Robust 3D Object Detection for Autonomous Vehicles using Sensor Fusion
Mid-Term Report

Mohan Rao Divate Kodandarama
University of Wisconsin-Madison
Madison, WI
divatekodand@wisc.edu

Aditya Rungta
University of Wisconsin-Madison
Madison, WI
arungta@wisc.edu

1. Problem Definition

In our work, we intend to explore the viability of augmenting the point cloud with pseudo-LiDAR data generated form monocular images. More formally, the problem is to predict accurate depth map from monocular images and project the points in 3d to augment the point cloud.

[11] utilizes a similar approch to augment the point cloud with Pseudo-LiDAR points. However, they make use of stereo images to generate Pseudo-LiDAR data.

2. Approach

Our approach for 3D Object Detection is outlined below.

1. **Estimate per pixel depth from monocular images -** We utilize [4] to estimate accurate per pixel depth from monocular images. Since it is very difficult to obtain ground truth data for depth of every pixel, this paper uses self-supervised learning by framing the problem as the minimization of a photometric re-projection error at training time.

2. **Project all images pixel to 3D space (LiDAR Coordinates)-** Once the per pixel depth estimate is available, all the images pixels can be projected into 3D space using the camera intrinsic matrix (available as a part of calibration data in [3]). More formally, the image pixel at location (i, j) with the estimated depth $D^{i,j}$ can be projected into 3D according to equation 1.

$$Q^{i,j} = D^{i,j} K^{-1} [i, j, 1]^T \qquad (1)$$

3. **Alignment and Filtering -** Unlike stereo, Pseudo-LiDAR generated by monocular images would not be aligned with the LiDAR (This is evident in Figure 4). Hence, we estimate the scale of the generated Pseudo-LiDAR data with respect to the point cloud by estimating the centroid of both Pseudo-LiDAR and the point cloud. We are also exploring the use of Iterative closes point (ICP) algorithm for this scenerio. After the alignment stage we filter out all the Pseudo-LiDAR points which are far from the original point cloud.

4. **Detection -** Once the augmented point cloud is available, use a approach similar to [6] for detecting 3D objects. However, instead of PointNet based backbone network, we use a backbone network based on [12] (Graph Convolution based feature extractor) to extract point cloud features for each Pillar.

2.1. Datasets

We are using the KITTI dataset for all our experiments.

3. Progress

Task	Status
Setup Monocular depth estimation Pipeline	Done
Generate Pseudo-LiDAR data from the estimated depth	Done
Alignment and Filtering of Pseudo-LiDAR	In Progress
Design of Graph Conv based feature extractor	Done
Detection Pipeline	In Progress

Figure 1. Progress

3.1. Challenges

Since the depth map estimated from the monocular images is only an approximation of the true depth map, the generated Pseudo-LiDAR data needs to be aligned with the point cloud. Further, Pseudo-LiDAR away from the point cloud need to be filtered out. Since, these operations are in the inference path, they have to be performed efficiently.

4. Results

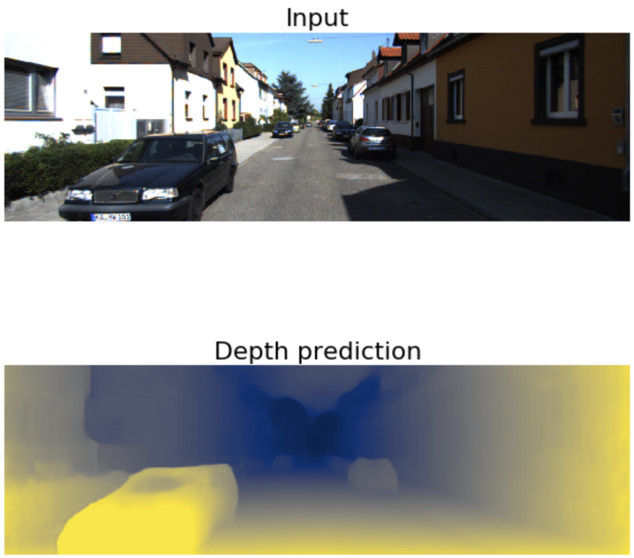

Figure 2. Estimated depth map for an image from the KITTI dataset

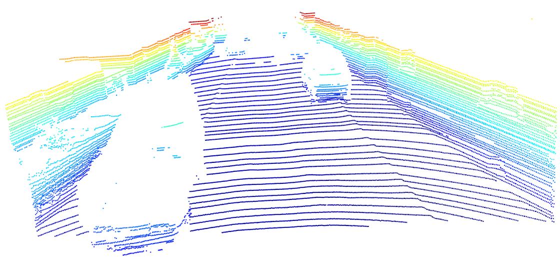

Figure 3. LIDAR point cloud corresponding to the above image

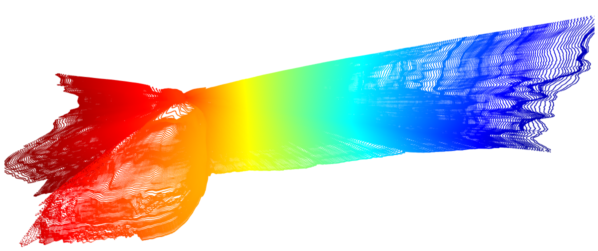

Figure 4. Pseudo-LiDAR generated by projecting the RGBD (Estimated Depth) in 3D

Figure 2 show the estimated depth for an image from KITTI training split. Figure 4 shows the generated Pseudo-LiDAR data. It is evident that the generated Pseudo-LiDAR data does not perfectly align with the point cloud data and requires alignment and filtering.

References

[1] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.

[3] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[4] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018.

[5] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Lake Waslander. Joint 3d proposal generation and object detection from view aggregation. *CoRR*, abs/1712.02294, 2017.

[6] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *CoRR*, abs/1812.05784, 2018.

[7] Gregory P. Meyer, Ankit Laddha, Eric Kee, Carlos Vallespi-Gonzalez, and Carl K. Wellington. Lasernet: An efficient probabilistic 3d object detector for autonomous driving. *CoRR*, abs/1903.08701, 2019.

[8] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016.

[9] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017.

[10] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. *CoRR*, abs/1812.04244, 2018.

[11] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. *CoRR*, abs/1812.07179, 2018.

[12] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *CoRR*, abs/1801.07829, 2018.

[13] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection, 2017.