Department of Informatics
Technical University of Munich

Exposé on Master's Thesis

Name: Joseph M. Birkner
Matr. Nr.: 03704462
EMail: joseph.birkner@tum.de
Supervisor: Prof. Dr.-Ing. habil. Alois C. Knoll
Institute: Chair of Robotics, Artificial Intelligence and Real-time Systems
Title: Monocular 3D Traffic Perception Using HD Maps as an Auxiliary Feature
Date: August 5, 2022

1 Topic

This work is conducted within the scope of the **Providentia++** project. The project aims to provide cars with Digital Twins in the cloud, facilitating enhanced traffic prediction and micro-routing abilities for safer autonomous driving. Under the Providentia project, Sensor packages, encompassing RGB Cameras, Radar, and Lidar, are installed along major roads and intersections. We are evaluating, which minimal Sensor-suite would be suitable for a wider rollout of the sensing infrastructure under minimum cost.

The "Digital Twin" data-point of a vehicle consists of multiple components: First and foremost, position $(x|y|z)$, size $(w|h|d)$ and orientation (θ) are key variables which define a vehicle's spatial state. Since the RGB camera sensor is the cheapest among the installed package, 3D object detection from 2D video is an important area of research within Providentia.

A working initial approach for monocular 3D detection for Providentia++ was developed by Leon Blumenthal [Blu] in the scope of their Bachelor's thesis, using 3D projection of the lower edge of 2D vehicle segments. This approach works very well for vehicles which are moving in straight lines, as the orientation can be fixed to a constant value. However, more work needs to be done for reliable monocular detection when observing traffic scenes with heavily varying vehicle orientations, especially scenes such as complex intersections.

The goal of this work is to improve the orientation estimation of turning vehicles by exploiting clues about their heading from HD maps of the observed road scene. Both map-matching and heading estimation will also benefit from considering the prior trajectory of a vehicle. The trajectory can be obtained by chaining observations of an individual vehicle across multiple video frames, thereby introducing a time component to the observation. Once bounding box estimates between frames are related through vehicle identities, it will also become possible to stabilize predictions about their spatial state via Kalmann Filters, Recurrent Neural Networks or other methods.

2 Approach

2.1 Architecture

Considering prior work, we have identified several ways in which we amy incorporate map data into the 3D detection process. Within the taxonomy provided by [Ma+22], we are mainly looking at *Result-based Lifting Methods*. Within such approaches, the detection pipeline is divided into a 2D detection and a 3D lifting stage. The 2D detection stage is facilitated by a 2D instance segmentation model

such as YoloV5 or YoloActEdge. The 3D lifting stage can then focus on 3D keypoint estimation for each detected instance, which is a much narrower task than full end-end monocular 3D detection.

While there are many recent papers on the topic, we have picked two approaches among recent work which we are going to consider for implementation and further development in this thesis: TrafficNet [RAM21] and UrbanNet [CW21]. For TrafficNet, HD map data could directly be used to seed the initial heading estimate. In the paper, this is currently done by matching the vehicle position to the nearest road boundary detected from sattelite imagery (see fig. 1). TrafficNet also provides a very fine-tuned rulebook for smoothing 3D detections for individual vehicles over time.

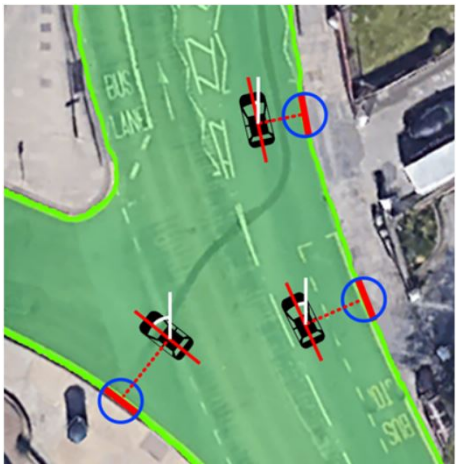

Figure 1: Figure 12 regarding Angle Estimation from Traffic-Net [RAM21]

The drawback of TrafficNet is that the bounding box height and length predictions are based on hard-coded prior assumptions for their detected object categories. This is solved a bit more elegantly in UrbanNet, which employs a convolutional neural network to detect 3D keypoints for vehicles within their 2D image bounding box (check out fig. 2). See fig. 3 for an overview of the UrbanNet architecture.

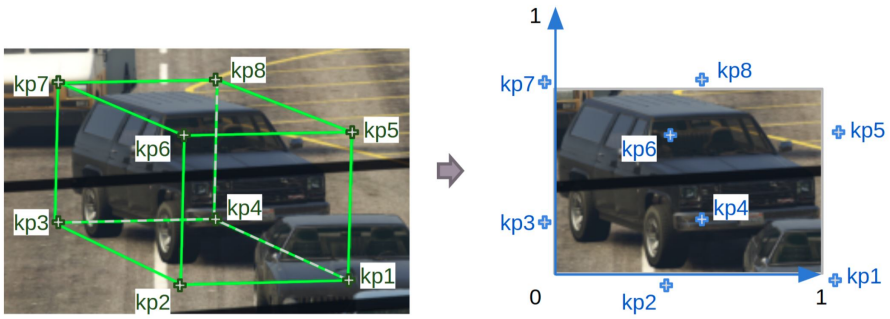

Figure 2: Figure 4 regarding Keypoint Estimation from UrbanNet [CW21]

For our system design, we are looking to combine the best of these two approaches: Strong temporally informed, map-guided heading estimation, with dynamic neural keypoint extraction.

2.2 Map Format

We have chosen Lanetlet2 [Pog+18] as the input map format for our system, as the underlying model is very focused on lanes. This approach is well-suited to inform heading estimation and map-matching of vehicle trajectories.

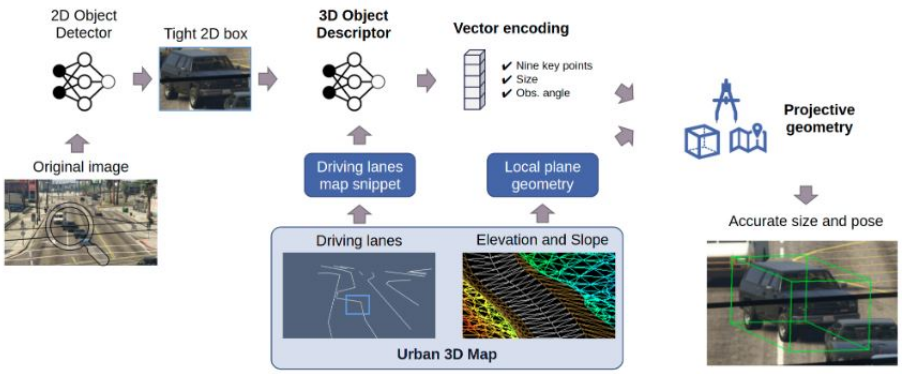

Figure 3: Figure 2 regarding Architecture from UrbanNet [CW21]

2.3 Datasets

Labeled Datasets for the monocular 3D detection task are required for training of supervised machine-learning based components and evaluation of the final system. We are considering to use the following datasets for Training and Evaluation:

- The Providentia A9 dataset [Cre+22]
- The Roadside Perception (ROPE) 3D dataset [Ye+22]
- The KITTI dataset [GLU12]
- The nuScenes dataset [Cae+20]

3 Work Plan and necessary Resources

My work is divided into Research, Development and Evaluation, and Analysis.

4 Literature

[Blu] Blumenthal, L. “Real-Time Monocular 3D Object Detection to Support Autonomous Driving”.

[Cae+20] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. “nuscenes: A multimodal dataset for autonomous driving”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11621–11631.

[CW21] Carrillo, J. and Waslander, S. “Urbannet: Leveraging urban maps for long range 3d object detection”. In: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE. 2021, pp. 3799–3806.

[Che+21] Chen, H., Huang, Y., Tian, W., Gao, Z., and Xiong, L. “Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10379–10388.

[Cre+22] Creß, C., Zimmer, W., Strand, L., Fortkord, M., Dai, S., Lakshminarasimhan, V., and Knoll, A. “A9-dataset: Multi-sensor infrastructure-based dataset for mobility research”. In: *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2022, pp. 965–970.

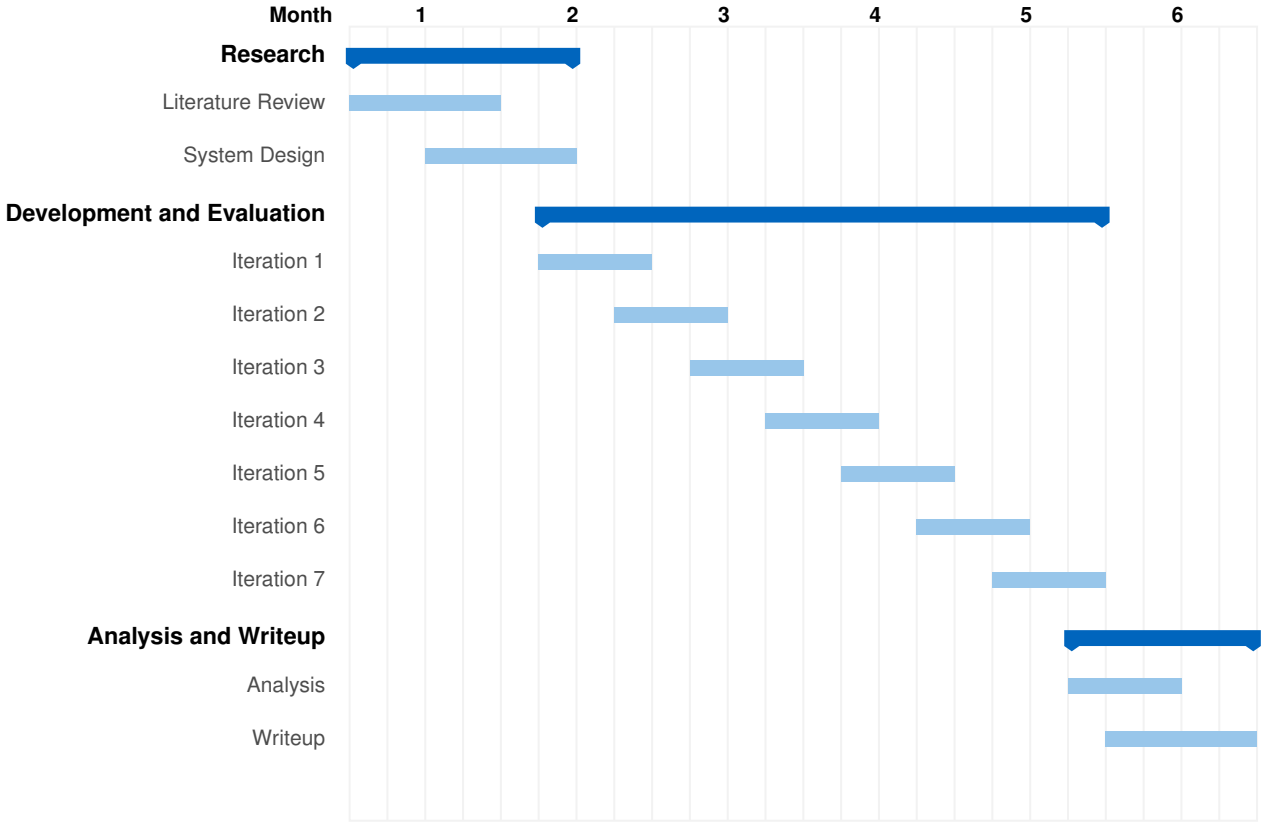

Figure 4: Time schedule.

[GLU12] Geiger, A., Lenz, P., and Urtasun, R. “Are we ready for autonomous driving? the kitti vision benchmark suite”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3354–3361.

[GMJ19] Gkioxari, G., Malik, J., and Johnson, J. “Mesh r-cnn”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 9785–9795.

[Guo+21] Guo, E., Chen, Z., Rahardja, S., and Yang, J. “3D Detection and Pose Estimation of Vehicle in Cooperative Vehicle Infrastructure System”. In: *IEEE Sensors Journal* 21.19 (2021), pp. 21759–21771.

[Hey+21] Heylen, J., De Wolf, M., Dawagne, B., Proesmans, M., Van Gool, L., Abbeloos, W., Abdelkawy, H., and Reino, D. O. “MonoCInIS: Camera Independent Monocular 3D Object Detection using Instance Segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 923–934.

[KLR18] Kundu, A., Li, Y., and Rehg, J. M. “3d-rcnn: Instance-level 3d object reconstruction via render-and-compare”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3559–3568.

[LJ22] Li, P. and Jin, J. “Time3D: End-to-End Joint Monocular 3D Object Detection and Tracking for Autonomous Driving”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3885–3894.

[Ma+22] Ma, X., Ouyang, W., Simonelli, A., and Ricci, E. “3d object detection from images for autonomous driving: a survey”. In: *arXiv preprint arXiv:2202.02980* (2022).

[MPVG22] Marinello, N., Proesmans, M., and Van Gool, L. “TripletTrack: 3D Object Tracking Using Triplet Embeddings and LSTM”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4500–4510.

[Mou+22] Mouawad, I., Brasch, N., Manhardt, F., Tombari, F., and Odone, F. “Time-to-Label: Temporal Consistency for Self-Supervised Monocular 3D Object Detection”. In: *arXiv preprint arXiv:2203.02193* (2022).

[Pog+18] Poggenhans, F., Pauls, J.-H., Janosovits, J., Orf, S., Naumann, M., Kuhnt, F., and Mayr, M. “Lanelet2: A high-definition map framework for the future of automated driving”. In: *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE. 2018, pp. 1672–1679.

[RAM21] Rezaei, M., Azarmi, M., and Mir, F. M. P. “Traffic-Net: 3D Traffic Monitoring Using a Single Camera”. In: *arXiv preprint arXiv:2109.09165* (2021).

[Rüe+22] Rüegg, N., Zuffi, S., Schindler, K., and Black, M. J. “BARC: Learning to Regress 3D Dog Shape from Images by Exploiting Breed Information”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3876–3884.

[Ye+22] Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., and Ding, E. “Rope3D: The Roadside Perception Dataset for Autonomous Driving and Monocular 3D Object Detection Task”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 21341–21350.

[Yua+21] Yuan, W., Lv, Z., Schmidt, T., and Lovegrove, S. “Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 13144–13152.