# PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection

ArXiv 2020

## Summary

In this paper PV-RCNN is presented as a novel 2-stage architecture for 3D object detection combining voxel-based and point-based feature learning. The authors argue that the combination of efficient 3D-CNNs that generate high-quality proposals and the flexibility of PointNet-like architectures for the computation of local features can be leveraged for highly-accurate 3D object detection. The authors present a thorough evaluation on the KITTI Object Detection Benchmark and report state-of-the-art results on several categories. Furthermore, an ablation study is presented that shows the effectiveness of the proposed components.

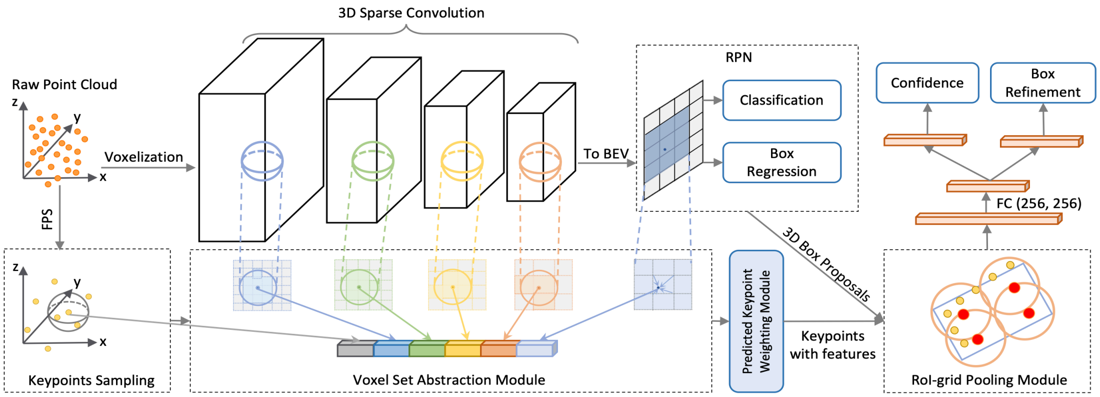

Figure 1: Illustration of the PV-RCNN architecture. The keypoint representation is chosen to reduce the memory footprint of the method and increase the contextual information in the feature aggregation step. This is used for the keypoint feature aggregation as well as for the proposal feature generation prior to the box refinement step.

## Main Contributions

- **Point-Voxel Feature Learning** The proposed architecture is a two-stage detector featuring a *Voxel-to-Keypoint Module* for scene encoding and a *Keypoint-to-Grid-RoI Module* for feature abstraction. A 3D-CNN based on sparse convolutions is used in the first step to compute a multi-scale feature representation. Based on the extracted features, an RPN computes bounding box proposals from a BEV representation of the voxelized feature maps. For the scene encoding, a sub-sampling process is employed that samples a number of representative keypoints to reduce computational complexity for each of which a feature representation is computed by leveraging the features extracted by the 3D-CNN.

  The second stage consists of a set abstraction module which is used to aggregate multi-scale features for each keypoint based on which proposal refinement is performed. Thus, PV-RCNN leverages the potential of voxel-based methods for accurate propsal generation and the quality of point-based local information aggregation for highly-accurate 3D object detection.

## Implementation Details

- **Voxel Set Abstraction** The VSA module is used to aggregate multi-scale features extracted by the 3D-CNN and assign them to the keypoints sampled from the grid to represent the entire scene. The VSA module can be thought of as a voxel-based adaptation of PointNet++'s set abstraction levels. It makes use of different sizes of receptive fields and extracts features from different levels of the 3D-CNN to obtain a rich multi-scale representation. These individual features are stacked and amended with features extracted from the raw input point cloud as well as the 2D-BEV representation from the RPN. The computational pipeline is illustrated in Figure 1.

- **Predicted Keypoint Weighting** As the keypoints are sampled randomly using Farthest Point Sampling, a number of background points might be included. In order to increase the influence of foreground points on the subsequent box refinement step, a weighting module is introduced. The *Predicted Keypoint Weighting* module uses a three-layer MLP to predict a foreground confidence score for each keypoint and is trained based on a segmentation ground truth extracted from the bounding boxes using focal loss.

- **RoI-grid Pooling** The RoI pooling module is required to compute a feature representation of each bounding box based on the scene encoding in the keypoints' features. For each box proposal, a $6 \times 6 \times 6$ grid is sampled uniformly for each of which a feature vector is aggregated from the surrounding keypoints using different sizes of receptive fields. This is displayed in Figure 3. The grid point features of each box proposal are then transformed into a 256-dimensional feature vector which acts as the final descriptor of the proposal based on which the refinement is computed as residuals w.r.t. the box proposal.

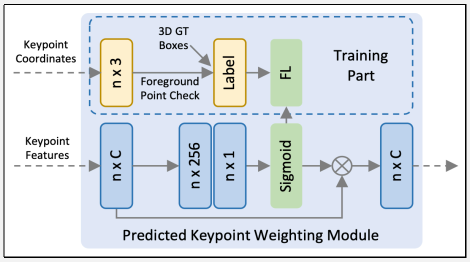

Figure 2: Illustration of the Predicted Keypoint Weighting Module.

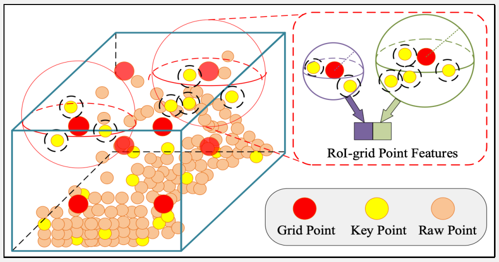

Figure 3: Illustration of the RoI-grid Pooling Module.

## Evaluation

- **KITTI 3D and BEV Object Detection** The authors present a thorough comparison of PV-RCNN with a large number of previous arts. PV-RCNN achieves state-of-the-art performance on the car class for all three difficulties, increasing the previous best scores by more than 1.5% mAP. New best scores are also obtained for detections on the cyclist class.

- **Scene Encoding Analysis** The authors perform ablation experiments indicating the importance of keypoint encoding. As a bridge between the first and second stage of the architecture, the keypoint encoding reduces the GPU footprint and further benefits the performance due to larger receptive fields.

## 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/1912.13192.pdf`