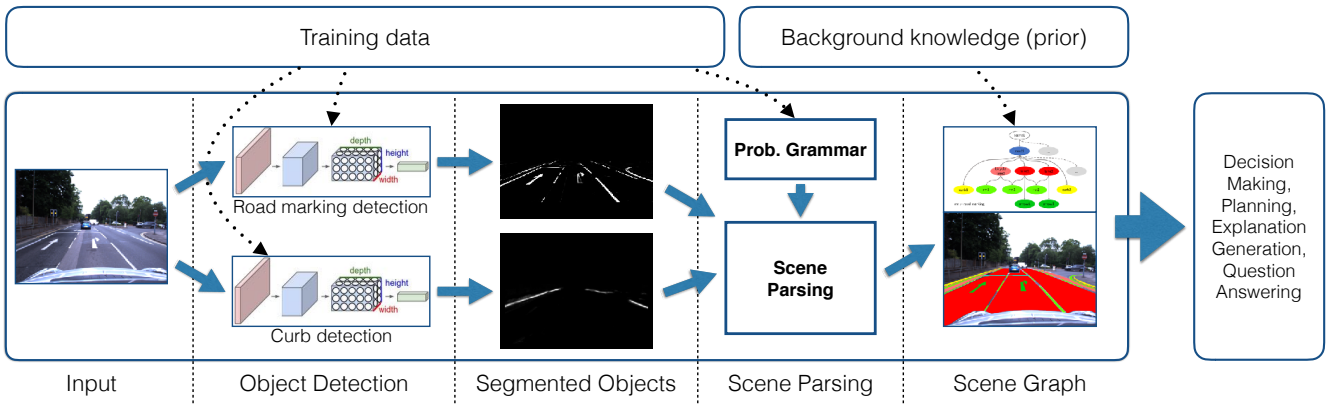


Fig. 2. Scene parsing approach based on road marking and curb detections. The approach has two main steps: (1) given an image, road marking and curb segments are detected by deep networks, and (2) given a set of detected segments, the scene is parsed using an adapted version of the Earley algorithm and a learnt probabilistic grammar. The resulting scene graph is integrated with domain knowledge and can used for planning and decision making.

ever, these networks suffer from interpretation and debugging difficulties and often fail to include prior information, high-level reasoning, or constraints (in the output space). Recently, some works have tried to improve some of these disadvantages by introducing spatial and semantic reasoning frameworks that can be trained in an end-to-end way [16]–[18]. In this work, we simply use deep networks as an effective way for segmenting an input image. However, our future goal is to extend this approach and to feed geometric, spatial, and semantic constraints back to the deep networks during learning.

III. Approach Overview

Our approach constructs a symbolic, graph-based description of the road layout given an image of a road scene (see Fig. 1). When interpreting the image, our approach considers two types of information: object detections and common road configurations based on learnt prior models.

Fig. 2 depicts the overall pipeline of our approach. We first segment the image by detecting curbs and road markings using trained deep networks (Sec. IV). These pixel-wise segmented images are clustered and the resulting entities are considered as input for a parsing process which generates a hierarchical scene representation (*scene graph*) (Sec. V). The parser takes object detections (and their uncertainty) and prior information of road scenes into account. Our probabilistic approach is in particular suitable for integrating incomplete and uncertain information from object detection pipelines. Each valid parse tree is scored by a probability which allows us to disambiguate between alternative representations. Intuitively, the score captures three aspects: (1) hierarchy (2) geometric features of detected entities, and (3) spatial relations between entities in the hierarchy. As we represent scene graphs using logical representations they can be linked to background knowledge and used for auditable planning and decision making.

IV. Scene Perception

This section describes how road markings and curbs are detected in a given image of a road scene. The resulting pixel-based images are clustered and segmented entities are obtained which are considered as input for the scene interpretation process described in Sec. V.

A. Road Marking Detection

Road markings are a critical component for (autonomous) driving especially in urban environments. The road rules are captured by their underlying meaning and they guide all traffic participants through potentially dangerous situations. Therefore, real-time detection and interpretation of road markings is an important cue for high-level scene understanding and aids planning and decision making.

Detecting all painted road markings (not just lane separators) on the road surface, which dictate the traffic rules for that particular urban setting, is a challenging problem for several reasons. Firstly, there are visual challenges such as occlusions, varying lighting, and changing weather conditions. Secondly, road markings vary from country to country and are often degraded. Lastly, there are no large datasets available for training with accurate ground-truth labels for road markings.

Road marking detection in images can be seen as a semantic segmentation problem. State-of-the-art methods for these tasks implement deep networks, which are able to learn specific scene context and thereby cope with the challenges stated above, as long as sufficient training data is available. Manually generating training data is extremely labour expensive, because of the required pixel-level detail in combination with the aforementioned visual issues. Therefore, we create road marking annotations in a weakly-supervised way, by leveraging complementary sensor modalities (i.e. LiDAR).

For generating the annotations, we exploit the property that road markings are highly reflective and must lie on the road surface. Firstly, we utilise the LiDAR point cloud to coarsely segment the road surface from the image. A dense CRF is then optimised to identify the road marking image pixels by

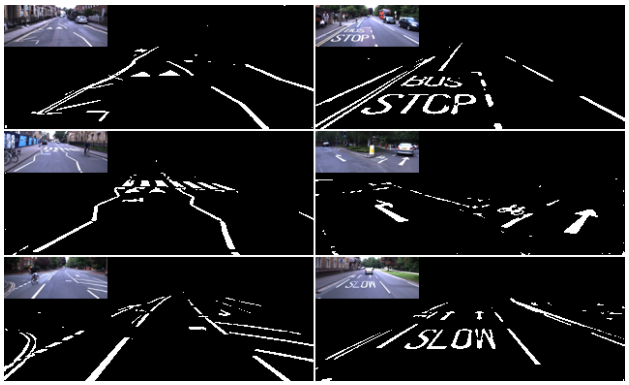


Fig. 3. Road marking detection performed by a deep semantic segmentation network in real-time. Before the detections are employed to generate the scene graph, they are mapped to top-down view.

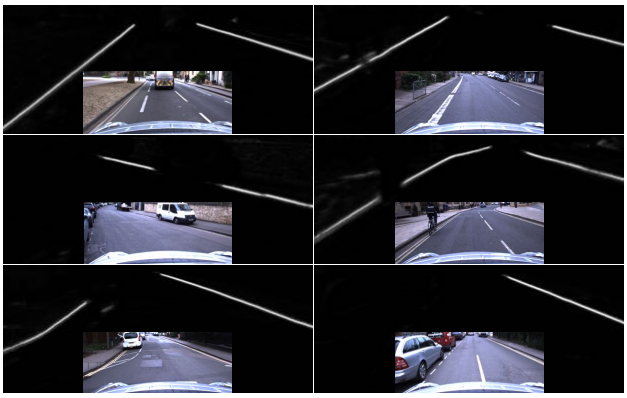


Fig. 4. Curbs are detected by a fully convolutional network. The network can detect visible curbs without making any assumptions about their 3D structure, shape or appearance.

corresponding them with the high-reflectance LiDAR points, which are not affected by varying lighting.

We employ these annotations to train a deep semantic segmentation network (inspired by U-net [19]) for road marking detection using only a monocular camera. The results demonstrate that the network segments the road markings from the image without any preprocessing steps, as shown in Fig. 3.

We direct the reader to [20] for a more detailed description of this method.

B. Curb Detection

Curbs (road boundaries) play an important role for autonomous cars as they intentionally and legally delimit driveable space. Curb detection using monocular images is a challenging problem. Road boundaries have narrow and long shapes which are not easily detectable. Deep networks often require large amounts of training data to obtain high-performance, well-generalised models. Due to colour, appearance, shape, perspective, illumination and background clutter, the training data should incorporate great variability changes. However, image by image hand labelling of the ground truth data is a time-consuming process. To avoid this problem and obtain a large amount of training samples, we generated 3D points cloud from 2D laser data and annotated points in the point cloud corresponding to road boundaries. Note that the 2D laser is attached vertically to the rear of a test car, which makes road boundaries easy to spot and annotate in the point cloud. The annotated points are projected to images of forward facing camera of the car. Lines are drawn between consecutive points to annotate road boundaries in-between the points. This way, hundreds of labelled images are obtained within an hour (approximately 750 images). A 10 kilometres dataset from the Oxford RobotCar Dataset introduced by [21] was annotated to generate several thousand semi-annotated masks. A vision based localiser was used to boost the number of training images by projecting labels from the annotated dataset to other traversals. However, some of the generated curbs masks contain annotations for occluded areas of curbs, such as over parked cars. To remove those redundant annotations, we trained U-net [19] with the raw masks and then run the inference with RGB images from the training data to generated output of detected curbs. The trained U-Net model can segment visible areas of curbs, but produces blurry outputs over occluding obstacles. Applying a threshold to the outputs gives us masks for detected visible curbs. We obtain labels for visible curbs by applying an AND operation between the thresholded outputs and raw labels. Finally, we train the U-net with visible curbs only (Fig. 4). A detailed description of our work on curb detection is given in [22].

V. Scene Interpretation

In the previous section, we explained how an input image is segmented into two classes: road markings and curbs. Before we describe how we learn a probabilistic grammar to parse these segmentations and construct a scene graph from them, we first introduce scene graphs formally.

A. Representation

Our motivation with this work is to support autonomous vehicles in their decision making, planning, and explanation generation. In particular, we aim at a representation that is interpretable (by machines and humans alike), extendable, and suitable for different inference tasks. To this end, we introduce *scene graphs* as a way to represent road scenes semantically using well-defined concepts and relations which are grounded in the vehicle's perception system.

Formally, scene graphs are represented in Description Logic; an overview is given in [23]. A scene is described by a set of instances of meaningful classes and their relations. For example, a *scene* is composed of a *road* which has two *curbs* and several *lanes* which in turn are bounded by several *road markings*. This hierarchical decomposition of a scene is important as we will explain later in Sec. V-C. In general, however, scene graphs can be linked flexibly to other information resources due to its underlying logical representation as we have shown in previous work [24]. For example, they can be linked to the outcome of detection and tracking algorithms of traffic participants and/or domain knowledge

TABLE I
SCENE GRAPH TAXONOMY

Class	Description
Scene	Root node of a scene graph. A *Scene* has at least one road (*Road*), but can have multiple.
Road	A road is delimited by at most two curbs (*Curb*) and has one or more lanes (*Lane*).
Curb	A curb is composed of one or multiple curb segments (*CurbSeg*).
Lane	A lane is bounded by road markings along the carriage way (*RMAlong*). Additionally, lanes can have road markings that are across the carriage way (*RMAcross*), and other road markings such as symbols and text (*RMOther*).
RMAlong	Road marking along the carriage way.
RMAcross	Road marking across the carriage way.
RMOther	Road marking of a symbol or text.
RMSeg	A road marking segment is a set of clustered pixels detected by the network described in Sec. IV-A. It can be one of three types: *RMALong*, *RMAcross*, or *RMOther*.
CurbSeg	A curb segment is a set of clustered pixels detected by the network described in Sec. IV-B.

TABLE II
LEARNT PROBABILISTIC CONTEXT-FREE GRAMMAR

Production rule
Scene $\rightarrow$ *Road*
Road $\rightarrow$ *Curb Lane*
Lane $\rightarrow$ *RMAlong RMAcross RMOther*
RMAlongCW $\rightarrow$ *RMSeg*
RMAcrossCW $\rightarrow$ *RMSeg*
RMOther $\rightarrow$ *RMSeg*
Curb $\rightarrow$ *CurbSeg*

defined by the Highway Code. This kind of knowledge can be encoded as logical rules within Description Logic.

A brief description of the most important concepts is given in Tab. I. It is important to note that entities that represent road marking segments (*RMSeg*) and curb segments (*CurbSeg*) are both linked to the output of the segmentation networks described in the previous section. Hence, instances of these types are grounded in image space. This is important as it allows us to reconstruct concepts higher-up in the hierarchy (e.g. Lanes) based on those low-level segmentations. In particular, we represent detected segments using axis-aligned and minimal area bounding boxes. More high-level concepts are represented as the bounding box of their children. Note that all other concepts are assigned based on the learnt grammar.

In the next section, we explain how we learn a probabilistic grammar for road scenes based on the introduced concepts.

B. Probabilistic Grammar

We adopt the approach by [11] and learn a probabilistic context-free grammar for road scenes from a set of annotated examples. To this end, we consider a set of scene graphs that have been manually annotated according to the concepts introduced in the previous section and based on the detections of road markings and curbs (Sec. IV). We learn the structure of the production rules and their probability from the frequency observed in the annotated set. The production rules are shown in Tab. II[1]

For each annotated scene graph we compute a set of geometric properties and spatial relations between instances that share the same parent node. We start the computation at the leaf nodes and propagate the results up the hierarchy. In our implementation, we consider several geometric properties including: length, width, and area for both axis-aligned and minimal bounding boxes. Furthermore, we consider the ratios between these properties to compute scores for the *axis-alignedness*, *alongness*, and *acrossness* of an instance. We also consider spatial relations between instances that share the same parent node (e.g. two boundaries of a lane). For these instances, we compute several relations including: the connectivity of the bounding boxes based on the Region Connection Calculus [25] and their relative angle and distance based on the Ternary Point Calculus [26]. In total we consider 18 geometric properties and 14 spatial relations. However, the details of how these properties and relations are not described here for brevity. Overall, the individual features are not critically important (and can be replaced). However, they provide us with the ability to assess the overall probability of the scene by considering all instances of a tree t given its geometric description and its relations. For each geometric property and relation we learn a probability distribution, namely $P_{geo}(x)$ and $P_{rel}(x)$, based on the annotated data using Kernel Density Estimation (based on Gaussian kernels). By computing the probability of each individual property and relation we can compute the overall probability of a tree based on the grounded representation as follows:

$$P(s|t,g) = \prod_{x \in t} P_{geo}(x) P_{rel}(x) \tag{1}$$

whereby s denotes a scene, t a tree, and g a grammar.

C. Scene Parsing

To reconstruct the layout of a road scene we use an extended version of a probabilistic Earley parser [27]. In general, the Earley algorithm is a dynamic programming approach that is able to handle ambiguous grammars. It combines top-down predictions and bottom-up recognitions to effectively parse its input. The algorithm has three main steps: *predict*, *scan*, and *complete*. In the predict step, rules are expanded according to the grammar. This step guides the overall search in a top-down way (initially the root node is expanded). In the scan step, the next input symbol is read and compared to the next one that was predicted. If a production rule is completed, the complete step has found a valid parse of a subtree and overall search is advanced. This type of

[1]Note, that we have omitted the learnt probabilities as we have learn different rules for different cardinalities.

hybrid search using top-down down reasoning and bottom-up perception for scene understanding can be very effective in real-world scenarios as we have shown earlier [28].

Our adapted version of the parser takes the learnt probabilistic grammar and a sequence of curb and road marking segments as input. The segments form the lexicon of our grammar and their probabilities are determined according to $P_{geo}(X)$ as defined in the previous section.

After the parser has recognised the input, a forest of parse trees can be retrieved. In our implementation we use a shared packed parse forest (SPPF) to store the ambiguous parse trees [29]. Parse trees are evaluated according their probabilities computed as follows:

$$P(t|s,g) = P(t|g)P(s|t,g) \tag{2}$$

whereby t denotes a parse tree, s the scene, and g the grammar. $P(t|g)$ is the product of all probabilities according to the production rules and $P(s|t,g)$ represents the data likelihood of seeing this scene given the tree and the grammar. Eventually, the best parse tree t^* can be chosen according to the overall probability:.

$$t^* = \arg\max_{t \in \mathcal{T}} P(t|s,g) \tag{3}$$

whereby t denotes a parse tree in the parse forest $\mathcal{T}$, s the scene, and g the grammar.

VI. Experiments

In this section we present the experimental setup and discuss qualitative results of our approach.

A. Experimental Setup

In this work, we evaluated the overall pipeline as depicted in Fig. 2. A given input image is processed by the road marking and the curb detection networks. The output of these networks is a probability distribution of segments in the image space. Using Inverse Perspective Mapping (IPM), we transform each of the segmented images into a birds eye view (see Tab. III). For each class, we then find clusters that represent these entities by their bounding boxes and compute a set of geometric features. Based on their visual and geometric probability these segmented entities are added to the lexicon of the grammar.

The Earley algorithm predicts the structure of the scene based on the learnt grammar and parses the segments from left to right in image space. We evaluated the generated parse trees according to their probability. However, given the high ambiguity of rules in the learnt grammar, we have selected a few examples manually (Tab. III). In the next section we discuss several theses examples and point to interesting and/or problematic aspects.

B. Qualitative Results

Tab. III depicts the qualitative results for several scenes. The table shows the input image; the different segments produced by the networks and the clustering step (road markings in green; curbs in orange); and the generated scene graphs (or parts of it).

Scene (a) In this scene (see Fig. 1), the segmentation captures curbs on both sides of the road as well as road markings along the carriage way. However, a stop line as well as the bicycle symbol are not detected. By integrating some domain knowledge from the Highway Code in form of rules, we can refine the scene graph by inferring that there is a bicycle lane on the left-hand side as the lane's width is too narrow for a standard car lane. These rules are encoded within Description Logic and can infer classes which were not labelled in any of the examples. However, we are not able to infer the same on the right-hand side as we do not have any meaningful segment that describes the boundary of the bicycle lane on the right-hand side. The detection of road markings and curbs in roads other than the main road is typically more challenging as they are perceived at the edge of the camera's field of view.

Scene (b) In this scene the parser detects two road markings on a lane. Given their size and spatial relation we can infer that these entities are road markings that introduce speed humps on the road.

Scene (c) This scene is interesting as there are curb structures in the middle of the road. Furthermore, the left lane has two stop-lines. However, it is important for an autonomous vehicle to infer that is has to stop in front of the first one. Note, that such an inference can only be drawn when local context of the scene is considered, but not from the single segment alone. These are situations in which we believe that background knowledge and AI reasoning techniques can have a great impact when interpreting scenes.

Scene (d) In this scene both curbs and road markings are well detected (except for the degraded dotted line across the road). However, this scene provides an interesting and rare case as the road markings for the car (zig-zag line) and the bicycle lane overlap. Momentarily this cannot be represented by our grammar as we made the assumption that lanes are next to each other.

In future work we will also perform a quantitative analysis of our approach, in particular with respects to its real-time capabilities. In general the Early algorithm is well-suited for real-time applications as its worst time complexity is $O(n^3)$. However, retrieving and processing a potential exponential number of parse trees might be challenging.

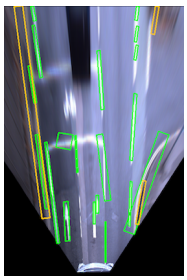
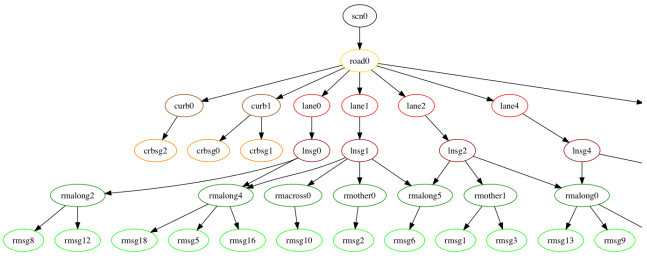
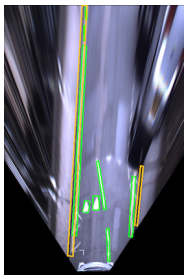
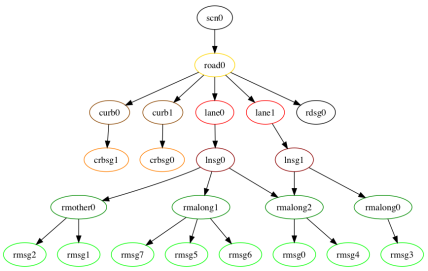
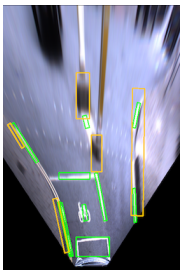
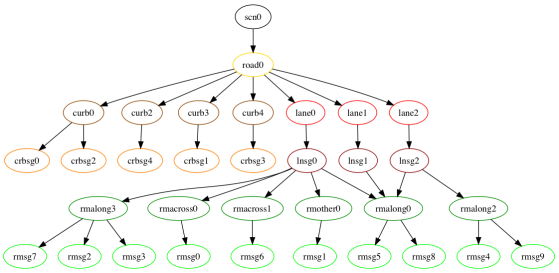
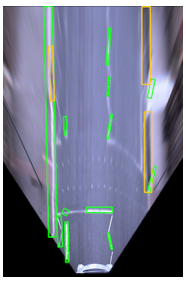
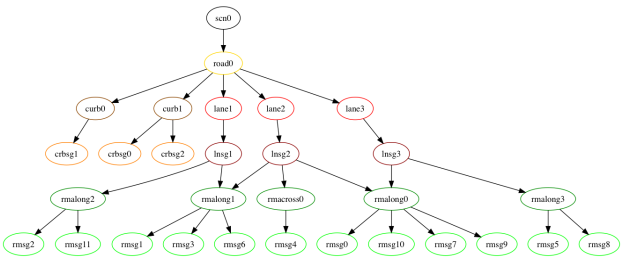
VII. Discussion

In this section we would like to provide a brief overview of the application space of the scene graph.

1) Urban traffic scenes are highly structured since they are built consistently according to specified road rules. By incorporating these rules, certain nodes in the scene graph can be classified. For instance, a bicycle lane is easily distinguished from a car lane by comparing the width. In this way, the scene graph allows for classification of road objects/segments without requiring expensive manual labels.

2) The segmented scenes given by the scene graph can be employed to bootstrap deep learning models. As stated above classification labels which can be used for training

TABLE III
QUALITATIVE RESULTS

ID	Original (RGB)	Segments (IPM)	Scene Graph (partial)
(a)			
(b)			
(c)			
(d)			

purposes can be acquired without expensive manual annotation. Furthermore, the scene graph provides an informed indication about the likely location of road objects (e.g. curbs, road markings). This could be used when training deep networks for instance to guide attention or to adjust the loss and thereby improve performance. In this way, important prior information about the environment is included in a deep learning approach (which is non-trivial).

3) Scene graphs can be used for (cost-based) planning for autonomous vehicles as they reason about the lane geometry and can infer road marking classes based on contextual spatial relations. For instance, a solid boundary of a bicycle lane should only be crossed in case of emergency. Besides, actions are now interpretable because we can review the representation inferred from the segmentation.

4) The scene graph is able to predict/hallucinate missing objects because of the learned spatial and semantic constraints. For example, two-way roads with missing lane markings in the middle will not fit the learned representations (nor the road rules). The scene graph can predict the most likely lane geometry in that case.

We think that these examples are interesting uses cases with exciting technological challenges for applications of scene graphs.

VIII. CONCLUSION

In this paper we presented an approach for scene understanding of complex urban environments. To this end, we

proposed *scene graph*, a hierarchical, graph-based representation, and a parsing pipeline that generates and evaluates scenes graphs based on partially segmented images, a learnt probabilistic grammar, as well as geometric and relational models. Furthermore, we have presented and discussed several example scenarios in which scene graphs can provide meaningful insights in the overall structure of the environment. The construction and interpretation of interpretable and auditable scene graphs can play essential role in many tasks of autonomous vehicles including planning, decision making, and explanation generation. Hence we believe that this functionality can have wide impact in the context of autonomous driving and mobile robotics in general.

ACKNOWLEDGMENT

The work has been supported by the EPSRC/UK Research and Innovation Programme Grant EP/M019918/1 (Mobile Autonomy: Enabling a Pervasive Technology of the Future).

We acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[3] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3d traffic scene understanding from movable platforms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 1012–1025, 2014.

[4] J. Ziegler, P. Bender, M. Schreiber, H. Lategahn, T. Strauss, C. Stiller, T. Dang, U. Franke, N. Appenrodt, C. G. Keller *et al.*, "Making bertha drive an autonomous journey on a historic route," *IEEE Intelligent Transportation Systems Magazine*, vol. 6, no. 2, pp. 8–20, 2014.

[5] J.-B. Bordes, F. Davoine, P. Xu, and T. Denœux, "Evidential grammars: A compositional approach for scene understanding. application to multimodal street data," *Applied Soft Computing*, vol. 61, pp. 1173–1185, 2017.

[6] F. Dierkes, M. Raaijmakers, M. T. Schmidt, M. E. Bouzouraa, U. Hofmann, and M. Maurer, "Towards a multi-hypothesis road representation for automated driving," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*. IEEE, 2015, pp. 2497–2504.

[7] D. Töpfer, J. Spehr, J. Effertz, and C. Stiller, "Efficient road scene understanding for intelligent vehicles using compositional hierarchical models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 441–451, 2015.

[8] J. Spehr, D. Rosebrock, D. Mossau, R. Auer, S. Brosig, and F. M. Wahl, "Hierarchical scene understanding for intelligent vehicles," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE, 2011, pp. 1142–1147.

[9] S. Kashetty Venkateshkumar, M. Sridhar, and P. Ott, "Latent hierarchical part based models for road scene understanding," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[10] B. Mathibela, P. Newman, and I. Posner, "Reading the road: Road marking classification and interpretation," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2072–2081, 2015. [Online]. Available: http://dx.doi.org/10.1109/TITS.2015.2393715

[11] T. Liu, S. Chaudhuri, V. G. Kim, Q.-X. Huang, N. J. Mitra, and T. Funkhouser, "Creating consistent scene graphs using a probabilistic grammar," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 33, no. 6, Dec. 2014.

[12] A. Ess, T. Mueller, H. Grabner, and L. van Gool, "Segmentation-based urban traffic scene understanding," in *Proceedings of the British Machine Conference, pages*, 2009, pp. 84–1.

[13] A. Joshi and M. R. James, "Generation of accurate lane-level maps from coarse prior maps and lidar," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 1, pp. 19–29, 2015.

[14] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker, "Learning to look around objects for top-view representations of outdoor scenes," *arXiv preprint arXiv:1803.10870*, 2018.

[15] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," *arXiv preprint arXiv:1705.07115*, 2017.

[16] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," *arXiv preprint arXiv:1803.11189*, 2018.

[17] X. Liang, H. Zhou, and E. Xing, "Dynamic-structured semantic propagation network," *arXiv preprint arXiv:1803.06067*, 2018.

[18] N. Nauata, H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Structured label inference for visual understanding," *arXiv preprint arXiv:1802.06459*, 2018.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[20] T. Bruls, W. Maddern, A. A. Morye, and P. Newman, "Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data," in *Robotics and Automation (ICRA), 2018 IEEE International Conference on*. IEEE, 2018.

[21] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017. [Online]. Available: http://dx.doi.org/10.1177/0278364916679498

[22] T. Suleymanov, P. Amayo, and P. Newman, "Inferring road boundaries through and despite traffic," in *The 21st IEEE International Conference on Intelligent Transportation Systems*, November 2018.

[23] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds., *The Description Logic Handbook: Theory, Implementation, and Applications*. New York, NY, USA: Cambridge University Press, 2003.

[24] M. Tenorth, L. Kunze, D. Jain, and M. Beetz, "Knowrob-map - knowledge-linked semantic object maps," in *2010 10th IEEE-RAS International Conference on Humanoid Robots*, Dec 2010, pp. 430–435.

[25] D. A. Randell, Z. Cui, and A. G. Cohn, "A spatial logic based on regions and connection," in *KR*. Morgan Kaufmann, 1992, pp. 165–176.

[26] R. Moratz and M. Ragni, "Qualitative spatial reasoning about relative point position," *Journal of Visual Languages & Computing*, vol. 19, no. 1, pp. 75–98, 2008, spatial and Image-based Information Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1045926X06000723

[27] J. Earley, "An efficient context-free parsing algorithm," *Commun. ACM*, vol. 13, no. 2, pp. 94–102, Feb. 1970. [Online]. Available: http://doi.acm.org/10.1145/362007.362035

[28] L. Kunze, C. Burbridge, M. Alberti, A. Tippur, J. Folkesson, P. Jensfelt, and N. Hawes, "Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Chicago, Illinois, US, September, 14–18 2014.

[29] E. Scott, "Sppf-style parsing from earley recognisers," *Electronic Notes in Theoretical Computer Science*, vol. 203, no. 2, pp. 53 – 67, 2008, proceedings of the Seventh Workshop on Language Descriptions, Tools, and Applications (LDTA 2007). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1571066108001497