ICIEA 2022

16 - 19 Dec 22
Chengdu, China

ICIEA22-000215
3D Object Detection Based on Feature Fusion of Point Cloud Sequences

Zhenyu Zhai, Qiantong Wang*, Zongxu Pan,
Wenlong Hu, Yuxin Hu
Aerospace Information Research Institute, CAS

Abstract

Continuous frames of point-cloud-based object detection is a new research direction. Currently, most research studies fuse multi-frame point clouds using concatenation-based methods. The method aligns different frames by using information on GPS, IMU, etc. However, this fusion method can only align static objects and not moving objects. In this paper, we proposed a non-local based multi-scale feature fusion method, which can handle both moving and static objects without GPS- and IMU-based registrations.

Introduction

In dense point cloud scenes, the geometric shape of the object is relatively complete. However, these lidar techniques, which use more laser beams, are expensive as well. Reducing the cost of lidar techniques is a problem in the large-scale application of automatic driving. The autopilot company nuTonomy tried to use cheap 32-line lidar and released the NuScenes dataset. Unlike the KITTI dataset, which uses 64-line lidar, the NuScenes dataset is built with 32-line lidar, exacerbating the sparsity of point clouds. Therefore, NuScenes officially recommends concatenating 10 calibrated point cloud frames to obtain denser point clouds. Compared with single-frame point cloud, multiple frames provide a denser description of the surrounding environment as a result of multi-view observations.

Currently, multi-frame-based object detectors inevitably face the problem of registration between different frames. Usually, most of them align different frames via GPS and IMU, etc. However, registration can align static objects but not moving objects. Consequently, such fusion will cause motion blur.

Method

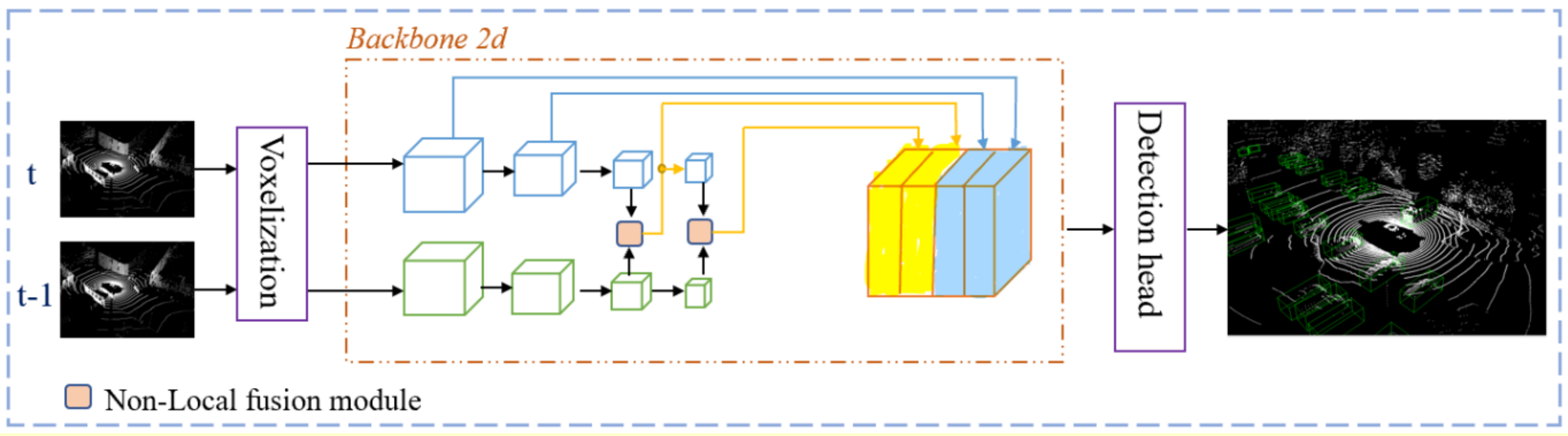

Figure 1. Overall structure. The network includes three parts: voxelization module, feature extraction and fusion module, and detection head.

Our network is an improvement on the single-frame network PointPillars. The proposed method adopts PointPillars as the backbone. The overall structure of the proposed network, shown in Figure. 1, includes three parts: (1) Voxel representation; (2) Feature extraction and fusion; (3) Detection head. The input of the network is two frames in the point cloud time series (frame t-1 and frame t). First, the proposed method feeds two adjacent point clouds into the same feature extraction network. Then, a non-local model, modeling the relationship between objects within two frames, is adopted to fuse feature maps of two point clouds. Finally, the fused feature maps are fed into the detection head.

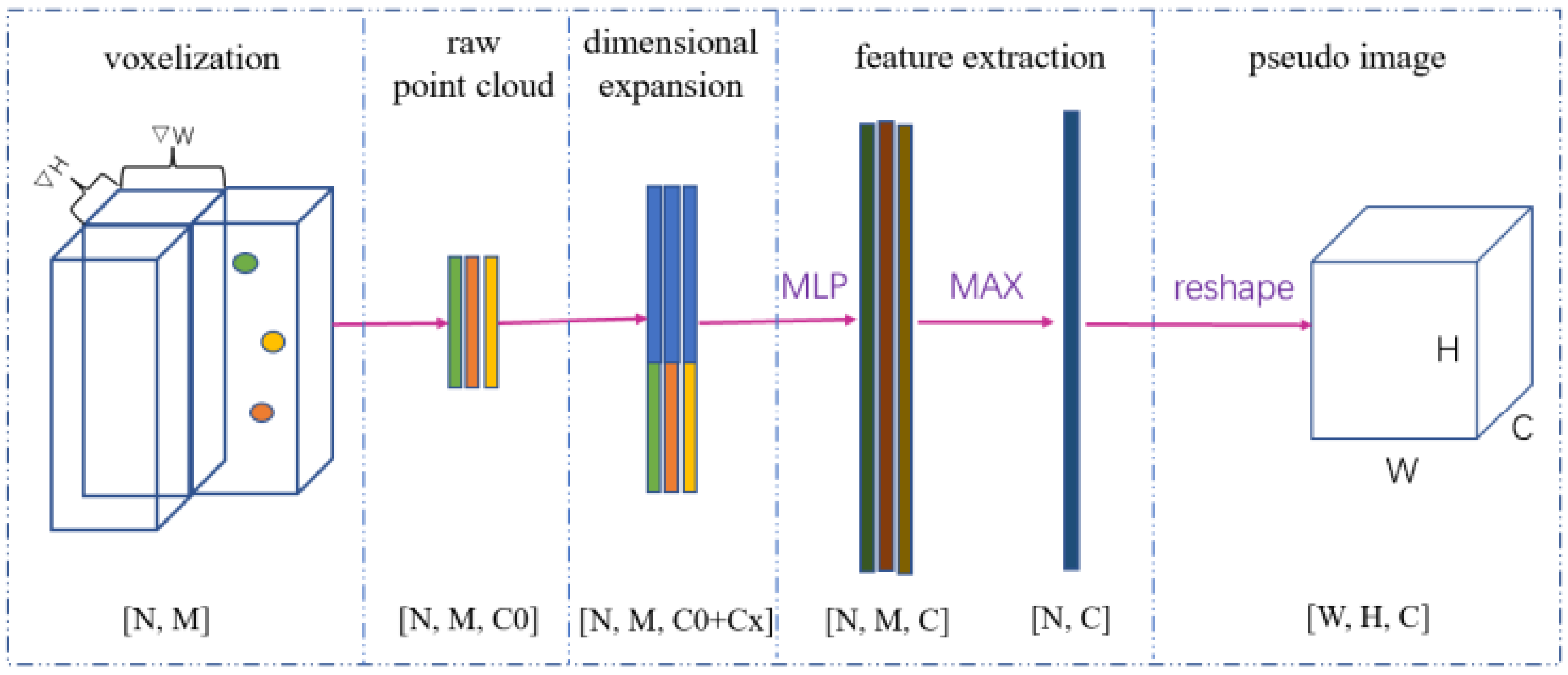

Figure 2. Point cloud voxelization.

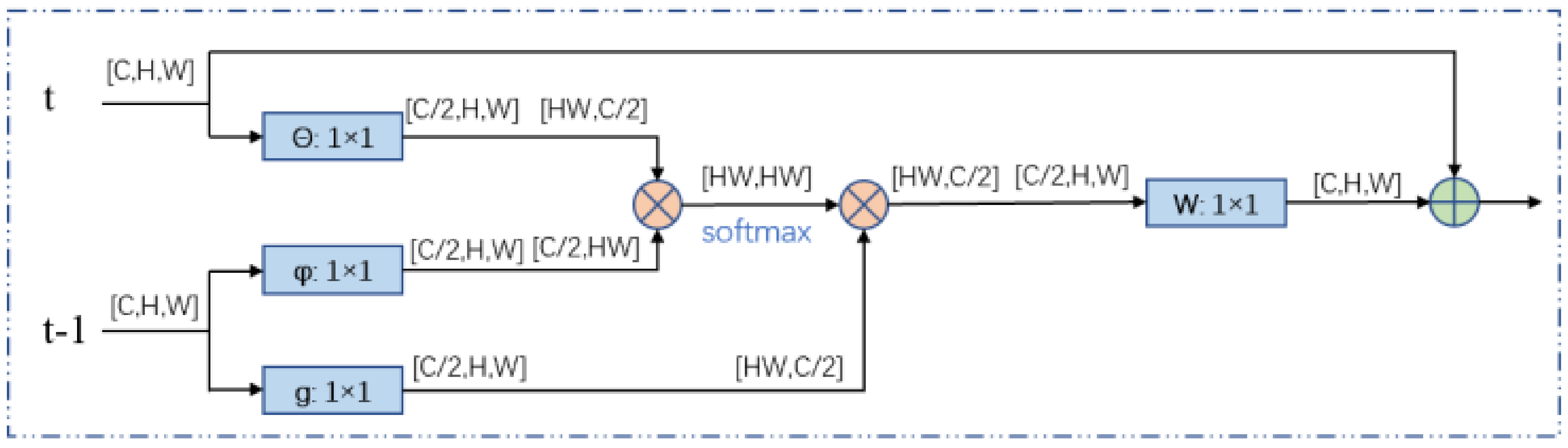

Figure 3. Non-local module. The blue symbols represent 1 × 1 convolutions, the orange symbols represent matrix multiplication, and the green symbols represent element-wise addition.

Results

Method	mAP	Car	Truck	Ctr. Vhl	Bus	Trailer	Barrier	Motorcycle	Bicycle	Pedestrian	Tr. Co.
PointPillars*	34.7	72.0	34.9	5.7	56.0	28.2	37.3	21.1	**1.4**	56.5	33.8
Ours	**36.6**	**72.8**	**36.3**	**7.2**	**56.8**	**31.0**	**42.9**	**23.1**	1.2	**58.7**	**35.8**

Summary

In this paper, we propose a non-local-based feature fusion method to fuse two frames of point cloud. The proposed method can handle both moving and static objects without external information-based registration.