JANT 한국항행학회논문지
Journal of Advanced Navigation Technology

기반 기술

J. Adv. Navig. Technol. 27(2): 190-196, Apr. 2023

3차원 객체 탐지를 위한 어텐션 기반 특징 융합 네트워크

Attention based Feature-Fusion Network for 3D Object Detection

유 상 현[1] · 강 대 열[2] · 황 승 준[2] · 박 성 준[2] · 백 중 환[2*]
[1]한국항공대학교 항공우주 및 기계공학부
[2]한국항공대학교 항공전자정보공학부

Sang-Hyun Ryoo[1] · Dae-Yeol Kang[2] · Seung-Jun Hwang[2] · Sung-Jun Park[2] · Joong-Hwan Baek[2*]
[1]School of Aerospace and Mechanical Engineering, Korea Aerospace University, Goyang, 10540, Korea
[2]School of Electronics and Information Engineering, Korea Aerospace University, Goyang, 10540, Korea

[요 약]

최근 들어, 라이다 기술의 발전에 따라 정확한 거리 측정이 가능해지면서 라이다 기반의 3차원 객체 탐지 네트워크에 대한 관심이 증가하고 있다. 기존의 네트워크는 복셀화 및 다운샘플링 과정에서 공간적인 정보 손실이 발생해 부정확한 위치 추정 결과를 발생시킨다. 본 연구에서는 고수준 특징과 높은 위치 정확도를 동시에 획득하기 위해 어텐션 기반 융합 방식과 카메라-라이다 융합 시스템을 제안한다. 먼저, 그리드 기반의 3차원 객체 탐지 네트워크인 Voxel-RCNN 구조에 어텐션 방식을 도입함으로써, 다중 스케일의 희소 3차원 합성곱 특징을 효과적으로 융합하여 3차원 객체 탐지의 성능을 높인다. 다음으로, 거짓 양성을 제거하기 위해 3차원 객체 탐지 네트워크의 탐지 결과와 이미지상의 2차원 객체 탐지 결과를 결합하는 카메라-라이다 융합 시스템을 제안한다. 제안 알고리즘의 성능평가를 위해 자율주행 분야의 KITTI 데이터 세트를 이용하여 기존 알고리즘과의 비교 실험을 수행한다. 결과적으로, 차량 클래스에 대해 BEV 상의 2차원 객체 탐지와 3차원 객체 탐지 부분에서 성능 향상을 보였으며 특히 Voxel-RCNN보다 차량 Moderate 클래스에 대하여 정확도가 약 0.47% 향상되었다.

[Abstract]

Recently, following the development of LIDAR technology which can detect distance from the object, the interest for LIDAR based 3D object detection network is getting higher. Previous networks generate inaccurate localization results due to spatial information loss during voxelization and downsampling. In this study, we propose an attention-based convergence method and a camera-LIDAR convergence system to acquire high-level features and high positional accuracy. First, by introducing the attention method into the Voxel-RCNN structure, which is a grid-based 3D object detection network, the multi-scale sparse 3D convolution feature is effectively fused to improve the performance of 3D object detection. Additionally, we propose the late-fusion mechanism for fusing outcomes in 3D object detection network and 2D object detection network to delete false positive. Comparative experiments with existing algorithms are performed using the KITTI data set, which is widely used in the field of autonomous driving. The proposed method showed performance improvement in both 2D object detection on BEV and 3D object detection. In particular, the precision was improved by about 0.54% for the car moderate class compared to Voxel-RCNN.

Key word : Attention module, KITTI dataset, Multi sensor data fusion, 2D object detection, 3D object detection.

http://dx.doi.org/10.12673/jant.2023.27.2.190

Received 22 February 2023; **Revised** 31 March 2023
Accepted (Publication) 20 April 2023 (30 April 2023)

***Corresponding Author; Joong-Hwan Baek**

Tel: +82-02-300-0125
E-mail: jhbaek@kau.ac.kr

Ⅰ. 서 론

특정 객체의 위치와 클래스 정보를 탐지하는 객체 탐지는 컴퓨터 비전 분야에서 중요하게 다뤄지고 있다. 특히, 합성곱 신경망의 등장과 더불어 2차원 이미지 내에서 객체를 탐지하는 2차원 객체 탐지 분야는 많은 발전을 이룩해 왔다. 그러나 상대적으로 3차원 객체 탐지 분야는 그 발전이 더디었는데, 3차원 좌표에서 위치를 정확히 추정하기 위해서는 정확한 깊이 정보가 필요하기 때문이다[1]. 센서에 따라 깊이 정보를 얻는 방식이 다르기 때문에 3차원 객체 탐지의 방식은 대표적으로 카메라 기반, 라이다 기반, 카메라-라이다 융합 방식으로 나뉜다[2].

이 중에 라이다 기반 방식은 센서로부터 물체의 3차원 위치 정보를 직접적으로 얻을 수 있다는 장점이 있다. 따라서 라이다 포인트 클라우드 데이터만을 입력으로 하는 많은 3차원 객체 탐지 네트워크가 등장하였고 카메라만을 이용하였을 때보다 성능적인 이점이 존재함에 따라 3차원 객체 탐지 분야의 주류가 되었다.

라이다 기반의 3차원 객체 탐지 분야는 포인트 기반 방식과 그리드 기반 방식으로 나누어진다. 포인트 기반 방식은 정확하고 성능이 높지만, 데이터양이 많기 때문에 연산 비용이 높아 처리 속도가 느리다는 단점이 있다. 그리드 기반 방식은 포인트 클라우드 데이터를 복셀이나 기둥 등으로 변환하는 과정을 거친 후 활용하는 방식이다. 정보의 손실이 일어나지만, 메모리 지역성 (memory locality)이 높아 CNN (convolutional neural network)을 활용하기 적합하여 처리 속도의 이점을 취할 수 있다[3]. 그리드 기반 방식은 희소 3차원 합성곱 (sparse 3D convolution)을 이용하여 다운샘플링을 거치고 마지막 단에서 특징 맵을 추출하는 3차원 백본 네트워크 구조를 따른다. 그리드 기반 방식은 이를 통해 라이다 기반 3차원 객체 탐지 네트워크에서 대표적인 프레임워크로 주로 활용된다[4].

그러나 다운샘플링 과정에서 공간적인 정보의 손실이 발생해 포인트 클라우드 데이터의 정확한 위치 정보를 이용할 수 없게 된다[5]. 따라서 기존의 3차원 백본 네트워크는 마지막 단만을 이용하여 특징 맵을 추출하기 때문에 위치 정보가 다소 부정확한 3차원 박스를 생성하는 문제가 있다.

이런 문제점을 보완하기 위해 본 연구에서는 고수준 특징과 높은 위치 정확도를 동시에 획득하기 위하여 3차원 백본 네트워크의 다양한 단의 정보를 효과적으로 융합하는 방법을 제안한다. 다양한 단의 멀티 스케일 특징 정보를 효과적으로 결합하기 위하여 어텐션 기법을 활용하는 결합 모듈 (integrating module)을 제안한다.

추가적으로 최근 라이다 기반 3차원 객체 탐지 네트워크의 성능을 높이는 방법으로 라이다 특징과 이미지 특징을 융합하는 카메라-라이다 특징 융합 방식이 활발히 연구되고 있다[6]. 라이다 특징과 이미지 특징을 융합하는 방식에는 총 3가지가 있으며, 특징을 추출하는 시점에 따라 전기 융합 (early fusion), 깊은 융합 (deep fusion), 후기 융합 (late fusion)으로 나뉜다[7]. 상대적으로 전기 융합 방식과 깊은 융합 방식은 두 데이터 간의 교차 양식 정보를 추출할 수 있는 큰 잠재력을 지녔다는 장점이 있지만, 데이터 간의 정렬 문제와 구조적인 복잡도가 있다. 반면에 후기 융합 방식은 두 네트워크를 독립적으로 적용할 수 있다는 점에서 구조적으로 간단하게 설계할 수 있다[8].

후기 융합 방식은 일반적으로 새로운 탐지를 하기보다는 기존 탐지를 제거하는 것에 자주 이용되지만 제거하는 과정에서 참 양성 (true positive)도 제거될 수 있다는 문제점이 존재한다[8]. 따라서 본 연구에서는 참 양성 검출 결과는 유지하고 거짓 양성 (false positive)을 제거할 수 있는 후기 융합 방식을 사용한다. 라이다 포인트 클라우드 데이터 분석 결과, 라이다 센서로부터 먼 거리에 있는 객체들이 부족한 포인트 클라우드 데이터로 인해서 잘못 탐지되는 결과가 다수 존재하였다. 특히 이 거짓 양성 결과는 탐지된 3차원 객체들 군집과 떨어져서 존재하는 경우가 다수이다. 이에 따라 본 연구에서는 미리 학습시킨 2차원 객체 탐지 네트워크를 통해 검출한 2차원 상의 결과와 라이다 거리 정보를 융합하여 3차원 객체 탐지 결과의 신뢰도를 높이는 카메라-라이다 융합 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에서 사용한 라이다 기반 3차원 객체 탐지 네트워크인 Voxel-RCNN (voxel regions with convolutional neuron networks features)의 구조와 결합 모듈에 적용한 어텐션 모듈인 CBAM (convolutional block attention module)에 대해서 소개한다[3], [9]. 3장에서는 전체적인 네트워크 구조에 대해서 설명하고 결합 모듈과 카메라-라이다 융합 방식을 설명한다. 4장에서는 KITTI 데이터 세트를 이용하여 실험 결과를 도출하고 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

2장에서는 본 연구에서 기본 네트워크로 사용하는 3차원 객체 탐지 네트워크인 Voxel-RCNN과 어텐션 모델인 CBAM에 대해서 설명한다.

2-1 Voxel-RCNN

Voxel-RCNN은 본 연구에서 기본 네트워크로 사용하는 모델로 라이다 기반 3차원 객체 탐지 네트워크 중 그리드 기반 방식의 네트워크이며 2단계로 구성된다. 첫 번째 단계에서는 3차원 백본 네트워크를 통해 BEV (bird eye view) 특징 맵을 추출하고 2차원 백본 네트워크와 RPN (region proposal network)를 통해 3차원 박스를 예측한다. 두 번째 단계는 Voxel ROI (region of interest) pooling 단계로 3차원 백본 네트워크의 마지막 두 개의 단의 특징과 복셀의 좌표를 이용하여 예측된 3차원 박스 근처의 특징점을 추출하여 박스 보정을 하는 단계이다. 이 네트워크는 전형적인 그리드 기반의 라이다 3차원 객체 탐지 네트워크의 흐름을 따르며 포인트 기반 방식에서 사용하던 박스 보정

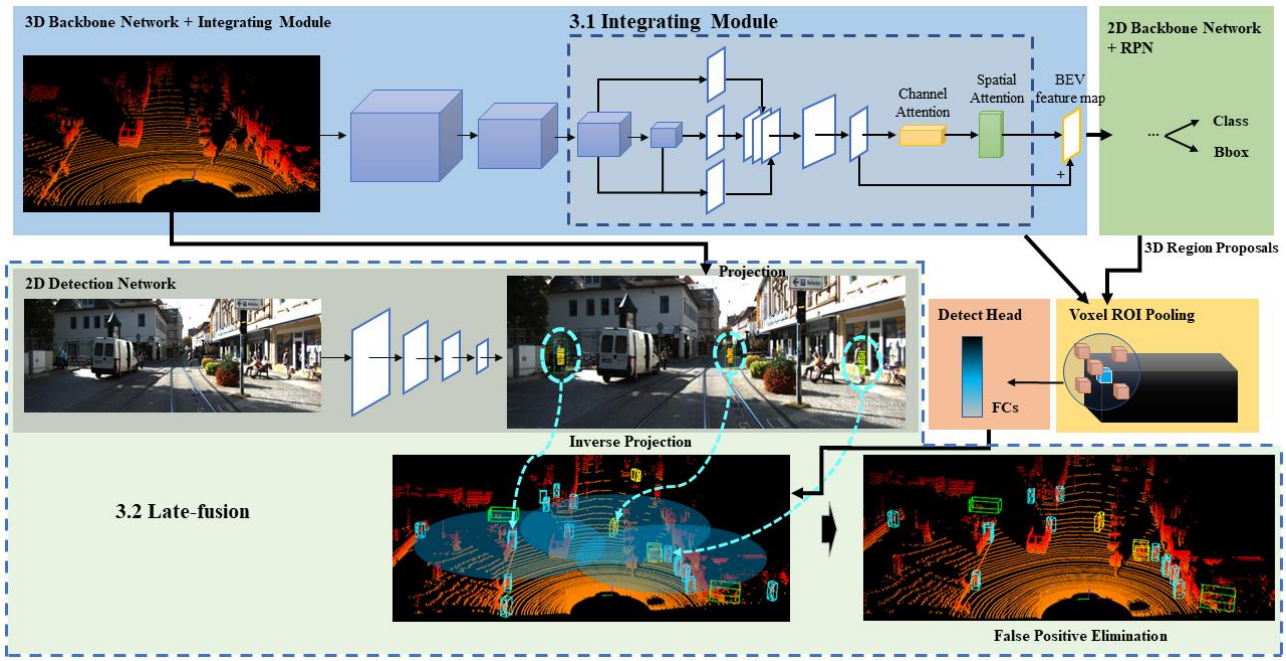

그림 1. 3차원 객체 탐지를 위한 어텐션 기반 특징 융합 네트워크 모델 구조
Fig. 1. Architecture of Attention based Feature-Fusion Network for 3D Object Detection

과정을 그리드 기반 방식에서 활용하여 속도와 정확성을 동시에 향상시킨다.

2-2 CBAM

CBAM은 간단한 아이디어이지만 범용성과 성능을 인정받은 모듈이다[9]. BAM (bottleneck attention module)의 후속 연구로써 SENet (squeeze and excitation Networks)처럼 채널 간의 관계를 살펴 어떤 채널에 집중할 것인지 인코딩하는 채널 어텐션과 전체 픽셀별로 어디에 더 집중할 것인지 인코딩하는 공간 어텐션을 연속적으로 적용하고 원래의 입력되는 특징 맵에 더하여 줌으로써 어텐션 모듈의 성능을 극대화하였다[10], [11].

먼저 채널 간의 집중도를 계산할 때 입력 특징 맵에 Average pooling과 Max pooling을 함께 적용하여 공유 MLP (multi layer perceptron)층을 이용하여 각 특징들을 추출한 후 더한다. 이후 Sigmoid를 적용하여 확률적 형태의 값으로 인코딩을 완료한다. 픽셀 간의 집중도를 파악할 때는 Average pooling과 Max pooling의 결과를 결합하여 2개의 채널로 만들어 낸 후 합성곱 연산을 통하여 기존의 차원과 동일한 형태로 만들어 준다. 마찬가지로 Sigmoid를 적용하여 픽셀 간의 집중도를 확률적 형태의 값으로 인코딩을 진행하는 방식으로 어텐션 모듈을 적용한다.

Ⅲ. 어텐션 기반 특징 융합 네트워크

3장에서는 전체적인 네트워크 구조를 먼저 설명하고 어텐션 기반의 결합 모듈과 카메라-라이다 융합 기법에 대해서 설명한다. 제안하는 모델의 전체 구조를 그림 1에 나타내며 전체 흐름은 다음과 같다. 먼저, 제안하는 결합 모듈을 추가한 3차원 백본 네트워크로 포인트 클라우드 데이터를 처리하여 BEV 특징 맵을 추출한다. 이후 2장에서 설명한 Voxel-RCNN의 흐름을 따른다. 2차원 백본 네트워크와 RPN을 통해 3차원 박스를 예측하고 Voxel-ROI pooling을 통해 3차원 박스 주변의 특징들을 융합한다. 융합한 정보를 Detect head로 전달하여 보정된 3차원 박스를 예측한다. 이후 미리 학습시킨 2차원 객체 탐지 네트워크를 이용하여 이미지 내에서 예측한 객체의 2차원 픽셀 좌표와 라이다 센서의 거리 정보를 이용한 카메라-라이다 융합 기법을 적용하여 거짓 양성을 제거한다.

3-1 어텐션 기반 결합 모듈

3차원 백본 네트워크는 포인트 클라우드 데이터를 받아서 네 개의 희소 합성곱 단을 거치며 BEV 특징 맵을 생성한다. 각 단을 거칠 때마다 절반으로 다운샘플링 하는 과정을 거치게 된다. 다운샘플링을 거칠수록 포인트 클라우드 데이터의 고수준 특징을 추출할 수 있지만 포인트 클라우드 데이터의 3차원 위치 정보는 손실되게 된다. 이에 착안하여 본 연구에서 제안하는 결합 모듈은 3차원 백본 네트워크의 마지막 두 개의 단을 이용한다. 그림 1에서와 같이 마지막 두 개의 단에 대한 각각의 특징 맵과 이 두 개의 특징 맵을 합성곱을 이용하여 융합한 한 개의 특징 맵까지 총 세 가지 특징 맵을 추출한다. 이때 마지막 두 개의 단을 이용하는 이유는 초기 두개의 단에 비하여 마지막 두 개의 단이 다운샘플링이 되어있어 GPU 메모리 소모 측면에서 유리하기 때문이다. 또한, 마지막 단의 고수준 특징과 융합하기

위해서는 저수준 특징을 융합하는 것보다는 비슷한 수준의 특징을 융합하는 것이 효과적이기 때문이다. 세 개의 특징 맵을 적층하여 채널을 합쳐주어 하나의 특징 맵으로 만든다. 이후 합성곱을 적용하여 한 개의 특징 맵의 크기로 채널을 줄인다. 이후 혼합된 특징 맵의 정보를 융합하기 위하여 어텐션 모듈을 도입한다. CBAM 모듈에서 제안하는 채널 어텐션과 공간 어텐션을 도입하여 세 개의 특징 맵의 정보를 융합한다. 이후 CBAM에서 추출한 정보를 입력 특징 맵에 더해준다.

결합 모듈을 추가한 3차원 백본 네트워크를 통해 각 단의 지역적 특징과 더불어 여러 단의 특징을 융합한 전역적 특징을 획득한다. 특히, 3차원 백본 네트워크의 세 번째 단에서 획득한 특징 맵의 경우 마지막 단에서 획득한 특징 맵보다 포인트 클라우드의 공간적인 특징을 더 잘 보존하고 있다. 이 특징에 대한 정보를 기존의 특징과 융합하기 때문에 더욱 정밀한 탐지가 가능하다.

3-2 카메라-라이다 후기 융합 방법

본 연구에서는 미리 학습된 2차원 객체 탐지 네트워크를 이용하여 3차원 객체 탐지의 검출 결과의 신뢰도를 높인다. 그림 2와 같이 2차원 객체 탐지 네트워크로 이미지 내의 검출 결과를 획득한 후 라이다 포인트 클라우드 데이터를 식 (1)의 투영행렬을 이용하여 이 영역에 투영한다[12]. KITTI 데이터 세트의 경우 식 (1)에서와 같이 기준 카메라에 대한 캘리브레이션 정보를 바탕으로 수행한다. 이 때 $P_{lidar\rightarrow ref}$는 라이다 좌표계에서 기준 카메라 좌표계로 변환해주는 투영행렬, $R_{ref\rightarrow rect}$는 기준 카메라의 왜곡을 보정해주는 회전 행렬, $P_{rect\rightarrow cam}$는 카메라 좌표계로 투영해주는 투영행렬이다. 따라서 이 세 행렬을 이용하여 동차 좌표계 시스템에서 식 (1)에서와 같이 변환을 수행한다[12].

$$P_{lidar\rightarrow cam} = P_{rect\rightarrow cam} \cdot R_{ref\rightarrow rect} \cdot P_{lidar\rightarrow ref} \quad (1)$$

이때, 이미지에 투영되는 라이다 포인트 클라우드 데이터의 범위는 검출 박스의 중심에서 너비와 높이의 1/4만큼의 영역에 해당하는 점으로 한다. 각 검출 박스에 대해서 투영된 라이다 포인트 클라우드 데이터의 깊이 값의 평균을 획득하여 이 값을 2차원 객체 탐지 결과에 대한 깊이 값으로 한다. 2차원 객체 탐지 결과의 픽셀 좌표, 깊이 값을 바탕으로 식 (2)의 역투영행렬을 이용하여 3차원 라이다 좌표계로 역 투영하여 2차원 객체 탐지 결과를 3차원 좌표로 변환한다[12].

그림 2. 카메라-라이다 융합을 위한 이미지 검출 결과
Fig. 2. Image detection for camera fusion

$$P_{cam\rightarrow lidar} = P^{-1}_{lidar\rightarrow ref} \cdot R^{-1}_{ref\rightarrow rect} \quad (2)$$

2차원 객체 탐지 네트워크를 통해 획득된 탐지 결과들의 3차원 좌표를 3차원 객체 탐지 결과 박스의 중심 좌표와 비교한다. 3차원 객체 탐지 결과가 2차원 객체 탐지 결과로부터 얻은 임계범위 안에 들어오지 않을 경우 결과는 거짓 양성으로 간주하여 제거한다.

Ⅳ. 실험 및 고찰

4-1 데이터 세트

본 실험의 결과는 기존의 Voxel-RCNN 네트워크와 제안된 모델의 성능 비교를 위해 자율주행 분야 연구에서 자주 이용되는 KITTI 데이터 세트를 이용하여 진행하였다[12]. 본 연구에서는 KITTI 데이터 세트의 7,481개를 기존 연구를 따라 3,712개의 학습 데이터와 3,769개의 검증 데이터로 나누었다. KITTI 데이터 세트의 검증 데이터에 대하여 평가를 진행하였고 차량 클래스에 대하여 Hard, Moderate, Easy로 나누어서 평가를 진행하였다. 평가 방식은 기존 연구와 같이 11개의 재현율 위치를 이용한 AP (average prevision) 평가 방식을 따랐다[3].

4-2 실험 환경

본 연구에서 이용한 3차원 백본 네트워크의 채널은 각각 16, 32, 64, 64레벨의 차원을 이용하였다. 그림 1에서와 같이 3차원 백본 네트워크의 마지막 두 개의 단에서 추출하는 특징 맵은 각각 2차원 합성곱 연산과 형태 변경을 통해 256차원을 지닌 특징 맵으로 변경된다. 이 두 개의 특징 맵을 이용한 세 번째 특징 맵의 형태도 합성곱 연산과 형태 변경을 통해 256차원을 가진다. 어텐션을 적용한 후의 특징 맵의 형태 역시 256차원이다. 실험을 위한 복셀 크기는 x, y 방향의 경우 0.05 m, z 방향의 경우 0.1 m로 실험하였다. 3차원 객체 탐지 네트워크의 경우 80 Epoch으로 학습시켰다.

2차원 객체 탐지 네트워크로는 Yolo v5 (you only look once) 모델을 이용하였고 학습 데이터는 3,712개의 학습 데이터를 이용하였다[13]. 120 Epoch으로 미리 학습시킨 모델을 이용하여 3차원 객체 탐지 네트워크의 거짓 양성을 제거한다. 본 연구는 Linux Ubuntu 18.04에서 수행되었으며 GPU는 RTX 3090를 이용하였다.

4-3 비교 평가

KITTI 데이터 세트의 차량 클래스에 대한 3차원 객체 탐지 정확도를 기존 연구와 비교한 결과를 표 1에 나타낸다.

표 1. KITTI 검증 데이터 세트의 차량에 대한 AP11 성능 비교 (AP11은 11개의 재현율 위치를 이용하여 계산)

Table 1. AP11 Performance comparison on the KITTI validation dataset for car class (AP11 calculated by 40 recall positions)

Method	Car $AP_{3D}(\%)$		
	Hard	Moderate	Easy
Point-based:			
Point-RCNN[14]	77.38	78.63	88.88
3DSSD[15]	78.67	79.45	89.71
PV-RCNN[16]	78.70	83.69	89.35
Voxel-based:			
VoxelNet[17]	62.85	65.46	81.97
PointPillars[18]	68.91	76.06	86.62
TANet[19]	68.91	76.64	87.52
SECOND[20]	77.22	78.62	88.61
$Part-A^2$[21]	78.54	79.47	89.47
Voxel-RCNN[3]	78.75	84.05	89.31
SA-SSD[22]	78.78	79.91	**90.15**
Proposed (w/o Late-fusion)	78.91	84.28	89.67
Proposed	**78.97**	**84.59**	89.67

표 2. KITTI 검증 데이터 세트의 차량에 대한 AP40 성능 비교 (AP40은 40개의 재현율 위치를 이용하여 계산)

Table 2. AP40 Performance comparison on the KITTI validation dataset for car class (AP40 calculated by 40 recall positions)

Method	Car $AP_{3D}(\%)$			Car $AP_{BEV}(\%)$		
	Hard	Mod	Easy	Hard	Mod	Easy
Voxel-RCNN	82.59	84.92	92.10	**90.55**	91.11	95.32
Proposed	**82.92**	**85.39**	**92.58**	89.11	**91.41**	**95.73**

표 1에서 결합 모듈을 적용한 모델이 기존의 Voxel-RCNN보다 Hard, Moderate, Easy 모든 부분에서 AP가 상승한 것을 확인할 수 있다. 논문[22]보다는 Easy 부분에서는 낮은 성능을 보였지만, 난이도가 높은 Hard, Moderate 부분에서 제안하는 알고리즘이 보다 높은 성능을 보였다. 이를 통해 결합 모듈이 3차원 백본 네트워크의 다중 스케일의 특징을 효과적으로 융합했다고 평가할 수 있다. 또한, 카메라-라이다 후기 융합의 결과 추가적인 AP의 상승을 통해 라이다 센서로부터 멀리 떨어진 작은 물체에 대한 희박한 포인트 클라우드 데이터로 인한 거짓 양성을 효과적으로 제거했다는 것을 확인할 수 있다. 표 2는 AP40에 대한 성능 평가 표이다. 표 2는 3D 검출 성능 뿐만 아니라 제안하는 알고리즘은 BEV 검출 성능도 기존 연구에 비해 향상됨을 보인다.

제안하는 결합 모듈 적용에 따른 정성적 비교 평가를 그림 3에 보인다. 그림 3의 입력 이미지의 좌측 하단의 물체를 기존의 Voxel-CNN은 차량으로 잘못 인식했지만, 결합 모듈을 추가한 제안된 모델에서는 차량으로 인식하지 않은 것을 확인할 수 있다.

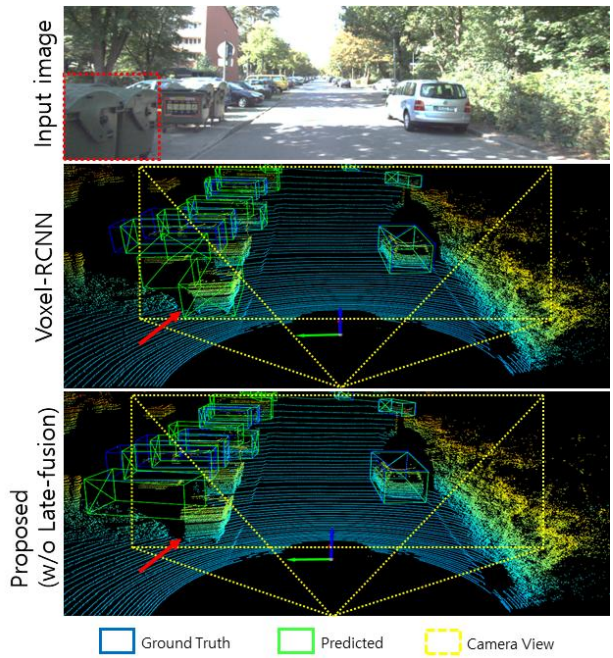

그림 3. 결합 모듈 적용에 따른 정성적 비교 평가

Fig. 3. Qualitative comparative evaluation of integrating module

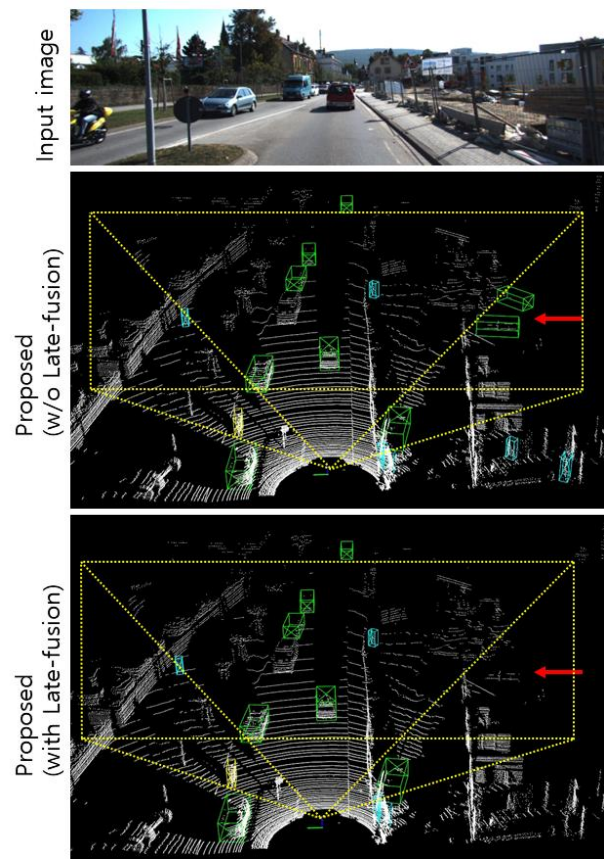

그림 4. 후기 융합 적용에 따른 정성적 비교 평가

Fig. 4. Qualitative comparative evaluation of late fusion

마지막으로 제안하는 후기 융합 적용에 따른 정성적 비교 평가를 그림 4에 보인다. 2차원 객체 탐지 결과에 따라 3차원 객체 탐지 결과의 거짓 양성을 제거하며 화살표 표시의 오탐지 객체가 제거됨을 확인한다.

Ⅴ. 결 론

본 연구에서는 기존의 그리드 기반 3차원 객체 탐지 네트워크를 개선하기 위하여 3차원 백본 네트워크의 다중 스케일의 특징을 효과적으로 융합하는 결합 모듈을 제안하였다. 제안된 결합 모듈을 통하여 3차원 백본 네트워크의 고수준 특징과 포인트 클라우드 데이터의 공간적인 특징을 효과적으로 융합하여 더 정확한 3차원 박스 제안을 가능하게 한다. 또한, 미리 학습시킨 2차원 객체 탐지 네트워크를 활용하여 이미지에서의 2차원 객체 탐지 결과와 라이다 센서의 거리 정보를 융합하여 3차원 객체 탐지 결과의 신뢰도를 높이는 방법을 제안하였다. 제안된 카메라-라이다 융합 방식을 통해 먼 거리의 부족한 포인트 클라우드 데이터로 인해 잘못 탐지되는 거짓 양성을 줄일 수 있다. KITTI 데이터 세트를 이용한 실험 결과에서 3차원 객체 탐지뿐만 아니라 BEV 상의 2차원 객체 탐지에서도 AP 상승을 통해 본 연구에서 제안하는 기법의 유용성을 증명하였다.

본 연구에서는 카메라-라이다의 후기 융합 방식을 이용하여 이미지 융합을 진행하였다. 추후 두 센서 데이터 간의 교차 양식 정보를 효과적으로 융합할 수 있는 깊은 융합 방식을 후기 융합 방식에 결합하여 보다 나은 성능 향상을 유도할 예정이다.

Acknowledgments

본 연구는 경기도 지역협력 연구센터 사업의 일환으로 수행하였음. [GRRC항공2017-B04, 지능형 인터랙티브 미디어 및 공간 융합 응용 서비스 개발]

References

[1] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby and A. Mouzakitis, “A Survey on 3D Object Detection Methods for Autonomous Driving Applications,” *IEEE Transactions on Intelligent Transportation Systems,* Vol. 20, No. 10, pp. 3782-3795, Oct. 2019.

[2] J. Mao, S. Shi, X. Wang, and H. Li, “3D object detection for autonomous driving: a review and new outlooks,” arXiv preprint arXiv:2206.09474, 2022. [Online]. Available: https://arxiv.org/abs/2206.09474

[3] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, “Voxel r-cnn: Towards high performance voxel-based 3d object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver: Canada, pp. 1201-1209, 2021.

[4] B. Graham, “Sparse 3D convolutional neural networks,” arXiv preprint arXiv:1505.02890, 2015. [Online]. Available: https://arxiv.org/abs/1505.02890

[5] C. He, H. Zeng, J. Huang, X. Hua, and L. Zhang, “Structure aware single-stage 3d object detection from point cloud,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle: WA, pp. 11873-11882, 2020.

[6] W. Liang, P. Xu, L. Guo, H. Bai, Y. Zhou, and F. Chen, “A survey of 3D object detection,” *Multimedia Tools and Applications*, Vol. 80, No. 19, pp. 29617-29641, Aug. 2021.

[7] R. Qian, X. Lai, and X. Li, “3D object detection for autonomous driving: a survey,” *Pattern Recognition*, Vol. 130, 2022.

[8] S. Pang, D. Morris, and H. Radha, “CLOCs: Camera-LiDAR object candidates fusion for 3D object detection,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas: NV, pp. 10386-10393, 2020.

[9] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *European Conference on Computer Vision*, Munich: Germany, pp. 3-19, 2018.

[10] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, “Bam: Bottleneck attention module,” arXiv preprint arXiv:1807.06514, 2018. [Online]. Available : https://arxiv.org/abs/1807.06514

[11] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City: UT, pp. 7132-7141, 2018.

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, Vol. 32, No. 11, pp. 1231-1237, 2013.

[13] Ultralytics-Yolov5 [Online]. Available: https://doi.org/10.5281/ZENODO.7347926

[14] S. Shi, X. Wang, and H. Li, “Pointrcnn: 3d object proposal generation and detection from point cloud,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 770-779, Long Beach: CA, 2019.

[15] Z. Yang, Y. Sun, S. Liu, and J. Jia, “3dssd: Point-based 3d single stage object detector,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle: WA, pp. 11040-11048, 2020.

[16] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle: WA, pp. 10529-10538, 2020.

[17] Y. Zhou, and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, pp. 4490-4499, 2018.

[18] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Long Beach: CA, pp. 12697-12705, 2019.

[19] Z. Liu, X. Zhao, T. Huang, R. Hu, Y. Zhou, and X. Bai, "Tanet: Robust 3d object detection from point clouds with triple attention," in *Proceedings of the AAAI conference on artificial intelligence*, New York: NY, pp. 11677-11684, 2020.

[20] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, Vol. 18, No. 10, pp. 3337, Oct. 2018.

[21] S. Shi, Z. Wang, X. Wang, and H. Li, "Part-aˆ 2 net: 3d part-aware and aggregation neural network for object detection from point cloud," arXiv preprint arXiv:1907.03670 2.3, 2019. [Online]. Available: https://arxiv.org/abs/1907.03670

[22] C. He, H. Zeng, J. Huang, X. S. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud." in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle: WA, pp. 11873-11882, 2020.

유 상 현 (Sang-Hyun Ryoo)
2023년 2월: 한국항공대학교 기계공학 (공학사)
※관심분야 :영상처리, 컴퓨터 비전, 자율주행

강 대 열 (Dae-Yeol Kang)
2023년 2월: 한국항공대학교 전자공학 (공학사)
※관심분야 : 컴퓨터 비전, SLAM, 로보틱스

황 승 준 (Seung-Jun Hwang)
2012년 2월 : 한국항공대학교 정보통신공학 (공학사)
2014년 2월 : 한국항공대학교 일반대학원 정보통신공학 (공학석사)
2022년 8월 : 한국항공대학교 일반대학원 항공전자정보공학(공학박사)
2022년 9월 ~ 현재 : 영상음향공간 융합기술 연구센터 연구원
※관심분야 : 영상처리, 컴퓨터비전, 패턴인식

박 성 준 (Sung-Jun Park)
2019년 2월 : 한국항공대학교 전자및항공전자공학 (공학사)
2021년 2월 : 한국항공대학교 일반대학원 항공전자정보공학 (공학석사)
2021년 ~ 현재 : 한국항공대학교 일반대학원 항공전자정보공학 박사과정
※관심분야 : 컴퓨터비전, 영상처리

백 중 환 (Joong-Hwan Baek)
1981년 2월 : 한국항공대학교 항공통신공학 (공학사)
1987년 7월 : 오클라호마주립대학원 전기 및 컴퓨터공학 (공학석사)
1991년 7월 : 오클라호마주립대학원 전기 및 컴퓨터공학 (공학박사)
1992년 ~ 현재 : 한국항공대학교 항공전자정보공학부 교수
※관심분야 : 영상처리, 패턴인식, 멀티미디어, 가상현실