Deep Hough Voting for 3D Object Detection in Point Clouds

Charles R. Qi[1] Or Litany[1] Kaiming He[1] Leonidas J. Guibas[1,2]
[1]Facebook AI Research [2]Stanford University

Abstract

Current 3D object detection methods are heavily influenced by 2D detectors. In order to leverage architectures in 2D detectors, they often convert 3D point clouds to regular grids (i.e., to voxel grids or to bird's eye view images), or rely on detection in 2D images to propose 3D boxes. Few works have attempted to directly detect objects in point clouds. In this work, we return to first principles to construct a 3D detection pipeline for point cloud data and as generic as possible. However, due to the sparse nature of the data – samples from 2D manifolds in 3D space – we face a major challenge when directly predicting bounding box parameters from scene points: a 3D object centroid can be far from any surface point thus hard to regress accurately in one step. To address the challenge, we propose VoteNet, an end-to-end 3D object detection network based on a synergy of deep point set networks and Hough voting. Our model achieves state-of-the-art 3D detection on two large datasets of real 3D scans, ScanNet and SUN RGB-D with a simple design, compact model size and high efficiency. Remarkably, VoteNet outperforms previous methods by using purely geometric information without relying on color images.

1. Introduction

The goal of 3D object detection is to localize and recognize objects in a 3D scene. More specifically, in this work, we aim to estimate oriented 3D bounding boxes as well as semantic classes of objects from point clouds.

Compared to images, 3D point clouds provide accurate geometry and robustness to illumination changes. On the other hand, point clouds are irregular. thus typical CNNs are not well suited to process them directly.

To avoid processing irregular point clouds, current 3D detection methods heavily rely on 2D-based detectors in various aspects. For example, [42, 12] extend 2D detection frameworks such as the Faster/Mask R-CNN [37, 11] to 3D. They voxelize the irregular point clouds to regular 3D grids and apply 3D CNN detectors, which fails to leverage sparsity in the data and suffer from high computation cost due to expensive 3D convolutions. Alternatively, [4, 55] project points to regular 2D bird's eye view images and then apply 2D detectors to localize objects. This, however, sacrifices geometric details which may be critical in cluttered indoor environments. More recently, [20, 34] proposed a cascaded two-step pipeline by firstly detecting objects in front-view images and then localizing objects in frustum point clouds extruded from the 2D boxes, which however is strictly dependent on the 2D detector and will miss an object entirely if it is not detected in 2D.

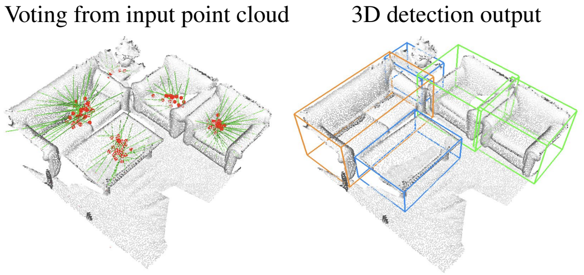

Figure 1. **3D object detection in point clouds with a deep Hough voting model.** Given a point cloud of a 3D scene, our VoteNet votes to object centers and then groups and aggregates the votes to predict 3D bounding boxes and semantic classes of objects.

In this work we introduce a *point cloud focused* 3D detection framework that directly processes raw data and does not depend on any 2D detectors neither in architecture nor in object proposal. Our detection network, *VoteNet*, is based on recent advances in 3D deep learning models for point clouds, and is inspired by the generalized Hough voting process for object detection [23].

We leverage PointNet++ [36], a hierarchical deep network for point cloud learning, to mitigates the need to convert point clouds to regular structures. By directly processing point clouds not only do we avoid information loss by a quantization process, but we also take advantage of the sparsity in point clouds by only computing on sensed points.

While PointNet++ has shown success in object classification and semantic segmentation [36], few research study how to detect 3D objects in point clouds with such architectures. A naïve solution would be to follow common practice in 2D detectors and perform dense object proposal [29, 37],