Multi-View 3D Object Detection Network for Autonomous Driving

CVPR 2017

Summary

In this paper, MV3D is presented as a novel 3D object detection architecture. At the first stage, the 3D Proposal Network generates bounding box proposals from a Bird's-Eye-View (BV) representation of the LiDAR data. The second stage consists of a Region-based Fusion Network that combines the BV representation with a Front-View (FV) LiDAR representation as well as RGB images. Based on a deep fusion scheme, the different input representations are leveraged to accurately refine the bounding box proposals. MV3D achieves state-of-the-art performance on the KITTI 3D Object Detection Benchmark, increasing the previously best score by more than 25%.

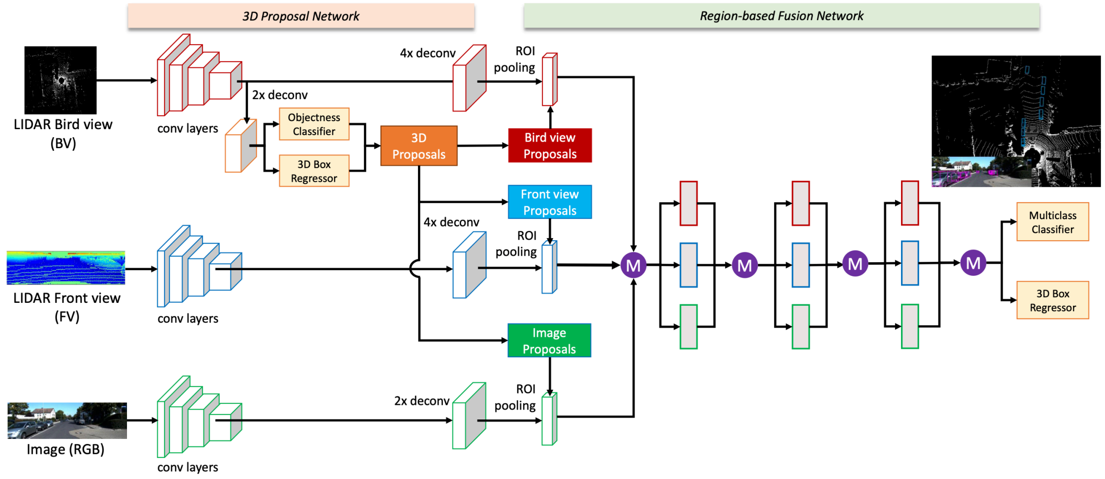

Figure 1: Illustration of the MV3D architecture. The BV representation is used for 3D proposal generation. A deep fusion scheme of BV, FV and RGB data is then used for bounding box refinement.

Main Contributions

- **3D Proposal Network** The authors propose to extract 3D bounding box proposals from the BV representation. Anchor boxes are designed based on size clusters of the ground truth boxes. Orientation regression is not performed during proposal generation in order to simplify the training process. Instead, a small number of orientations are used that correspond to the orientation of most road objects. To ensure sufficiently high resolution, bilinear upsampling is used, resulting in a down-sampling of the BV input by a factor of 4.

- **Region-based Fusion Network** In order to fuse information from BV, FV and RGB representations, the proposed bounding boxes are projected onto the respective input. ROI Pooling is then performed to ensure consistency of the feature dimensions. These multi-modal features are fused in a hierarchical fashion to predict semantic classes and bounding box refinements.

Implementation Details

- **Multi-View ROI Pooling** The features extracted from the three different data representations have to be unified before they can be fused. Thus, given the extracted proposals, a transformation is defined for each of the three different data representations. The corresponding ROIs are then combined with the extracted features to obtain fixed-length features for each proposal.

- **Deep Fusion** Contrary to conventional early-fusion or late-fusion schemes that either combine different features before or after applying feature transformations, the presented deep fusion scheme uses element-wise mean to combine the features before passing them through several feature transformation layers. Thus, a hierarchical fusion of the multi-modal features can be obtained. The different fusion schemes are displayed in Figure 2.

- **Network Regularization** The authors present two ways to regularize the Region-based Fusion Network during training. Drop-path training is used which corresponds to randomly dropping one of the input modalities. Moreover, auxiliary losses are used to supervise the fusion layers. This approach is illustrated in Figure 3.

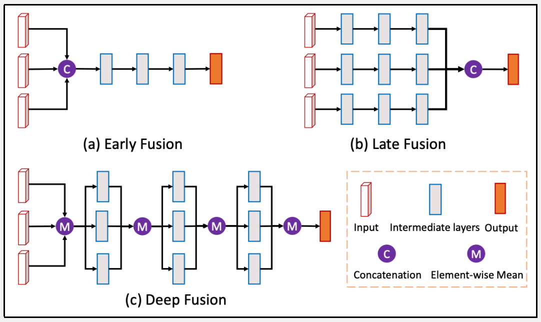

Figure 2: Illustration of different fusion schemes.

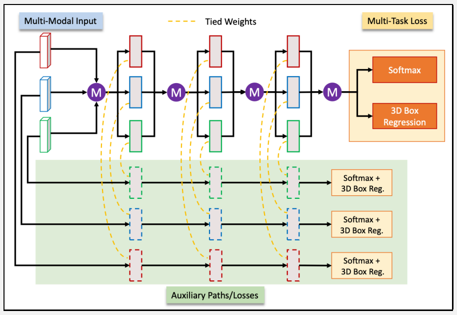

Figure 3: Illustation of the auxiliary losses. The dotted lines indicate weight sharing between the layers.

Evaluation

- **Ablation Study** Ablation experiments are conducted on the fusion scheme as well as the selection of multiple data representations. The results indicate that deep fusion, especially with auxiliary losses, as well as multi-modal features contribute to increased accuracy. It can be observed that the BV representation alone performs considerably better than FV or RGB. This can be attributed to the fact that objects in the BV representation are of the same size regardless of the distance to the sensor and clearly visible as no occlusion occurs.

- **KITTI 3D and BEV Object Detection** MV3D is evaluated on the KITTI 3D and BEV Object Detection Benchmark and is shown to outperform previous arts by large margins. This holds for predictions with and without image data. The baseline models include one purely LiDAR-based approach (VeloFCN), one model based on stereo images (3DOP) and one on monocular images (Mono3D).

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/1611.07759.pdf`