# A dataset for benchmarking vision-based localization at intersections

supplemental material for "Vision-Based Localization at Intersections using Digital Maps"

Augusto L. Ballardini, Daniele Cattaneo, and Domenico G. Sorrenti

***Abstract*—In this report we present the work performed in order to build a dataset for benchmarking vision-based localization at intersections, *i.e.,* a set of stereo video sequences taken from a road vehicle that is approaching an intersection, altogether with a reliable measure of the observer position. This report is meant to complement our paper "Vision-Based Localization at Intersections using Digital Maps" submitted to ICRA2019. It complements the paper because the paper uses the dataset, but it had no space for describing the work done to obtain it. Moreover, the dataset is of interest for all those tackling the task of online localization at intersections for road vehicles, *e.g.,* for a quantitative comparison with the proposal in our submitted paper, and it is therefore appropriate to put the dataset description in a separate report. We considered all datasets from road vehicles that we could find as for the end of August 2018. After our evaluation, we kept only sub-sequences from the KITTI dataset. In the future we will increase the collection of sequences with data from our vehicle.**

***Index Terms*—intersection, crossing, KITTI, Dataset, OpenStreetMap, Map-Alignment, RTK, stereo, localization**

## I. Introduction

In order to benchmark vision-based localization at intersections we need at least the following streams of data from a vehicle approaching an intersections: the streams from the two cameras of a stereo rig, and a reasonably accurate position estimate of the vehicle, to be used as GT (Ground Truth). Quite un-expectedly, given the number of datasets from road vehicles available today, collecting such sequences resulted very difficult. We analyzed the sequences of the many available datasets, and checked whether everything needed was usable. Unfortunately, it turned out that, for various reasons, most of the material publicly available was not usable. This report describes how we ended up with a few more than forty usable sequences of a vehicle approaching an intersections, all from the KITTI dataset residential sequences. Although all sequences are from the same German city, which could compromise the generality of the dataset *w.r.t.* the variety of the world-wide intersections, the dataset does include different intersection geometries, (slightly) different lightning, and different traffic conditions. In spite of the effort put in building the dataset, and despite its uniqueness, we believe this dataset should be integrated with sequences from other countries, and more extreme light and traffic conditions. Nevertheless, as for today, we believe it is the best mean to benchmark a proposal about vision-based vehicle localization at intersections.

Dept. Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano - Bicocca, Milano, Italy; email: augusto.ballardini@unimib.it, daniele.cattaneo@disco.unimib.it, domenico.sorrenti@unimib.it

We analyzed the following datasets: KITTI [1], MALAGA [2], Oxford RobotCar [3], Rawseeds [4], and ApolloScape-Auto [5].

- KITTI: used, a stereo rig and GPS-RTK are available;
- Malaga: unused for the lack of GPS-RTK;
- Oxford RobotCar: unused for the lack of GPS-RTK;
- Rawseeds: no intersection between real roads is available, only intersections between roads in private areas, whose resemblance to real intersections has been considered not good enough;
- ApolloScapeAuto: unused for the lack of GPS-RTK.

This report describes the following aspects, which are relevant parts of the work performed on the KITTI dataset in order to extract the largest set of sequences of approaches to intersections: the intersection model is introduced in Section II, the issues with the camera calibration are reported in Section III, the issues with the alignments in OpenStreetMap are reported in Section IV, results about the sequences ending into the dataset are reported in Section V. A short conclusion is then followed by the aerial view of the KITTI residential sequences, in Section V.

## II. Intersection Model

The configuration of the road arms coming to an intersection bases on an intersection model, proposed in [6], which is an enhancement of the one presented in [7], [8]. It is depicted in Figure 1, and has the following parameters:

- The distance **c** of the intersection center, *w.r.t.* the hypothesized vehicle position;
- The number **n** of road segments / arms involved in the intersection;
- For each road segment $i$, the width $w_i$ and its orientation $r_i$ *w.r.t.* the road segment where the vehicle is.

This model allows us to represent almost every type of simple intersection, except the ones that can be better represented by roundabouts. A known limitation is the case of road segments entering the intersection with a lateral offset, as this offset is not available in the model. This makes it impossible to accurately represent conditions like the one in Figure 2.

Further limitations arises with less frequent, but still existing configurations, *e.g.,* containing traffic islands or other raised areas as in Figure 3. Other less frequent limitations concern

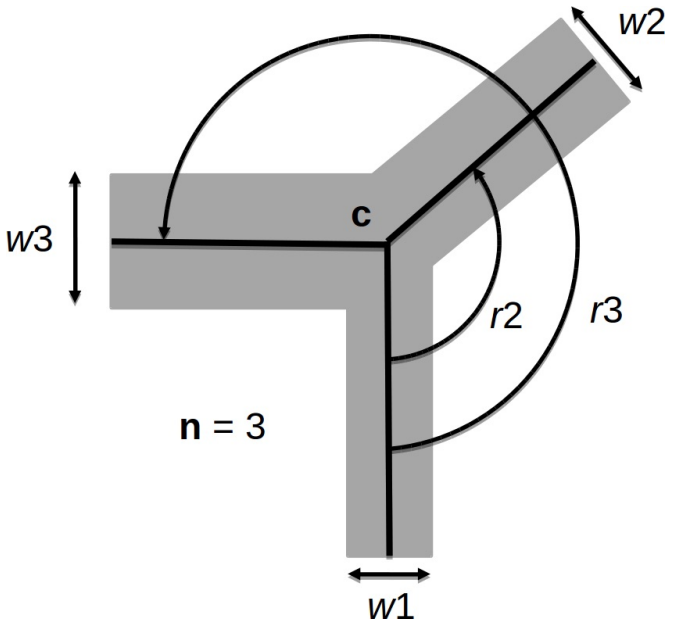



Fig. 1. The intersection model used to generate the intersections: $c$ represents the intersection center, $n$ the number of road segments, $w_i$ and $r_i$ the width and the orientation *w.r.t.* the road segment where the vehicle is.

Fig. 2. Intersections that can not be represented using the model proposed in Figure 1: arm laterally displaced *w.r.t.* one another.

intersection topologies that are available in the KITTI residential sequences, but cannot be dealt with the proposed model. As an example, see Figure 4. The frames associated with these intersections are not included in the dataset.

## III. Camera Calibration

Camera calibration has been found being the source of a significant problem. When computing the Point Cloud (PC) from the images of the colour cameras, we have been obtaining

Fig. 3. Intersections that can not be represented using the model proposed in Figure 1: traffic island or other raised areas.

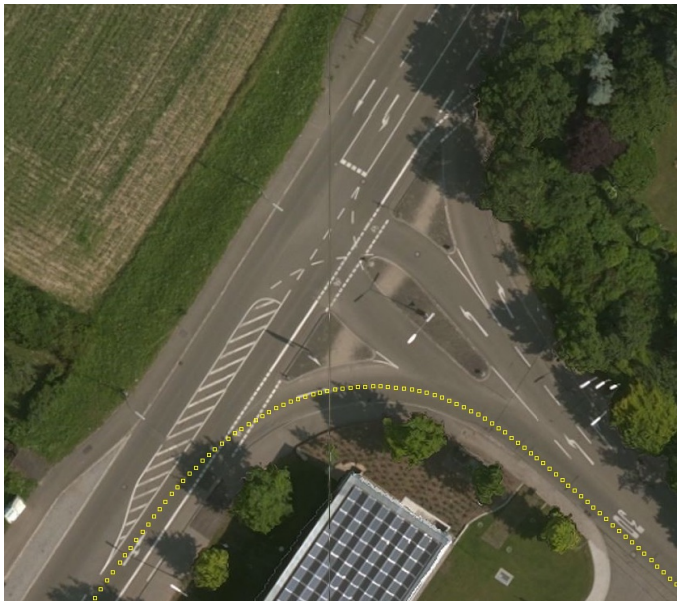


Fig. 4. The figure depicts intersections that cannot be represented using the model proposed in Figure 1: multiple laterally displaced arms plus traffic island or other raised areas; the yellow dotted line does not carry any particular meaning, please just dis-regard it.

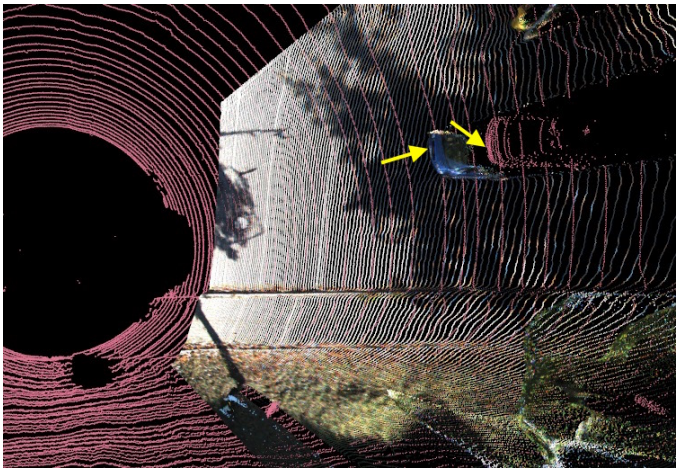


Fig. 5. Misalignment between LIDAR and camera-based PCs. The camera-based PC has been computed using the colour camera images and the projection parameters obtained for the colour cameras, while the other PC is from the LIDAR. The two has been superimposed using the poses of the two sensors (the stereo rig and the LIDAR) *w.r.t.* the vehicle. At the car bumper the misalignment is just more than 1.9m.

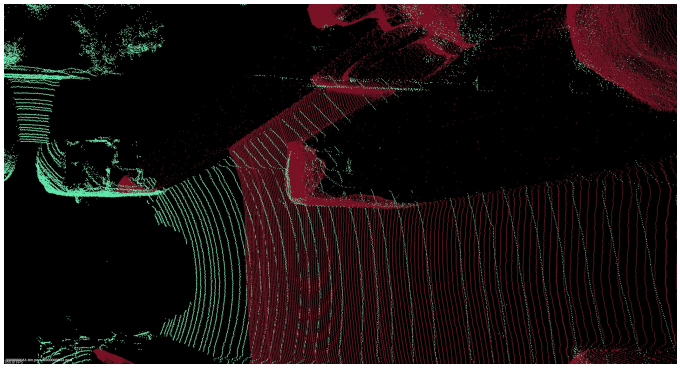


Fig. 6. Misalignment between LIDAR and camera-based PCs. The camera-based PC has been computed using the colour camera images and the projection parameters obtained for the black-and-white cameras, while the other PC is from the LIDAR. The two has been superimposed using the poses of the two sensors (the stereo rig and the LIDAR) *w.r.t.* the vehicle. Here the misalignment is less than 0.2m.

reconstructions of the intersections displaced from the correct positions. In order to find out where the problem was, we superimposed the LIDAR PC to the camera-based PC, by exploiting the poses of the two sensors *w.r.t.* the vehicle, an example can be seen in Figure 5. The figure clearly shows the different reconstruction produced by the camera processing pipeline. It can be noticed that the errors seem also to depend on the distance, so the error is not just a translation.

Thinking that a mistake could have been happened while putting online the dataset, we firstly re-computed the projection parameters of the colour cameras, using the online KITTI camera calibration tool as well as the KITTI calibration images. Unfortunately, this did not solve the problem.

We then discovered that these errors were not present for the reconstructions produced by the processing pipeline used in a previous paper of ours [6]. The work in this paper, which was dealing with the problem of recognition of the topology of the intersection, was actually computing the 3D reconstruction from the black-and-white cameras of the vehicle, by exploiting the SGBM [9] stereo approach.

As a matter of coincidence, we then discovered that the - apparently non-sensical - substitution of the projection parameters of the colour cameras with the projection parameters of the black-and-white cameras, resulted in a reconstructed PC that was more accurately aligned with the LIDAR PC, see Figure 6.

Given the time pressure, we gave up to discovering the reasons for the misalignment between the PCs, and proceeded using the projection parameters estimated for the black-and-white cameras by the KITTI team as if they were the projection parameters of the colour cameras.

## IV. OpenStreetMap Alignment

Unfortunately, we also had to deal with alignment issues of the map. OpenStreetMap has some global map alignment problem, *i.e.,* some parts of map are not aligned to each other. The problem manifests with the GT trajectory of the vehicle being apparently out of the road, *e.g.,* on curbs or inside private areas. This is a known issue and there are also literature contributions on how to increase the quality of the alignments. There are different approaches dealing with this issue, see *e.g.,* [10]–[13]. Notice that there is no clear way to determine whether the culprit is the GT or the map.

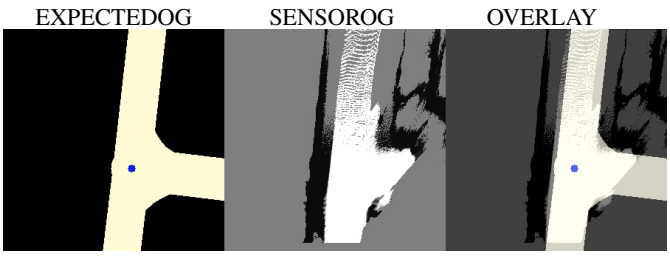
We started by superimposing the LIDAR PCs onto the aerial view, by means of the position GT, which is available in KITTI for each LIDAR PC. Of course, the misalignments between, *e.g.,* the boundaries of the buildings in the aerial view *w.r.t.* the same into the LIDAR PC cannot safely be attributed to an error in the pose GT or in the aerial map. This superimposition has been performed both with JOSM (the OpenStreetMap editor), requiring a cumbersome manual image mosaicing, and also with MAPViz, a ROS tool for superimposing layers of geo-localized data. Unfortunately, in both cases we could not obtain geometrically consistent results.

A further option is to use the capability of JOSM to set some alignment points in the aerial images, while the global position of the alignment points could be retrieved from a specific server. At this point we could modify the GT, by checking the alignment of buildings and roads. This would have been a difficult and very cumbersome task.

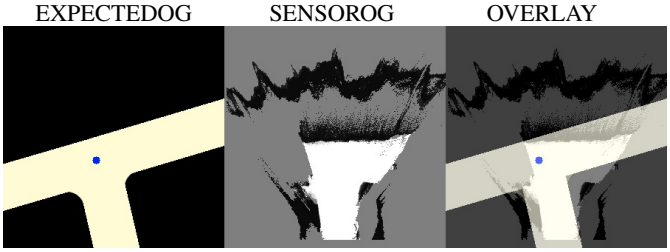
Alternatively, we could join, in a SLAM-like fashion, all aerial images, and then draw the roads and the GT position into this map (and re-entry all such data into OpenStreetMap), again a difficult and very cumbersome task.

In the end, we decided not to try to increase the quality of the maps, neither to correct the GT. Instead, by exploiting the spatial locality of the problem (*i.e.,* localization while approaching an intersection), we selected only the sequences taking place in areas where the alignment was good enough, according to the criteria described below. This has been a much simpler task. The criteria to select a sequence is as follows.

(1) We select a frame where the vehicle is very near to the intersection;




(a) Example of correct alignment




(b) Example of wrong alignment

Fig. 7. For each row, from left to right: the EXPECTEDOG created using the GPS-RTK and the digital map, the SENSORSOG and finally their overlay. A misalignment between the GPS-RTK pose (and the associated EXPECTEDOG) and the map is highlighted in the Figure 7b. It is worth to note that alignment error might come from degraded GPS-RTK measures and/or digital map misalignments.

(2) we compute the SENSOROG, *i.e.,* the occupancy grid built using the PC reconstructed from the cameras, see [6];
(3) we generate the EXPECTEDOG, *i.e.,* the occupancy grid obtained from OpenStreetMap when hypothesizing the vehicle in the GT position;
(4) we superimpose the two occupancy grids and,
(5) depending on whether the two are aligned, see Figure 7, we consider the approach to that intersection as usable or not.

Therefore, the localization performances will be benchmarked only in situations where it has been manually verified that the SENSOROGs, which are of course generated from the real positions of the vehicle, are correctly aligned *w.r.t.* the corresponding EXPECTEDOGs, computed considering the vehicle in the position GT, *i.e.,* a very good estimate of the real position of the vehicle.

In the end we obtained 48 well-aligned subsequences of approaches to different intersection geometries. The involved intersections are shown in Figure 8.

## V. Results and Discussion

The analysis of the sequences of the KITTI dataset, presenting subsequences of the vehicle approaching an intersection, is presented in Table I. In many cases the detection of the intersection has not been possible because of the lack of frames, *i.e.,* very few frames were recorded during the approach to the intersection, likely because the dataset was recorded at 10Hz. In other cases the light was too low or too much for the cameras to record usable images, or the topology of the intersection cannot be represented with the used intersection model, or the GT could not be verified against aerial imagery. All these cases has been discarded.

Fig. 8. In the figure the intersections where the 48 usable subsequences have been taken, are highlighted in yellow.

TABLE I
Results on KITTI residential sequences

| dataset name (chars ("2011_" deleted) | Lenght [mm:ss] | frames in sequence | passed intersect. | usable intersect. | issues |
|---|---|---|---|---|---|
| 09_26_drive_0019 | 00:48 | 487 | ? 0 | 0 | ? - |
| 09_26_drive_0020 | 00:09 | 92 | ? 0 | 0 | ? - |
| 09_26_drive_0022 | 01:20 | 806 | ? 3 | 0 | ? - |
| 09_26_drive_0023 | 00:48 | 480 | ? 0 | 0 | ? - |
| 09_26_drive_0035 | 00:13 | 137 | ? 1 | 1 | ? - |
| 09_26_drive_0036 | 01:20 | 809 | ? 2 | 0 | ? [1] |
| 09_26_drive_0039 | 00:40 | 401 | ? 2 | 2 | ? - |
| 09_26_drive_0046 | 00:13 | 131 | ? 1 | 1 | ? - |
| 09_26_drive_0061 | 01:10 | 709 | ? 3 | 0 | ? [1][N] |
| 09_26_drive_0064 | 00:57 | 576 | ? 4 | 4 | ? [N] |
| 09_26_drive_0079 | 00:10 | 107 | ? 0 | 0 | ? - |
| 09_26_drive_0086 | 01:11 | 711 | ? 0 | 0 | ? - |
| 09_26_drive_0087 | 01:13 | 735 | ? 1 | 0 | ? - |
| 09_30_drive_0018 | 04:36 | 2768 | ? 17 | 14 | ? - |
| 09_30_drive_0020 | 01:51 | 1111 | ? 6 | 0 | ? [1][N] |
| 09_30_drive_0027 | 01:51 | 1112 | ? 7 | 5 | ? [1] |
| 09_30_drive_0028 | 08:38 | 5183 | ? 34 | 0 | ? [1][2][3][N] |
| 09_30_drive_0033 | 02:40 | 1600 | ? | 0 | ? [1][N] |
| 09_30_drive_0034 | 02:03 | 1230 | ? 3 | 0 | ? - |
| 10_03_drive_0027 | 07:35 | 4550 | ? 41 | 12 | ? [1] |
| 10_03_drive_0034 | 07:46 | 4669 | ? | 9 | ? - |

[1] Intersection topology cannot be represented using the proposed intersection model
[2] Clear GT misalignments.
[3] Misalignment OSM *w.r.t.* GT or vice-versa.
[N] Missing aerial image for checking the position GT

## Conclusions

We analyzed many publicly available datasets in order to build a dataset of sequences for benchmarking methods for vision-based localization at intersection. The sequences have to include streams from a colour stereo rig, altogether with an accurate position estimate to be used as position GT.

We discarded all datasets, apart the residential sequences from the KITTI dataset. The analysis of these sequences led to the selection of 48 subsequences of the vehicle approaching an intersection, usable for the mentioned benchmarking task.



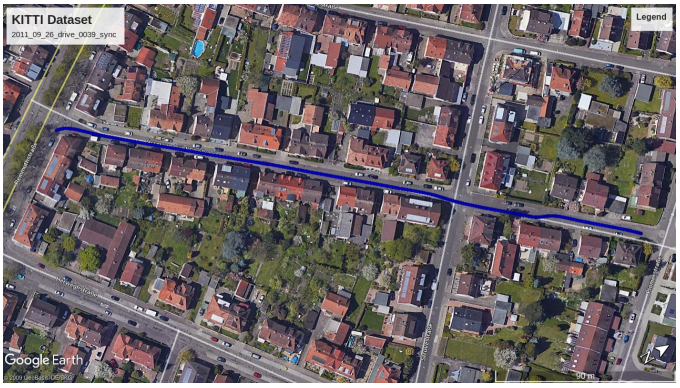


Even though this dataset is a valuable tool, it does not include all types of intersections, not all light conditions, and not all traffic levels. For these reasons it has to be taken as a first step toward the creation of a dataset usable for realistic benchmarking of localization at intersections. With "realistic" we mean that the coverage in terms of light conditions, intersection topology, and traffic level allow to consider a method, successful against the benchmark, to be worth trying in worldwide conditions.

A last word of appreciation goes to the KITTI team, whose work turned out to be the only usable dataset, so many years after its collection!


## Appendix: KITTI Residential Sequences

In this Appendix we just present the trajectory of the vehicle in the KITTI residential sequences, to have an overview of the intersections that are potentially available.

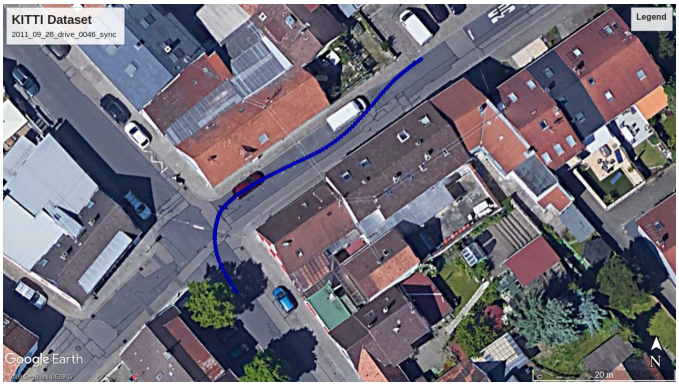


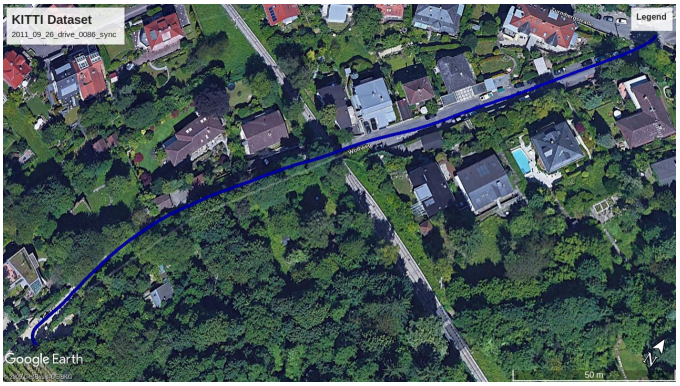

## References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[2] J.-L. Blanco, F.-A. Moreno, and J. Gonzalez-Jimenez, "The màlaga urban dataset: High-rate stereo and lidars in a realistic urban scenario," *International Journal of Robotics Research*, vol. 33, no. 2, pp. 207–214, 2014. [Online]. Available: http://www.mrpt.org/MalagaUrbanDataset

[3] 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


KITTI Dataset
2011_09_26_drive_0046_sync
Legend
Google Earth
20 m
N


KITTI Dataset
2011_09_26_drive_0086_sync
Legend
Google Earth
50 m


KITTI Dataset
2011_09_26_drive_0061_sync
Legend
Google Earth
100 m


KITTI Dataset
2011_09_26_drive_0087_sync
Legend
Google Earth
70 m

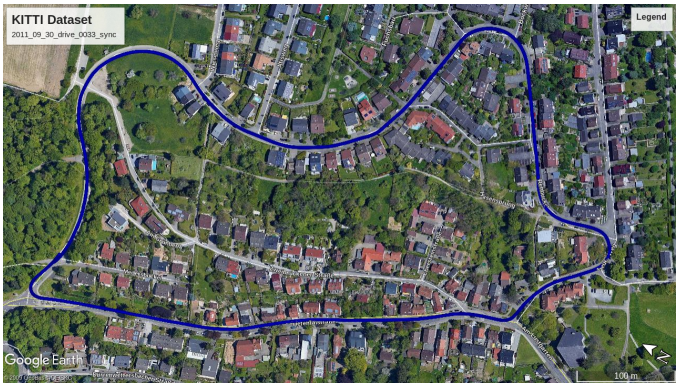

KITTI Dataset
2011_09_26_drive_0064_sync
Legend
Google Earth
100 m


KITTI Dataset
2011_09_30_drive_0018_sync
Legend
Google Earth
200 m


KITTI Dataset
2011_09_26_drive_0079_sync
Legend
Google Earth
20 m


KITTI Dataset
2011_09_30_drive_0020_sync
Legend
Google Earth
100 m








[4] G. Fontana, M. Matteucci, and D. G. Sorrenti, "Rawseeds: Building a benchmarking toolkit for autonomous robotics," in *Methods and Experimental Techniques in Computer Engineering*, ser. SpringerBriefs in Applied Sciences and Technology, F. Amigoni and V. Schiaffonati, Eds. Springer, Cham, 2014.

[5] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," *arXiv: 1803.06184*, 2018.

[6] A. L. Ballardini, D. Cattaneo, S. Fontana, and D. G. Sorrenti, "An online probabilistic road intersection detector," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017.

[7] A. Geiger, M. Lauer, and R. Urtasun, "A generative model for 3d urban scene understanding from movable platforms," in *IEEE Conf. Comp. Vis. Patt. Rec. (CVPR)*, June 2011.

[8] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3D traffic scene understanding from movable platforms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 1012–1025, 2014.

[9] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 2, 2008.

[10] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun, "Enhancing road maps by parsing aerial images around the world," in *International Conference on Computer Vision (ICCV)*, 2015.

[11] ——, "Hd maps: Fine-grained road segmentation by parsing ground and aerial images," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] G. Mattyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun, "Torontocity: Seeing the world with a million eyes," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.