A Dataset for High-level 3D Scene Understanding of Complex Road Scenes in the Top-View

Ziyan Wang[1] Buyu Liu[2] Samuel Schulter[2] Manmohan Chandraker[2,3]
[1]Carnegie Mellon University [2]NEC Laboratories America [3]UC San Diego

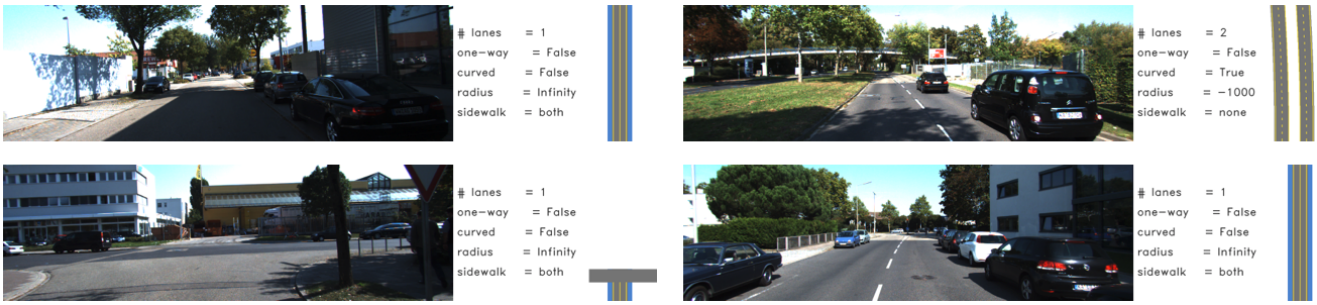

Figure 1: The figure shows four images from the KITTI [5] dataset along with or novel set of annotations for high-level scene understanding of complex road scenes. On the right of each RGB image, we provide a small subset of the scene attributes that we manually annotated. The attributes are rich enough to build an abstract but semantically and geometrically meaningful representation of the scene in a top-view, which is shown to the right of the attributes.

Abstract

We introduce a novel dataset for high-level 3D scene understanding of complex road scenes. Our annotations extend the existing datasets KITTI [5] and nuScenes [1] with semantically and geometrically meaningful attributes like the number of lanes or the existence of, and distance to, intersections, sidewalks and crosswalks. Our attributes are rich enough to build a meaningful representation of the scene in the top-view and provide a tangible interface to the real world for several practical applications.

1. Introduction

Understanding the 3D world from 2D images is an important but challenging task in computer vision with several immediate applications ranging from augmented reality over indoor robot navigation to autonomous driving and advanced driver assistance systems. With the recent push from the automotive industry, handling complex outdoor driving scenarios became of high interest. Several large-scale datasets with manual annotation for standard perception tasks like object detection [8, 12, 15] or semantic segmentation [2, 19] have been collected for this domain [1, 3, 5, 6, 10, 16, 18].

However, it is obvious that a comprehensive understanding of complex 3D scenes requires high-level reasoning beyond object detection or semantic segmentation. Thus, an increasing amount of works started to focus on inferring scene attributes or parameterized models of complex road scenes [4, 7, 9, 14, 17]. A modern dataset for such high-level reasoning tasks is lacking, though. Although most prior work infer some parameterized model [4, 7], no ground truth is given for either learning or evaluation. One exception is Seff and Xiao [14], who automatically acquire scene attributes from OpenStreetMaps (OSM) [11]. While being at a large-scale, the scene attributes are (i) limited to what OSM provides and (ii) noisy due to inconsistencies in annotation and a lack of quality control.

In this work, we provide a novel dataset for scene understanding and high-level reasoning in complex driving scenes.[1] It provides annotations for road scenes in the form of hand-labelled scene attributes like the topology of the road layout, the distances to intersections, the number of lanes or the existence of sidewalks and crosswalks. These attributes are semantically and geometrically meaningful, which is beneficial for higher-level modeling and decision making as they provide a tangible interface to the real world. The set of attributes is rich enough to reconstruct various scene layouts in a semantic top-view representation as in [17], see

[1]This is an extended version of our dataset presented in [17].

ID	Description
0	Can the scene be represented with the simulator?
1	Number of lanes
2	Which one is the ego lane?
3	Is the main road a one-way?
4	Additional lanes on opposing driving direction
5	Does a turnlane exist?
6	Is the main road curved (straight or radius)?
7	Rotation of main road (car makes a turn)
8	Delimiter width between driving directions on main road
9	Do sidewalks exist?
10	Delimiter width between main road and sidewalks
11	Crosswalk on main road w/o intersections
12	Distance to this crosswalk
13	Do side roads exist (no, left, right or both)?
14	Does the main road end with side-roads? (T-intersection)
15	Does the main-road delimiter end at intersection?
16	Distance to left side-road
17	Number of lanes of a left side-road
18	Distance to right side-road
19	Number of lanes of a right side-road
20	Is there a crosswalk before the intersection?
21	Is there a crosswalk after the intersection?
22	Is there a crosswalk on the left side-road?
23	Is there a crosswalk on the right side-road?

Table 1: Scene attributes that are being annotated.

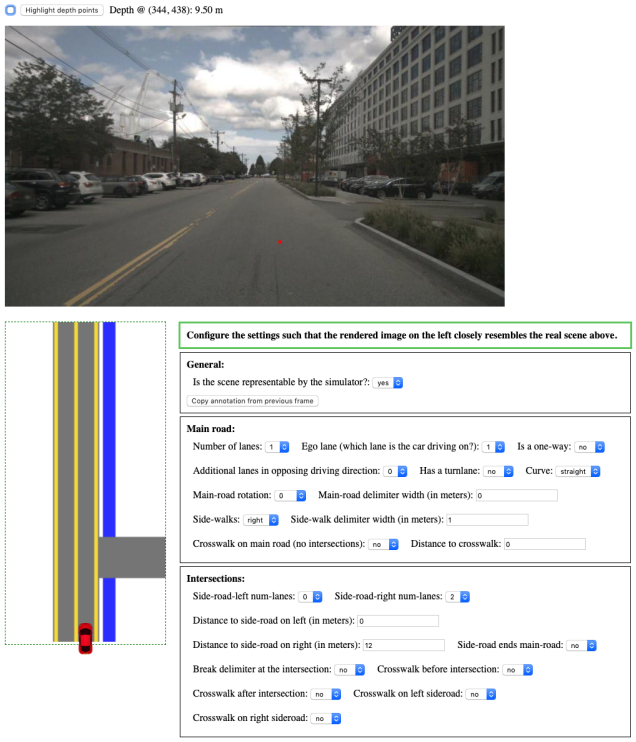

Figure 2: The annotation tool for our scene attributes: The user sees the RGB image as well as the sparse depth points when hovering over the image (red dot, all points can be highlighted too). We ask annotators to fill out the form below the image with all attributes that describe the scene. As soon as the annotator changes any value in the form, we use the attributes to render an abstract semantic top-view (left to the form), which gives immediate feedback.

the teaser figure for a few examples.

Our scene understanding approach presented in [13, 17] relies on ground truth annotation for semantic segmentation, depth prediction in the perspective view, as well as scene attribute annotation in the top-view. Instead of collecting a new set of images with accurate depth annotation, which is cumbersome, we augment KITTI [5] and nuScenes [1], which both already provide RGB images and depth estimates. Specifically, we provide our manual annotations for both semantic segmentation and top-view scene attributes. The dataset will be available at http://www.nec-labs.com/~mas/BEV.

2. Dataset

There are many existing datasets that focus on road scenes [5, 1, 6, 3]. We thus choose to augment those existing datasets instead of collecting a new set of videos. Specifically, we use KITTI [5] and nuScenes [1] as both provide RGB images along with well calibrated geometric data like 3D point clouds from a Lidar scanner, IMU and GPS data. This information is essential for learning scene understanding models [13, 17] as well as annotating higher-level attributes like ours.

The set of attributes that we annotate is summarized in Tab. 1. These attributes are relative to the car that is driving on the road and observing the scene. The annotation provides information about the layout of the road (number of lanes, one way street, side roads, distances to side roads), sidewalks and crosswalks. We refer the reader to [17] to see how these attributes are used in an actual scene model. Note that these annotations are converted into other variables in [17]. For instance, number of lanes and the ego-lane is converted into the number of lanes on the left and right of the ego lane, because it may be easier to learn for a neural network.

Besides these attribute annotations, we also provide semantic segmentation ground truth for a subset of images as well as pre-computed output from our occlusion-reasoning work [13], which serves as input to many models in [17].

2.1. Annotation process

Fig. 2 shows the web-based annotation tool that we use to collect the scene attributes defined in Tab. 1. The figure caption explains how the annotation tool works.

We ask our annotators to describe the scene as closely as possible with the available set of attributes and to use the depth estimates for any distance-related attribute. If a scene is not representable with the tool, the user should indicate

(a) Binary variables (b) Number of lanes (c) Existence of side roads (d) Distances to side roads

Figure 3: Four different statistics of our scene attribute annotations on the KITTI [5] training dataset. (a) The histogram shows the frequency how often certain binary variables occur in the scene. The numbers on the x-axis correspond to the IDs in Tbl. 1. (b-c) Histograms of multi-class attributes for the number of lanes (in one driving direction) and the existence of side roads (0=none, 1=left, 2=right, 3=both). (d) The distribution of the distance to the right side road (if it exists).

that in the form too. Since most of the data comes from a video sequence, we let annotators process the data in order and the tool copies all attributes from the previous frame automatically. Since many attributes stay constant over a long time, this feature reduces annotation cost significantly. Finally, we also perform various quality checks by letting an expert verify a subset (5%) of the data.

2.2. Dataset statistics

Our dataset consists of 18381 and 18785 images for KITTI [5] and nuScenes [1], respectively. All of them are annotated with the scene attributes from Tbl. 1. A subset of the images are annotated for semantic segmentation, 1307 for KITTI and 939 for nuScenes.We made sure the label space on both datasets contains categories that are relevant for predicting the scene attributes, like road, lane boundaries, sidewalks, crosswalks, traffic signs, cars and pedestrians.

For the KITTI dataset [5], we further provide some statistics of our annotated scene attributes in Fig. 3. For binary variables in Tbl. 1, we show the frequency of occurrence in the dataset. Multi-class variables and continuous variables are also shown as histograms. Complete statistics will be provided with the final dataset release.

2.3. Evaluation

In [17], we used standard metrics for evaluation, like plain accuracy and MSE. Because many attributes have a naturally biased distribution, it can be better to report F1-scores, but this can also be sensitive for extreme biases (see discussion in the supplemental of [17]).

Besides evaluating attributes separately, we also propose in [17] to actually render the semantic top-view given the predicted and ground attributes and measure the Intersection-over-Union (IoU) between them. While we are not able to release our own rendering function because it is proprietary, we want to note that it is easy to implement a different rendering function and we provide some considerations here:

- Initialize a canvas of size 256×128 corresponding to 60×30 meters.
- Place the ego-car in the bottom center of the canvas.
- Compute the number of lanes on left and right of ego-car from attributes 1-5.
- Each item (like lanes, sidewalks, crosswalks) can be easily represented with simple polygons.
- Compute a polygon of each lane considering the road properties (attributes 6-8).
- Align side roads, crosswalks and sidewalks relative to the main road.

2.4. Potential applications

We hope that our annotations enable various novel applications and tasks. In [17], we already showed that one can learn models to predict the scene attributes from perspective RGB images, which allows for reconstructing a basic layout of the road scene in the top-view. Such a representation can be visualized intuitively for humans and is free from occlusion caused in the perspective front view. Extending our attributes will likely enable predicting even more detailed layouts of road scenes.

The parametric nature of our annotation automatically provides a tangible interface to the 3D world that describes the topology and the spatial structure of the road, including intersections, sidewalks and crosswalks. It also gives semantic meaning to each element in the representation, which further makes it useful for several tasks like navigation, path planning and driveable area prediction. Compared with non-parametric representations like semantic segmentation, our annotations allow for directly accessing attributes like distances to elements like crosswalks or intersections.

3. Relation to other datasets

To push the ability of machine perception in autonomous driving, an increasing number of benchmarks has been established recently [1, 3, 5, 6, 10, 16, 18]. However, unlike our dataset, none of these provide high-level attribute annotations as described above in Sec. 2.

KITTI [5] provides a collection of benchmarks for 2D/3D object detection, 2D semantic segmentation, stereo and opti-

cal flow. Around 30000 frames are collected from different types of sensors like RGB camera, Lidar, IMU and GPS, subsets of which are annotated with bounding boxes or semantic segmentation. More recently, the nuScene [1] and ApolloScape [6] datasets were released, which both follow a similar settings as KITTI [5], but at a larger scale. Data is collected in Boston and Singapore for nuScenes and in China for ApolloScape. All three datasets contain RGB video sequences as well as geometric information and thus qualify for our additional annotations. We ultimately chose to augment KITTI because it is already well established in the community and nuScenes because of its well-aligned sensor suite and well-documented API.

Other datasets like CityScapes [3], Mapillary Vistas [10] or TorontoCity [16] all capture diverse road scenes at large-scale but do not provide high-accuracy geometric information from a Lidar scanner, for instance.

One dataset similar to our work is from Seff and Xiao [14], who propose to predict various attributes of road layouts from a single RGB image. However, the scene attributes in [14] are not enough for reconstructing the whole scene in the top-view. Our dataset contains a richer collection of scene attributes, which is sufficient to describe scenes with more complex road layouts as already demonstrated in [17]. Moreover, the annotation in [14] was automatically collected from OpenStreetMap [11] which is noisy and often inconsistent. Our data on the other hand is manually annotated.

4. Conclusion

High-level reasoning of complex road scenes from RGB images is an important ability for scene understanding, and may aid potential high-impact applications in the automotive industry. In this work, we provide a novel set of annotations that augment existing datasets (KITTI [5] and nuScenes [1]) with semantically and geometrically meaningful scene attributes. We hope that this data is useful to the community and will be used for various novel applications.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*, 2018.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016.

[4] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3D Traffic Scene Understanding from Movable Platforms. *PAMI*, 2014.

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[6] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *CVPR Workshop*, 2018.

[7] Lars Kunze, Tom Bruls, Tarlan Suleymanov, and Paul Newman. Reading between the Lanes: Road Layout Reconstruction from Partially Segmented Scenes. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2018.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, 2017.

[9] Gellért Máttyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. HD Maps: Fine-grained Road Segmentation by Parsing Ground and Aerial Images. In *CVPR*, 2016.

[10] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

[11] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015.

[13] Samuel Schulter, Menghua Zhai, Nathan Jacobs, and Manmohan Chandraker. Learning to Look around Objects for Top-View Representations of Outdoor Scenes. In *ECCV*, 2018.

[14] Ari Seff and Jianxiong Xiao. Learning from Maps: Visual Common Sense for Autonomous Driving. *arXiv:1611.08583*, 2016.

[15] Bharat Singh and Larry S. Davis. An Analysis of Scale Invariance in Object Detection - SNIP. In *CVPR*, 2018.

[16] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. In *ICCV*, 2017.

[17] Ziyan Wang, Buyu Liu, Samuel Schulter, and Manmohan Chandraker. A Parametric Top-View Representation of Complex Road Scenes. In *CVPR*, 2019.

[18] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

[19] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *CVPR*, 2017.