Reading between the Lanes: Road Layout Reconstruction from Partially Segmented Scenes (Extended Abstract)

Lars Kunze, Tom Bruls, Tarlan Suleymanov, and Paul Newman
Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK
{lars, tombruls, tarlan, pnewman}@robots.ox.ac.uk

***Abstract*—Autonomous vehicles require an accurate and adequate representation of their environment for decision making and planning in real-world driving scenarios. To this end, we introduce a hierarchical, graph-based representation, called *scene graph*, which is reconstructed from a partial, pixel-wise segmentation resulting from trained deep networks. By harvesting the power of deep learning and generating an interpretable representation of road scenes which can be linked to domain knowledge and AI reasoning techniques, we believe that this approach provides a vital step towards explainable and auditable models in the context of autonomous driving.**

I. Introduction

Autonomous vehicles need to perceive their surroundings accurately for safe navigation in complex urban environments. 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 reasoning and planning methods required for complex road manoeuvres. In this work, we introduce *scene graph*, a hierarchical, graph-based representation, 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 the literature, there exist several approaches that model (traffic) scenes using graphs. Work by [2] fuses different sensor modalities and hierarchical graphs containing relational knowledge, [14] connects road markings as a graph and optimises a CRF with handcrafted spatial features to predict their class, and [9] models lane separators as latent variables to infer lane geometries. While [5] presents a theoretical framework including uncertainties to reason about multiple hypotheses for lane geometry, other approaches consider real-world data [19, 17] as we do here.

Recently, deep semantic segmentation networks have achieved impressive results for pixel-wise scene understanding of images [1, 20]. However, these methods suffer from interpretation difficulties and often fail to include prior information or constraints. Several works have addressed this problem by introducing spatial and semantic reasoning frameworks that can be trained in an end-to-end way [4, 11, 15].

In contrast, all important objects influencing the decision making are detected separately in mediated approaches [7, 21, 8]. Similarly, we consider different aspects of a scene in a combined way whereby we model the geometry and relations of high-level concepts based on low-level segmentations.

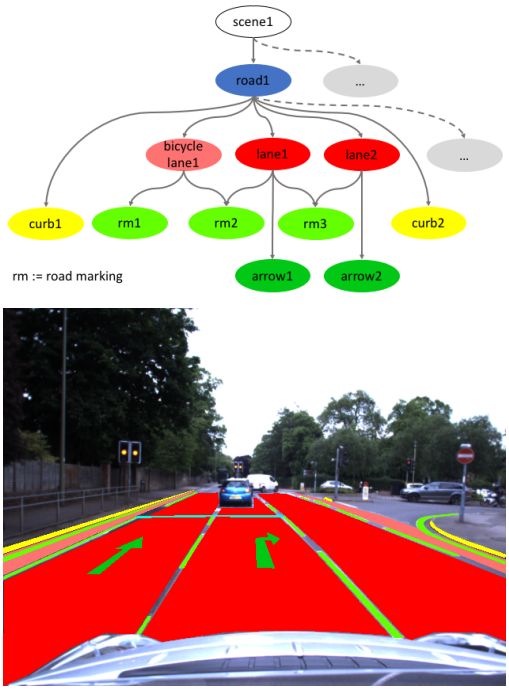


Fig. 1. Hierarchical scene graph representation (top) that was reconstructed from a partially segmented image (bottom). In this work we present a probabilistic scene parser that reconstructs the layout of road scenes from partial segmentations of road markings and curbs.

Fig. 1 shows an example scene graph for a segmented road scene. 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 an image by detecting curbs and road markings using trained deep networks (Sec. II). 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. III). The parser takes object detections and prior information of road scenes into account. Each valid parse tree is scored by a probability which allows us to disambiguate between alternative hypotheses. 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

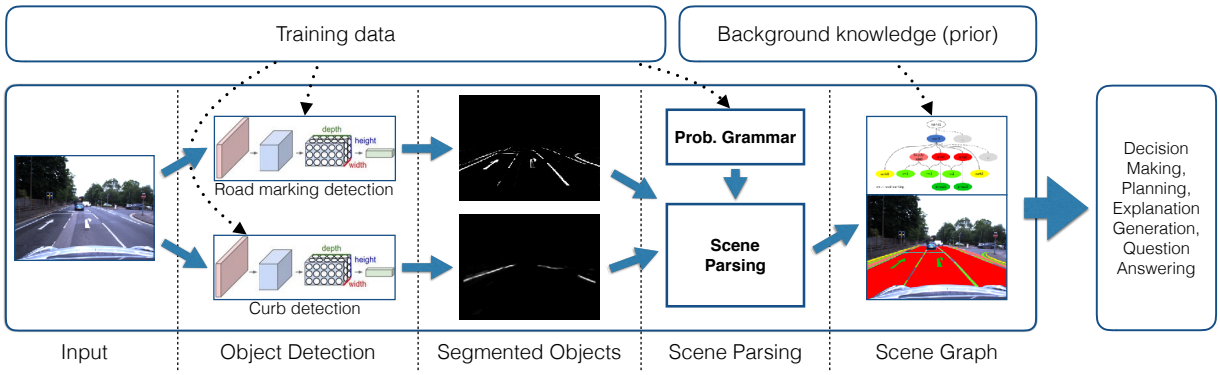


Fig. 2. Scene parsing approach based on road marking and curb detections. The approach has two main steps: (1) given an image, road marking and curb segments are detected by deep networks, and (2) given a set of detected segments, the scene is parsed using an adapted version of the Earley algorithm and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

representations they can be linked to background knowledge and used for auditable planning and decision making (Sec. IV).

II. Scene Perception

Road markings and curbs are critical components for (autonomous) driving especially in urban environments. Road rules are captured by their underlying meaning and they guide all traffic participants through potentially dangerous situations. Therefore, real-time detection and interpretation is important for scene understanding, planning and decision making.

Detecting road markings and curbs, which dictate the traffic rules, using monocular images is a challenging problem. Firstly, there are visual challenges such as occlusions, varying lighting, and changing weather conditions. Secondly, there are no large datasets available for training with accurate ground-truth labels for road markings and curbs. The problem can be seen as a semantic segmentation problem, for which deep network approaches are currently the state-of-the-art. Hence, we train two deep semantic segmentation networks (inspired by U-net [16]) on the Oxford RobotCar Dataset [13] whereby scenes were annotated semi-automatically using other sensor modalities such as Lidar (road markings) and 2D laser (curbs). A more detailed description of the weakly-supervised annotation process, training, and network architecture is given in [3]. The resulting segmented, pixel-based images are clustered and segmented entities are obtained which are considered as input for the scene interpretation process.

III. Scene Interpretation

In this work, 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.

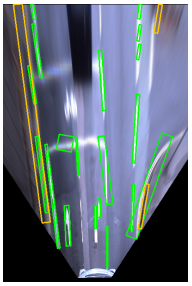
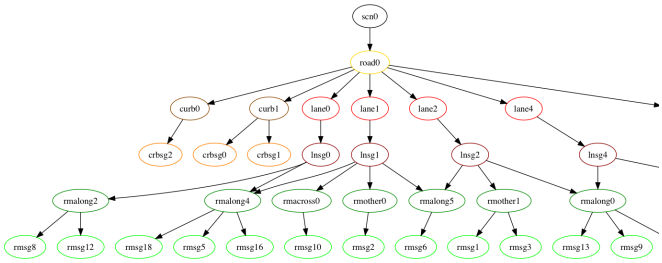
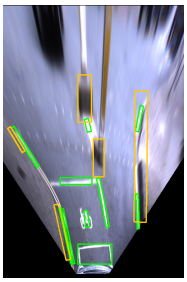
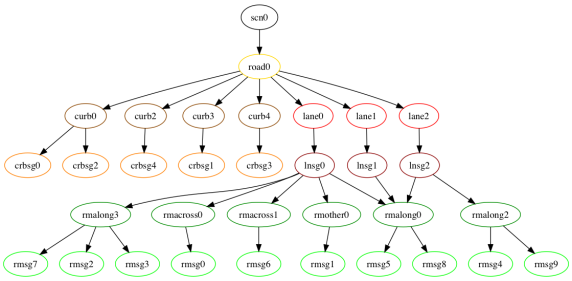
Formally, scene graphs are represented in Description Logic. A scene is described by a set of instances of meaningful classes and their relations. For example, a *scene* is composed of a *road* which has two *curbs* and several *lanes* which in turn are bounded by several *road markings*. It is important to note that segments of road markings and curbs are both linked to the output of the segmentation networks described above, and hence, grounded in image space. This is important as it allows us to reconstruct concepts higher-up in the hierarchy (e.g. Lanes) based on low-level segmentations. In general, scene graphs can be linked flexibly to other information due to its underlying logical representation as we have shown [18]. For example, they can be linked to the outcome of detection and tracking algorithms of traffic participants and/or domain knowledge defined by the Highway Code.

In this work, we adopt the approach by [12] 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 well-defined concepts and based on the detections of road markings and curbs (Sec. II). We learn the structure of the production rules and their probability from their frequency in the data.

For each annotated scene graph we compute a set of geometric properties and spatial relations between instances that share the same parent node. These geometric properties and relations provide us with the ability to assess the overall probability of the scene by considering all instances of a tree t. 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 $P(s|t,g) = \prod_{x \in t} P_{geo}(x) P_{rel}(x)$, whereby s denotes a scene, t a tree, and g a grammar.

To reconstruct the layout of a road scene we use an extended version of a probabilistic Earley parser [6]. In the algorithm's predict step, rules are expanded according to the grammar. This step guides the overall search in a top-down way. 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 hybrid search using

TABLE I
QUALITATIVE RESULTS

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

top-down down reasoning and bottom-up perception for scene understanding can be very effective in real-world scenarios as we have shown earlier [10].

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 previously.

After the parser has recognised the input, a forest of parse trees can be retrieved. Parse trees are evaluated according their probabilities: $P(t|s,g) = P(t|g)P(s|t,g)$.

$P(t|g)$ 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: $t^* = \arg\max P(t|s,g)$.

IV. QUALITATIVE RESULTS

We evaluated the overall pipeline as depicted in Fig. 2. Tab. I depicts the qualitative results for two scenes. It shows the input image; the different segments produced by the networks (road markings in green; curbs in orange); and the generated scene graphs (or parts of it).

Scene (a): In this scene (Fig. 1), the segmentation captures curbs on both sides of the road and segments parts of all road markings that are 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. Thereby we can infer classes which were not labelled in any of the examples.

Scene (b): 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 be very helpful when interpreting scenes.

In future work we will perform a quantitative analysis of our approach, in particular with respect to its real-time capabilities.

V. CONCLUSION

We presented an approach for scene understanding of complex, real-world environments. To this end, we proposed *scene graph*, a hierarchical, graph-based representation; 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. In general, we think that scene graphs have the potential to (1) reduce the amount of manual annotation by understanding scenes beyond labelled examples; (2) include prior information in the learning process of deep networks; (3) provide interpretable representations for planning and decision making; and (4) infer missing information based on prior models. These examples showcase interesting uses cases with exciting technological challenges for applications of scene graphs. Hence we believe that this functionality can have wide impact in the context of autonomous driving and mobile robotics in general.

Acknowledgments

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

References

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

[2] Jean-Baptiste Bordes, Franck Davoine, Philippe Xu, and Thierry Denœux. Evidential grammars: A compositional approach for scene understanding. application to multimodal street data. *Applied Soft Computing*, 61:1173–1185, 2017.

[3] Tom Bruls, Will Maddern, Akshay A. Morye, and Paul Newman. Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data. In *Robotics and Automation (ICRA), 2018 IEEE International Conference on*, page in press. IEEE, 2018.

[4] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. *arXiv preprint arXiv:1803.11189*, 2018.

[5] Frank Dierkes, Marvin Raaijmakers, Max Theo Schmidt, Mohamed Essayed Bouzouraa, Ulrich Hofmann, and Markus Maurer. Towards a multi-hypothesis road representation for automated driving. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 2497–2504. IEEE, 2015.

[6] Jay Earley. An efficient context-free parsing algorithm. *Commun. ACM*, 13(2):94–102, February 1970. ISSN 0001-0782. doi: 10.1145/362007.362035. URL http://doi.acm.org/10.1145/362007.362035.

[7] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3d traffic scene understanding from movable platforms. *IEEE transactions on pattern analysis and machine intelligence*, 36 (5):1012–1025, 2014.

[8] Avdhut Joshi and Michael R James. Generation of accurate lane-level maps from coarse prior maps and lidar. *IEEE Intelligent Transportation Systems Magazine*, 7(1):19–29, 2015.

[9] Suhas Kashetty Venkateshkumar, Muralikrishna Sridhar, and Patrick Ott. Latent hierarchical part based models for road scene understanding. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[10] Lars Kunze, Chris Burbridge, Marina Alberti, Akshaya Tippur, John Folkesson, Patric Jensfelt, and Nick 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.

[11] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. *arXiv preprint arXiv:1803.06067*, 2018.

[12] Tianqiang Liu, Siddhartha Chaudhuri, Vladimir G. Kim, Qi-Xing Huang, Niloy J. Mitra, and Thomas Funkhouser. Creating consistent scene graphs using a probabilistic grammar. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 33(6), December 2014.

[13] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. doi: 10.1177/0278364916679498. URL http://dx.doi.org/10.1177/0278364916679498.

[14] Bonolo Mathibela, Paul Newman, and Ingmar Posner. Reading the road: Road marking classification and interpretation. *IEEE Trans. Intelligent Transportation Systems*, 16(4):2072–2081, 2015. doi: 10.1109/TITS.2015.2393715. URL http://dx.doi.org/10.1109/TITS.2015.2393715.

[15] Nelson Nauata, Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Structured label inference for visual understanding. *arXiv preprint arXiv:1802.06459*, 2018.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[17] Jens Spehr, Dennis Rosebrock, Daniel Mossau, Richard Auer, Stefan Brosig, and Friedrich M Wahl. Hierarchical scene understanding for intelligent vehicles. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 1142–1147. IEEE, 2011.

[18] 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*, pages 430–435, Dec 2010. doi: 10.1109/ICHR.2010.5686350.

[19] Daniel Töpfer, Jens Spehr, Jan Effertz, and Christoph Stiller. Efficient road scene understanding for intelligent vehicles using compositional hierarchical models. *IEEE Transactions on Intelligent Transportation Systems*, 16 (1):441–451, 2015.

[20] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[21] Julius Ziegler, Philipp Bender, Markus Schreiber, Henning Lategahn, Tobias Strauss, Christoph Stiller, Thao Dang, Uwe Franke, Nils Appenrodt, Christoph G Keller, et al. Making bertha drive an autonomous journey on a historic route. *IEEE Intelligent Transportation Systems Magazine*, 6(2):8–20, 2014.