Frustum PointNets for 3D Object Detection from RGB-D Data

CVPR 2018

Summary

In this paper, Frustum PointNets are presented as a framework for 3D object detection based on region proposals from image data. By lifting 2D region proposals to 3D space, frustum point clouds can be extracted, reducing considerably the 3D search space. PointNets are leveraged to segment the proposal point cloud and predict accurate 3D bounding boxes. Frustum PointNets achieve state-of-the-art performance on the KITTI and SUN RGB-D 3D Object Detection Benchmarks, showing their potential as a generic framework for both, indoor and outdoor datasets.

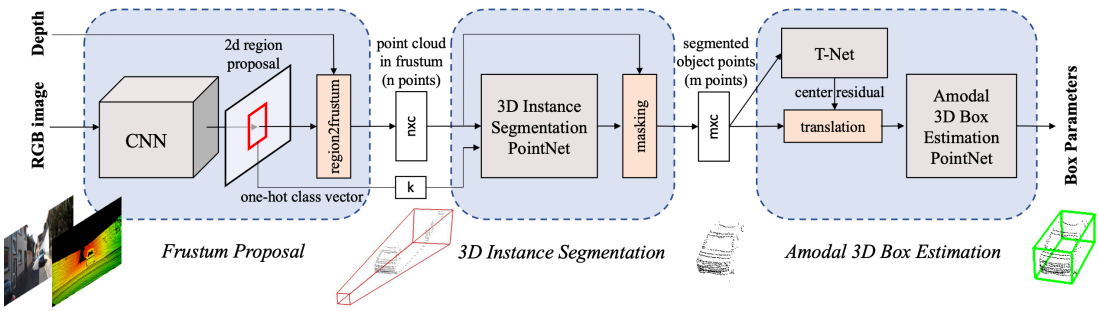

Figure 1: Illustration of the Frustum PointNet architecture. The pipeline can be divided into three stages: 1) 2D Proposal Generation and Semantic Classification, 2) Binary 3D Instance Segmentation to extract object points, 3) Amodal 3D Box Estimation.

Main Contributions

- **Novel 3D Object Detection Framework** The frustum framework is designed to leverage the density advantages of image data over LiDAR measurements as well as the maturity of 2D object detectors for accurate 3D object detection. It can be seen as a generic framework in the sense that any architecture can be chosen for both, the 2D proposal generation as well as the 3D instance segmentation task.
- **Frustum Proposal Generation** Given a camera projection matrix, any 2D RoI can be lifted to a frustum which defines the 3D search space for the respective object (see Figure 3). All LiDAR points inside a frustum, referred to as a frustum point cloud, are used for further processing to predict a 3D bounding box from the 2D proposal. The authors propose to rotate each frustum to be orthogonal to the image plane in order to increase rotation invariance of the algorithm (see Figure 2).
- **Amodal 3D Box Estimation** Given the predicted object points, the final 3D bounding box is regressed in a two-step process. First, a small spatial transformation network, called T-Net, is used to predict a center-offset from the coordinate system of the object points. This step is motivated by the fact that the observed points are located on the object surface and thus their mean doesn't necessarily correspond to the object center. In a second step, the bounding box is estimated using a combination of classification and regression for the heading and size prediction based on size templates and equally spaced heading bins.

Implementation Details

- **3D Instance Segmentation** Inspired by Mask-RCNN in the 2D domain, binary segmentation is performed on all points inside a frustum proposal to find points which belong to the respective object. To incorporate semantic information into the instance segmentation step, the class prediction of the 2D proposal network is concatenated with the point features. After removing background points, all object points are transformed to a local coordinate system by subtracting their centroid in order to be able to formulate the bounding box regression as a residual problem (see Figure 2).
- **Corner Loss for Joint Optimization of Box Parameters** The authors argue that having separate loss functions for center, size and heading prediction of the bounding boxes leads to suboptimal performance w.r.t to 3D IoU. They propose to add a regularization loss, referred to as corner loss, to promote consistency in the bounding box estimates. This loss computes the sum of distances of all corner points of the predicted bounding box to the ground truth box and is added to the multi-task loss as an additional component.

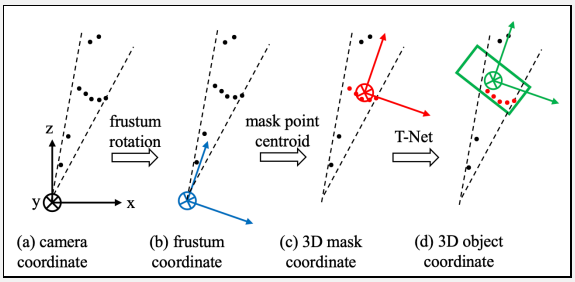

Figure 2: Illustration of the different coordinate transformations in the detection pipeline.

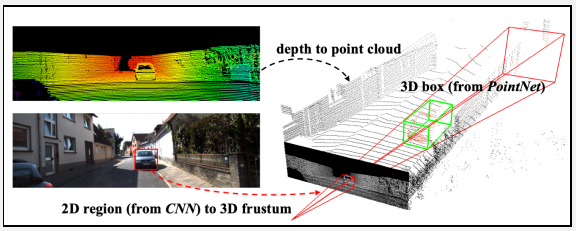

Figure 3: Illustration of the frustum-based 3D object detection concept.

Evaluation

- **Ablation Study** A number of ablation experiments are presented that illustrate the contribution of several design choices. For example, various transformations in the detection pipeline as well as the box estimation loss functions are shown to increase the accuracy in isolation as well as in combination.
- **KITTI 3D and BEV Object Detection** Frustum PointNets are evaluated on the KITTI Benchmark against several prior arts, with MV3D being the most relevant method. Frustum PointNets outperform MV3D by large margins on the task of 3D object detection on all three levels of difficulty (easy, moderate, hard). Increases in AP of more than 8% can be observed. For BEV object detection the improvements are less pronounced. For the moderate category an increase of 6% AP can be achieved. However, no increase can be observed for hard examples.
- **SUN RGB-D 3D Object Detection Benchmark** Usually, 3D object detection algorithms are evaluated either on indoor or outdoor datasets since design choices lead to superior performance in one or the other domain. Frustum PointNets show to work well on indoor datasets as well, outperforming the state-of-the-art on SUN RGB-D by considerable margins on the majority of object classes.

References

This summary is solely based on my understanding of the original paper. All images used here are taken from the original paper as well. The paper can be found under the following link:
`https://arxiv.org/pdf/1711.08488.pdf`