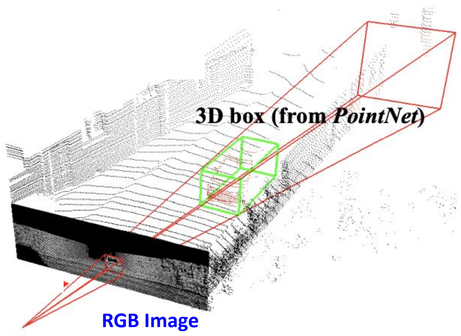
Replacing LIDAR: 3D Object Detection using Monodepth Estimation

Benjamin Goeing, Lars Jebe

Motivation

- **LIDAR** is very expensive, but currently irreplaceable for many autonomous driving and robotics applications
- The Goal of this project is to perform **3D Object Detection** (3D bounding box prediction and classification) from **a single RGB image:**



- This is a fundamentally **ill-posed problem** which can potentially be solved for homogenous datasets with strong priors

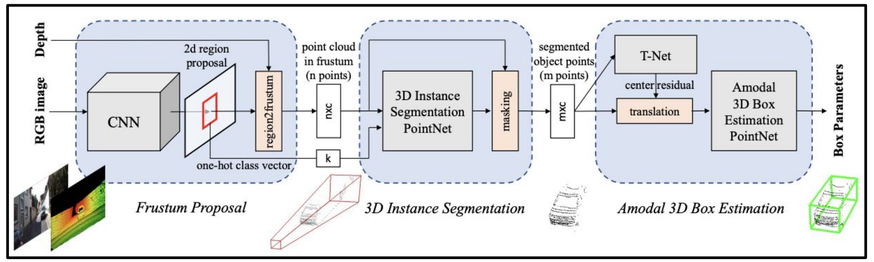
Related Work

- We use the Frustum Point Net [1] which makes a 3D bounding box prediction using an RGB image in combination with LIDAR data
- We replace the LIDAR data with unsupervised depth estimations from a single RGB image [2]
- Current results show that 3D object detection relies on LIDAR (up to 65% accuracy) whereas without LIDAR the best performance is < 8%

New Technique

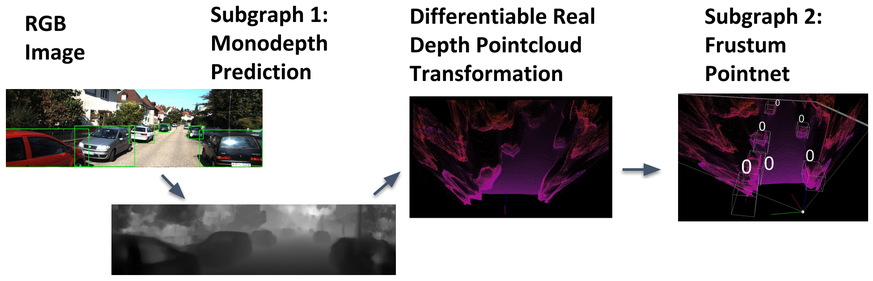
- The existing architecture relies on both a depth and RGB input in order to produce a 3D bounding box, as detailed in the figure below:

Existing Approach[1] :



Our Approach:

- We suggest to replace the depth data with an RGB monodepth prediction . The following diagram represents our approach:



Two Steps of Implementation

1. **Re-training of frustum pointnet (Subgraph 2)** — **Completed**
   - This steps involves retraining the model using the monodepth predictions from subgraph (1), which runs inference only
2. **End-to-end training** — **Work in Progress**
   - This step involves the end to end training of both graphs to achieve even higher performance

Experimental Results

- We evaluated our model using the KITTI Dataset, which contains a variety of street-view images

Baseline Performance (Using LIDAR)

- eval segmentation accuracy: 0.904
- eval segmentation avg class acc: 0.907
- eval box IoU (ground/3D): 0.743 / 0.696
- eval box estimation accuracy (IoU=0.7): 0.629

Step 1 Performance:

- eval segmentation accuracy: **0.715**
- eval segmentation avg class acc: **0.756**
- eval box IoU (ground/3D): **0.253 / 0.230**
- eval box estimation accuracy (IoU=0.7): **0.035**

Discussion

- Given the ill-posed nature of the problem, our approach shows promising results, but falls short of reaching near-Lidar quality
- One particularly noteworthy fact is that the distribution of points is **often correct on a relative basis**, but incorrect on an absolute one (e.g. object is shifted by a certain amount)

Next steps:

- Train subgraph (1) and (2) end-to-end, with a combined loss function for depth estimation and object detection

References

1. Qi, Charles R., et al. "Frustum pointnets for 3d object detection from rgb-d data." *CVPR* 2018.
2. Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." *CVPR* 2017.
3. Chen, Xiaozhi, et al. "Monocular 3d object detection for autonomous driving." *CVPR* 2016.
4. Zhou, Yin, and Oncel Tuzel. "Voxelnet: End-to-end learning for point cloud based 3d object detection." *CVPR* 2018.