ImVoteNet: Boosting 3D Object Detection in Point Clouds with Image Votes

Charles R. Qi*† Xinlei Chen*[1] Or Litany[1,2] Leonidas J. Guibas[1,2]
[1]Facebook AI [2]Stanford University

Abstract

3D object detection has seen quick progress thanks to advances in deep learning on point clouds. A few recent works have even shown state-of-the-art performance with just point clouds input (e.g. VOTENET). However, point cloud data have inherent limitations. They are sparse, lack color information and often suffer from sensor noise. Images, on the other hand, have high resolution and rich texture. Thus they can complement the 3D geometry provided by point clouds. Yet how to effectively use image information to assist point cloud based detection is still an open question. In this work, we build on top of VOTENET and propose a 3D detection architecture called IMVOTENET specialized for RGB-D scenes. IMVOTENET is based on fusing 2D votes in images and 3D votes in point clouds. Compared to prior work on multi-modal detection, we explicitly extract both geometric and semantic features from the 2D images. We leverage camera parameters to lift these features to 3D. To improve the synergy of 2D-3D feature fusion, we also propose a multi-tower training scheme. We validate our model on the challenging SUN RGB-D dataset, advancing state-of-the-art results by **5.7** *mAP. We also provide rich ablation studies to analyze the contribution of each design choice.*

1. Introduction

Recognition and localization of objects in a 3D environment is an important first step towards full scene understanding. Even such low dimensional scene representation can serve applications like autonomous navigation and augmented reality. Recently, with advances in deep networks for point cloud data, several works [33, 56, 41] have shown state-of-the-art 3D detection results with point cloud as the *only* input. Among them, the recently proposed VOTENET [33] work by Qi *et al.*, taking 3D geometry input only, showed remarkable improvement for indoor object recognition compared with previous works that exploit all RGB-D channels. This leads to an interesting research question: Is 3D geometry data (point clouds) sufficient for 3D detection, or is there any way RGB images can further boost current detectors?

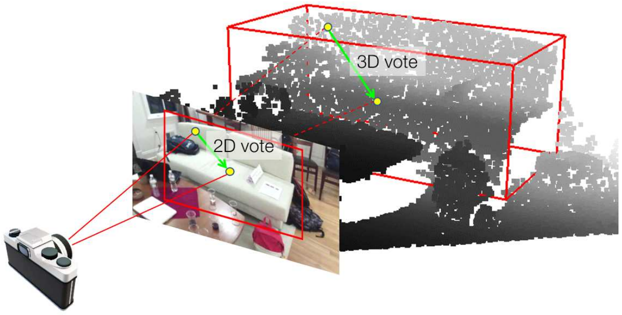

Figure 1. **Voting using both an image and a point cloud from an indoor scene.** The 2D vote reduces the search space of the 3D object center to a ray while the color texture in image provides a strong semantic prior. Motivated by the observation, our model lifts the 2D vote to 3D to boost 3D detection performance.

By examining the properties of point cloud data and RGB image data (see for example Fig. 1), we believe the answer is clear: RGB images have value in 3D object detection. In fact, images and point clouds provide *complementary* information. RGB images have higher resolution than depth images or LiDAR point clouds and contain rich textures that are not available in the point domain. Additionally, images can cover "blind regions" of active depth sensors which often occur due to reflective surfaces. On the other hand, images are limited in the 3D detection task as they lack absolute measures of object depth and scale, which are exactly what 3D point clouds can provide. These observations, strengthen our intuition that images can help point cloud-based 3D detection.

However, how to make effective use of 2D images in a 3D detection pipeline is still an open problem. A naïve way is to directly append raw RGB values to the point clouds – since the point-pixel correspondence can be established through projection. But since 3D points are much sparser, in doing so we will lose the dense patterns from the image domain. In light of this, more advanced ways to fuse 2D and 3D data have been proposed recently. One line of

*: equal contributions.
†: work done while at Facebook.