Accurate and Fast Monocular 3D Object Detection with Adaptive Feature Aggregation Centric Enhance Network

Peng-Wei Lin
Institute of Mechanical and Electrical Engineering
National Taipei University of Technology
Taipei 10608 Taiwan
109669006@ntut.org.tw

Chih-Ming Hsu
Department of Mechanical Engineering
National Taipei University of Technology
Taipei 10608 Taiwan
jmshiu@ntut.edu.tw

ABSTRACT

Three-dimensional (3D) object detection is crucial in autonomous driving. Monocular 3D object detection has become a popular area of research in autonomous driving because of its ease of deployment and cost-effectiveness. In real-world applications of autonomous driving, a detector must be both real-time and accurate. These can be achieved using deep learning. A one-stage center-based object detector is suitable for real-world applications. However, in center-based object detectors, object-centric estimation plays an important role because it significantly influences detection results. To address this issue, we proposed a real-time monocular 3D object detection neural network called the adaptive feature aggregate centric enhance network. The model is an anchor-free and center-based method. To enhance accuracy while maintaining inference speed, we propose an adaptive feature aggregation network that aggregates multiscale features with weighting. In addition, we proposed a centric enhance module for heatmap prediction to improve the accuracy of object localization and classification. Our model can achieve 35 frames per second using an Nvidia RTX3070 accelerator. Extensive experiments on the KITTI benchmark demonstrated that our method has good mean average precision (mAP) for small objects.

KEYWORDS

Real-time, 3D Object Detection, Monocular Image, Deep Learning, Attention Mechanism, Autonomous Driving.

METHODOLOGY To achieve accurate and fast detection, we proposed AFACENet for monocular 3D object detection. Our model adopted the Yolo v4 concept and divided the detector into three parts, the backbone, neck, and head. Our model adopted deep layer aggregation (DLA) 34 [1] as our backbone, the proposed AFAN as our neck, and the proposed CEM as our prediction heatmap head. In this section, we elaborate on the proposed AFAN and CEM. The overall architecture of the proposed AFACENet is shown in Figure 1

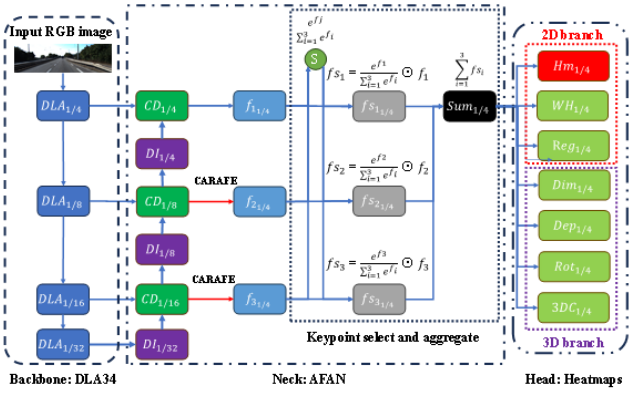

Figure 1: Overall architecture of the proposed AFACENet.

In this paper, we propose the aggregate feature attention network (AFAN) to calculate weights from different feature scales and then aggregate the features through concatenation while retaining information. In addition, our propose a centric enhancement module (CEM) to enhance the object recognition and localization abilities. Our model tests on the KITTI dataset [2], which is widely used in autonomous driving. Our contributions can be summarized as follows: Our AFAN can aggregate multiscale features by weighting by softmax function and concatenation to emphasize important features. It uses CARAFE and Interpolation for upsampling to balancing the performance in different categories. Our proposed CEM to can enhance classification and localization ability. Our AFACENet shows the performance of the KITTI dataset, particularly for car, which is far away from sensor. Our AFACENet achieves 35FPS real-time detection speed by using an Nvidia RTX3070 GPU.

REFERENCES

[1] F. Yu, D. Wang, E. Shelhamer and T. Darrell, "Deep Layer Aggregation," in Proc IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 2403-2412, doi: 10.1109/CVPR.2018.00255.

[2] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, 3D Detection Evaluation 2017. [Online]. 2017. Available: http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=3d.