VR3Dense: 3D Object Detection and Dense Depth Reconstruction from Voxel-Representation

Shubham Shrivastava

shubhams@stanford.edu

1 Introduction

3D Object Detection is s one of the most significant part in autonomous vehicle perception. An autonomous vehicle needs to be aware of its surrounding objects and should be capable of predicting their future trajectory. Most autonomous vehicles in their development phase today are occupied with LiDARs, Cameras, and RADARs which allow them to perceive the environment. While 2D object detection methods using cameras ([1], [2], [3], [4]) have matured quite a bit, it does not really capture the accurate geometry of the scene and hence does not help an autonomous vehicle plan its actions. Few recent papers have also proposed the use of monocular camera to detect and localize object 3D poses ([5], [6], [7], [8]), however, these have proven to be inferior to LiDAR base object detection methods. LiDARs have been one of the most crucial sensors for obtaining accurate 3D scene representation, however, its high-cost and sparsity in the point-cloud data have discouraged a small set of automotive companies from utilizing it for perception tasks. Although expensive, it is irrefutable that LiDAR point-clouds provides the means of obtaining most accurate 3D scene representation.

Several LiDAR-based 3D object detection methods ([9], [10], [11], [11]), have proven to perform extremely well and achieve state-of-the-art results on existing dataset benchmarks. Researchers utilize various representations of point-clouds such as bird-eye view (BEV), voxels, and stixels - each with its own advantages and disadvantages. While methods such as Li et. al.[12], and PIXOR[9] projects LiDAR point-cloud on a 2D plane and then use 2D CNN for object detection, other methods such as Vote3deep[13] and Voxelnet[14] represents point-cloud as a set of 3D voxels and then use 3D convolution to extract features. These methods often employ separate heads for objectess probability prediction and pose prediction. Another stream of research such as VoxNet[10], converts a point-cloud segment into a 3D volumetric grid and then perform detection. Recently, PointNet [11] introduced a different approach to point-representation which eliminates the need to manually structure the points in a predefined manner and was further improved by PointNet++[15].

In this work, we will explore 3D object detection from LiDAR point-clouds by first representing them as voxels and then using a 3D CNN to regress object pose parameters. We believe that the there are a lot of information that can be extracted from structured point-cloud representation. Furthermore, other sensors like cameras can also benefit from this structured representation and compliment each other, e.g. LiDAR points feature extracted using 3D convolution layers can be transformed and fused with camera features extracted using 2D convolutions. This work will be focused towards building a complete 3D object detection pipeline from scratch. In addition, we will explore fusion of LiDAR and Camera data for dense depth map reconstruction. Qualitative and quantitative analysis will be performed on the KITTI dataset [16].

2 Technical Approach

3D object detection from voxelized point-cloud can be done by first extracting 3D features, and then regressing the object pose parameters. Encoding the entire point-cloud data within a certain region-of-interest into a set of structured voxels will allow easier extraction of spatial features. A function approximator can then be taught to predict object existence along with their most probable position and orientation within a localized volume. This can be achieved by dividing the voxelized region-of-interest into a set of volumes, each tasked with detecting an object and their corresponding pose (x, y, z, l, w, h, yaw). Since, a prior knowledge

of the environment is known (e.g. cars mostly drive on the road plane), we assume the roll and pitch to be zero. Certain loss functions have such as GIoU [17] have recently shown to consistently improve the object detection performance - and is a subject of exploration in this work.

LiDAR data is sparse, and this results in a very few points representing far objects. Injecting other information about the scene, such as an RGB image, into the network might help the network infer missing depth data. To this end, we propose to use a separate network to extract 2D image features from the corresponding RGB image and fuse this with LiDAR in order to obtain a richer scene representation. Furthermore, this work proposes to use an encoder-decoder architecture to reconstruct dense depth map from just RGB images, with some supervision from the sparse point-clouds. This would encode a denser rgb-to-depth embeddings in the auto-encoder latent space, which can further be combined with the fully-connected layer of LiDAR based object detection branch. Both LiDAR and RGB have their own advantages and this way of combining data could influence overall task of perception system within autonomous vehicles.

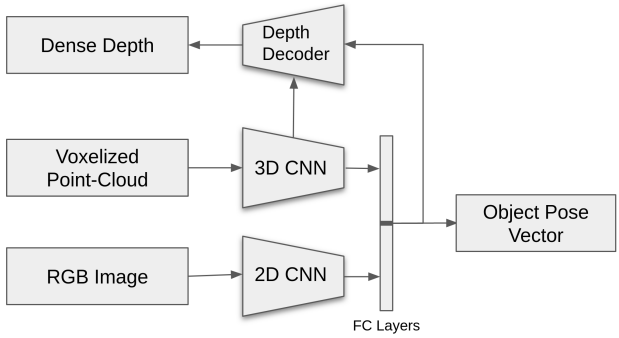

Figure 1: High-level architecture of VR3Dense.

3 Milestones

Some of the milestones for this project are listed below:

- [$Week 1$] Build a set of LiDAR point-cloud utility functions including *io*, *voxelization*, *visualization*, *etc.*
- [$Week\ 2-4$] Build a model (including loss functions) using 3D CNN to extract 3D features and regress pose parameters. Also, build the framework using PyTorch to allow easy data-loading, training, testing, evaluation, data-logging, checkpointing, etc.
- [$Week\ 4-5$] Evaluate 3D object detection quantitative performance on KITTI 3D object test dataset [16]. Additionally, perform qualitative analysis on KITTI raw dataset [18] with continuous frames; use a multi-object tracker (AB3DMOT [19]) for tracking of each object and smooth transition through frames.
- [$Week\ 5-6$] Implement RGB-to-Depth encoder-decoder architecture with feature-level fusion from the object detection branch and supervision using LiDAR point-cloud projected onto the 2D plane.

3.1 Dataset and Evaluation

We plan to use KITTI object *training* dataset [16] to train both 3D object detection as well as dense depth reconstruction. For quantitative analysis of 3D object detection, we will be using the *val split* and *testing* dataset - this will be done using KITTI object detection evaluation kit. Dense depth reconstruction can be quantitatively analyzed by computing metrics such as $AbsRel$, $SqRel$, $RMSE$, $RMSE\ log$ $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$ on the pixels for which we have a corresponding LiDAR point available.

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," *Lecture Notes in Computer Science*, p. 21–37, 2016. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46448-0_2

[3] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," 2020.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020.

[5] S. Shrivastava and P. Chakravarty, "Cubifae-3d: Monocular camera space cubification for auto-encoder based 3d object detection," 2021.

[6] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9287–9296.

[7] Z. Qin, J. Wang, and Y. Lu, "Monogrnet: A geometric reasoning network for monocular 3d object localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8851–8858.

[8] T. He and S. Soatto, "Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8409–8416.

[9] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," 2019.

[10] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928.

[11] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," 2017.

[12] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," 2016.

[13] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," 2017.

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

[15] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," 2017.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[17] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," 2019.

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

[19] X. Weng, J. Wang, D. Held, and K. Kitani, "Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics," August 2020.