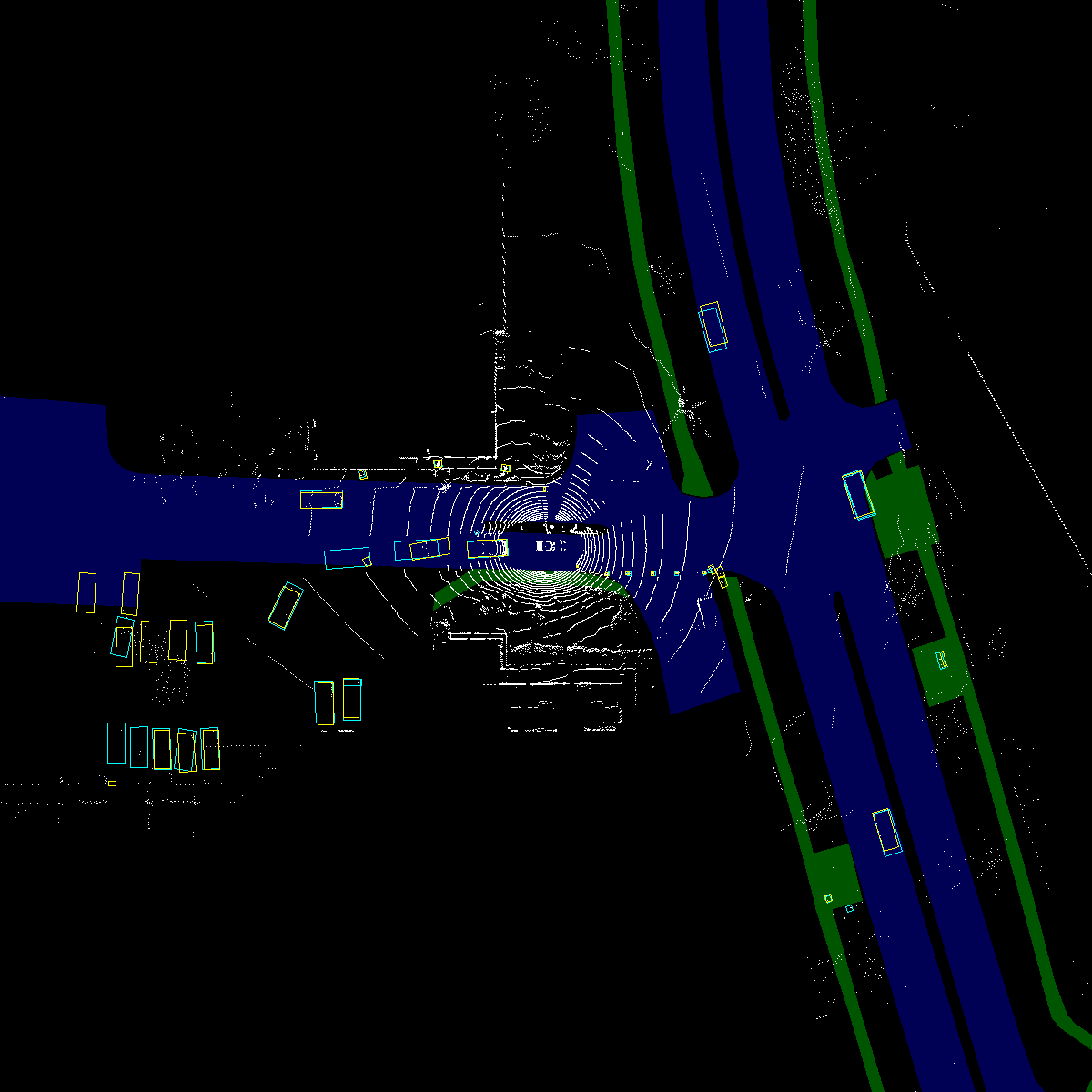
# MapFusion: A General Framework for 3D Object Detection with HDMaps

## Abstract

3D object detection is a key perception component in autonomous driving. Most recent approaches are based on Lidar sensors only or fused with cameras. Maps (e.g., High Definition Maps), a basic infrastructure for intelligent vehicles, have not been well exploited for boosting object detection tasks. In this paper, we propose a simple but effective framework - MapFusion to integrate the map information into modern 3D object detector pipelines. In particular, we design a FeatureAgg module for HD Map feature extraction and fusion, and a MapSeg module as an auxiliary segmentation head for the detection backbone. Our proposed MapFusion is detector independent and can be easily integrated into different detectors. The experimental results of three different baselines on a large public autonomous driving dataset demonstrate the superiority of the proposed framework. By fusing the map information, we can achieve 1.27 to 2.79 points improvements for mean Average Precision (mAP) on three strong 3d object detection baselines.

## Introduction

Autonomous driving (AD) has drawn significant attention over the past years. Despite recent progress, AD is still considered as one of the most challenging tasks. Developing autonomous driving systems needs to integrate many advanced robotics techniques such as perception, planning, and control. Perception module interprets the surrounding environment and provides inputs for downstream planning and control modules. For object detection tasks, cameras are the most commonly used sensors and many advanced camera-based 2D detectors have been developed. However, autonomous driving vehicles need to recover 3D locations of surrounding objects accurately and reliably, which are extremely hard for 2D image-based detectors, due to the loss of distance information in the perspective projection process from 3D space to 2D image. Therefore, active 3D sensors such as LiDAR devices, are prerequisites for most AD systems. With the development of deep learning on the 3D point clouds, many LiDAR-based 3D object detectors have been proposed. Based on the different input representations, approaches can be generally categorized as point-based and voxel-based methods. Compared with point-based methods, the latter is much more efficient and its computation time is independent of the size of the point cloud. The performance for existing LiDAR-based 3D object detection approaches, however, can be further improved, especially on reducing the false positives and false negatives for detected objects. This problem comes from two main reasons. First, without enough texture information, it is difficult to distinguish between foreground and background objects. Second, the number of scanned points for small objects or objects far away is very sparse, which results in the failure of object detection and recognition. Fusing with other sensors e.g., camera, radar is a commonly used strategy to handle this kind of problem. However, as image quality is easily affected by the environment illumination and weather situation, solely relying on fusion with images can’t provide stable detection results.



*Caption: "Detection results comparison" | Explanation: "Comparison between the detection outputs without MapFusion (a) and with MapFusion (b). The integration significantly reduces false positives as marked by red cycles. Yellow rectangles denote ground truth and cyan rectangles denote detected objects." | Ref: PDF p.1*

## RELATED WORK

With the development of range sensors and AD techniques, 3D object detection in driving scenarios draws more and more attention. To solve the problem, one of the commonly used strategy is projecting the 3D point cloud into 2D (e.g., bird-eye-view or front-view) to obtain the corresponding 2D detection result, then the final result can be obtained by re-projecting the 2D BBox into 3D. Benefiting from the development of graphics processing resources, volumetric convolutional approaches become another representative direction for 3D object detection. Voxelnet is a pioneering method, which employs 3D convolution to detect the 3D objects by converting the LiDAR point cloud to voxels. Inspired by Voxelnet, SECOND and PointPillars have been proposed, which use different 3D voxel representations for 3D object detection. By using structure information of 3D point cloud, a novel framework, which can improve the localization precision of single-stage detectors. Meanwhile, CenterPoint proposes to represent, detect, and track 3D objects as points, which detects centers and other attributes of the objects, then refines the estimated information using additional point features on the objects. Meanwhile, PointNet proposes a novel technique for point cloud feature extraction. Based on PointNet, several state-of-the-art methods have been proposed for 3D object detection. To avoid destroying the hidden information about free space, a novel strategy proposes to utilize 3D ray casting and batch-based gradient learning strategies for 3D object detection. Besides, effective intersection over union (IoU) operations have been proposed to generalize the losses in 3D object detection, which improve the accuracy of 3D object detection.

## PROPOSED APPROACH

The overview of the proposed framework, MapFusion, can be roughly divided into two parts: the standard 3D object detection block and the map feature extraction block. The standard LiDAR-based 3d object detection pipeline is described in the top red dotted box. The input LiDAR point cloud is sent to a 3d feature extractor such as 3D Sparse Convolution and outputs the features for voxels. The map feature extraction block is denoted as the blue dotted box, which takes the HDMap as the input. After a 2D feature extractor, the map feature with the same size of voxel feature is extracted. Then the 3D point cloud features and the map information are aggregated with a concatenation operation for each voxel. Then, a detection head including the region proposals, box regression, and categories classification follows the fused features. In addition, an auxiliary segmentation head, namely MapSeg, is added to further improve the feature extraction capability. Our MapFusion is an end-to-end framework that can be easily integrated into any standard 3d object detection pipeline with only slight modifications.

[Image missing: page2\_img1.jpg] Caption: "Overview of MapFusion framework" | Explanation: "This figure shows the modular design of our MapFusion framework. The red dotted box represents the standard 3D object detection pipeline, and the blue dotted box represents the innovative map feature extraction block integrated into the system." | Ref: PDF p.2

## EXPERIMENTAL RESULTS

MapFusion is a general framework and can be easily integrated into mainstream 3d object detection methods. We first introduce the baseline detectors that are used to evaluate MapFusion. Three state-of-the-art point cloud-based 3d object detectors are compared here: SECOND utilizes the sparse convolution to significantly increase the speed of both training and inference; PointPillars uses a pillar (vertical columns) representation and regards the pillars as a pseudo image. Then a standard 2D detection backbone can then be employed; CenterPoint is a strong anchor-free baseline, which ranks the top among all LiDAR-only method in public nuScenes and Waymo dataset. We evaluate MapFusion on nuScenes dataset with the three baseline detectors. The proposed Map Fusion can effectively improve the performance of the baselines in both nuScenes detection score (NDS) and mean Average Precision (mAP), which sufficiently proves the effectiveness of the proposed MapFusion. Using MapFusion, for NDS, the improvements of SECOND, PointPillars and Cen