3D Object Detection using Dynamic • Static Motion Information from Multi-camera Image Sequences

Yongho Son°[1)], Jungho Kim[1)], Jun Won Choi°°[2)]

Dept. of Artificial Intelligence°[1)], Dept. of Electronics Engineering°°[2)]

Hanyang University

{yhson°, jhkim}@spa.hanyang.ac.kr, junwchoi@hanyang.ac.kr°°

ABSTRACT

This paper proposes a novel three-dimensional object detection method that leverages both dynamic and static motion information from time-series image data captured by camera sensors. First, a bird's-eye view (BEV) feature map is generated from consecutive image frames using a depth estimation–based network. Among the BEV feature maps across multiple time steps, those from past frames are aligned to the current BEV coordinate system using ego-motion information. The aligned past BEV feature maps and the current BEV feature map are then aggregated in the spatiotemporal domain through a recurrent neural network that extracts dynamic motion information. The final BEV feature map is obtained by applying a residual connection from the current BEV feature map to the aggregated features, and it is used to produce the final 3D object detection results. For performance evaluation, experiments were conducted on the nuScenes dataset [3], demonstrating that the proposed method achieves higher detection accuracy compared to conventional approaches.

1. Introduction

With the recent advances in autonomous driving technology, extensive research has been conducted on perception tasks. Among these efforts, a research direction has emerged that generates bird's-eye-view (BEV) feature maps representing the surrounding environment from multiple on-vehicle cameras to perform 3D object detection. However, when using camera sensors, various types of noise such as occlusion, motion blur, and camera defocus can occur. To overcome these limitations, this study proposes a spatiotemporal fusion method. The proposed approach consists of temporal fusion, which captures object motion information across time frames, and spatial fusion, which refines object representations. By compensating for BEV feature map noise through the spatiotemporal fusion module, the proposed method achieves superior 3D object detection performance compared to existing techniques.

2. Proposed method

The proposed 3D object detection algorithm uses multi camera images as input. The overall architecture is shown in Figure 1. First, sequential camera image data are fed into a convolutional backbone network to extract camera feature maps. To estimate the depth of objects and surrounding context, a depth estimation network is employed, and it is trained with LiDAR point cloud data as ground truth. The features extracted by the backbone are projected into 3D space using the estimated depth information [1], and voxel pooling along the z axis is applied to generate bird's eye view feature maps. These bird's eye view feature maps exist for each time frame, and the feature maps from different time steps are aligned to the current bird's eye view position by compensating for ego motion.

2.2) Dynamic information extraction (Dynamic Flow)

For each time frame, the BEV feature map encodes diverse cues for 3D object detection. When aligning BEV feature maps using ego motion, static objects align well, whereas dynamic objects exhibit varying displacement depending on their velocity. To address this, a ConvLSTM augmented with Deformable Convolutional Networks (DCN) [2] is employed. By learning kernel offsets, DCN adapts to object motion during alignment. The LSTM comprises a forget gate, input gate, and output gate, and is further extended with a motion gate to better capture motion specific features. The motion gate takes as input the difference between the hidden state at time t−1 and the BEV feature map at time t.

2.3) Static information extraction (Static Flow)

The detector is trained to localize objects at the current time step t, thus information from the current frame plays a crucial role. Therefore, to reliably preserve static content from the current frame, the current BEV feature map is connected via a residual connection to the BEV feature map in which dynamic motion information has been extracted.

Method	Aggregation Strategy		Performance	
	Static Flow	Dynamic Flow	mAP(%)	NDS(%)
Baseline			20.8	24.2
Ours	✓		21.2	24.7
		✓	21.5	25.1
	✓	✓	**22.6**	**25.7**

Table 1. Quantitative comparison

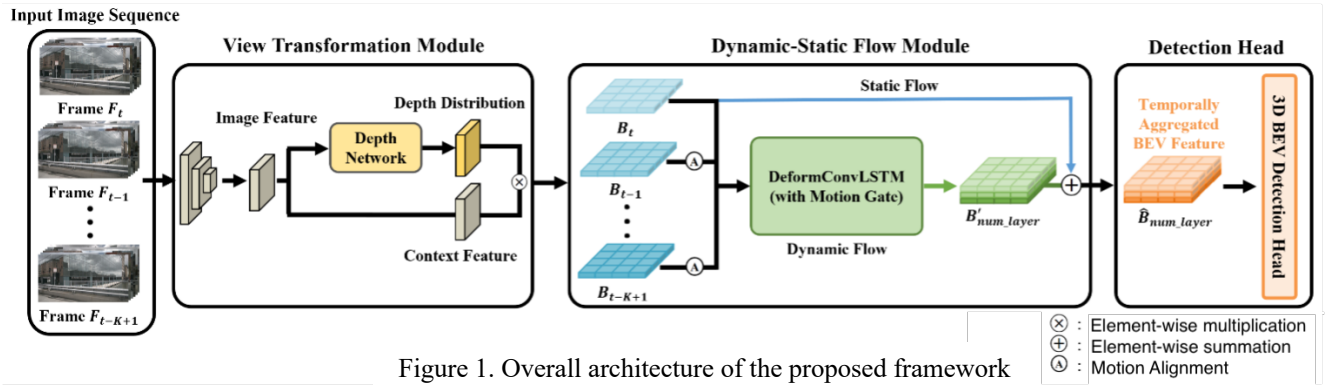

Figure 1. Overall architecture of the proposed framework

To match the channel dimensions between the two BEV feature maps, a 1×1 convolution is applied to expand the channels. The resulting final BEV feature map is then fed to the detection head to produce 3D object detection outputs.

3. Experiment

To evaluate the proposed 3D object detection model, experiments were conducted using the nuScenes autonomous driving dataset [3]. The input image resolution was set to 256×704, and the model was trained using one-seventh of the uniformly sampled training set. The overall 3D object detection performance was assessed on the full validation set, and the results are summarized in Table 1. When utilizing dynamic motion information and static motion information, the proposed method achieved performance improvements of 0.5% and 0.9% points in NDS [3], respectively, compared to the baseline. In particular, the spatiotemporal fusion approach demonstrated a 1.5% performance gain over BEVDepth [1], a representative 3D object detection technique.

4. Conclusion

This study proposes a 3D object detection algorithm that takes multi-camera images as input. First, feature representations extracted by a backbone network are projected into 3D space using estimated depth information, and bird's-eye-view (BEV) feature maps are generated through voxel pooling along the z-axis. These BEV feature maps exist for each time frame and are aligned to the current BEV coordinate system using motion information. To extract dynamic motion features, a ConvLSTM integrated with Deformable Convolutional Networks (DCN) is employed. Subsequently, a 1×1 convolution is applied to expand the feature dimension, and a residual connection is introduced to the dynamically refined BEV feature map. The final BEV feature map, which contains both dynamic and static motion information, is then used by the detection head to perform object detection. The proposed algorithm was evaluated on the nuScenes dataset [3], and the results confirm the effectiveness of the proposed spatiotemporal fusion methodology.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No.2020R1A2C2012146).

5. Reference

[1] Li, Y., et al. "BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection," In Proceedings of the AAAI Conference on Artificial Intelligence.

[2] Dai, J., et al. "Deformable convolutional networks," In Proceedings of the IEEE international conference on computer vision (pp. 764-773).

[3] Caesar, H., et al. "nuScenes: A multimodal dataset for autonomous driving," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp.11621-11631).