자율주행 환경에서 카메라 기반 3D 객체 탐지를 위한 Depth 추정 방식 비교 분석

김가현[1], 이승훈[2], 김경현[3], 김영빈[2,3]
[1]중앙대학교 예술공학부
[2]중앙대학교 첨단영상대학원, [3]중앙대학교 AI학과
e-mail : kahyun0817@cau.ac.kr, remomare@cau.ac.kr, khyun8072@cau.ac.kr, ybkim85@cau.ac.kr

A Survey on Depth Estimation Methods for Camera-Based 3D Object Detection in Autonomous Driving

Kahyeon Kim[1], Seunghoon Lee[2], Kyeonghyun Kim[3], Youngbin Kim[2,3]
[1]College of Art & Technology, School of Computer Art, Chung-Ang University
[2]Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University
[3]Department of Artificial Intelligence, Chung-Ang University

Abstract

Camera-only 3D object detection has gained significant attention as a cost-effective alternative to LiDAR-based methods in autonomous driving. Unlike LiDAR, cameras lack direct depth sensing, making depth estimation a key challenge. This paper surveys recent methods in this field, categorized into depth-free, depth-supervised, and structure-regularized approaches. We compare representative models from each category on benchmarks like nuScenes, analyzing their strengths and limitations. While depth-supervised and structure-aware models offer improved spatial precision, they still struggle with long-range accuracy and object-level consistency. To address these limitations, we discuss potential directions such as depth refinement and multi-sensor fusion. Through this analysis, we aim to derive model design directions for achieving robust detection performance in autonomous driving environments.

I. 서론

3D 객체 탐지는 입력 데이터로부터 물체의 범주와 함께 3차원 공간상의 위치, 크기, 방향 등을 추론하는 컴퓨터 비전 과제이다. 특히, 자율주행 시나리오에서는 도로 환경을 실시간으로 인식하고 주행 결정을 내려야 하기 때문에 3D 객체 탐지는 자율주행 인지 시스템의 핵심 모듈로 간주된다. 2D 객체 탐지는 이미지 내 물체의 위치만을 2차원 상에서 표현하는 데 그치지만, 3D 객체 탐지는 실제 세계 공간에서의 물체의 거리, 방향, 위치 등을 포함한 보다 정밀한 인식을 목표로 한다 [1].

레이저를 이용해 주변 물체까지의 거리를 정밀하게 측정하는 LiDAR 기반 탐지 모델은 고정밀 깊이 정보를 제공한다. 하지만 이러한 시스템은 비용이 높고, 차선이나 교통 신호와 같은 시각적 요소를 포착하는 데에는 한계가 있다 [2], [3].

이러한 한계로 인해, 최근에는 RGB 이미지만을 활용하는 camera-only 방식이 주목받고 있다. RGB카메라는 비용이 저렴하고 널리 보급되어 있으며 풍부한 시각 정보를 제공한다는 장점이 있지만, 단일

시점의 2D 이미지에서는 물체까지의 거리를 직접적으로 측정할 수 없어 정밀한 3D 위치 추정에는 구조적 제약이 따른다. 따라서 정확한 3D 위치 추정을 위해서는 2D 이미지로부터 장면의 구조적 정보를 복원하는 depth 추정 과정이 필수적이다. 이 과정은 2D 이미지의 픽셀만으로 3D 공간을 복원해야 하므로 본질적으로 ill-posed 문제에 해당한다. 또한, 예측된 depth 정보의 정확도는 최종 3D 객체탐지 성능에 결정적인 영향을 미친다.

본 논문에서는 최근 camera-only 기반 3D 객체 탐지 기법들을 정리하고, 각 기법들이 depth 추정과 3D 공간 표현을 어떻게 처리하는지에 대해 분석하고자 한다.

II. 본론

자율주행 시나리오에서 camera-only 기반 3D 객체 탐지의 성능은 depth 추정 방식에 따라 크게 달라진다. 최근 연구들은 depth 정보를 처리하는 다양한 전략을 제안하고 있다. 일부는 depth 정보를 명시적으로 예측하거나 구조적인 공간 제약을 활용한다. 혹은 별도의 depth supervision 없이도 3D 공간 정보를 학습하려는 접근을 취한다. 본 절에서는 이러한 모델들을 depth-free, depth-supervised, structure-regularized의 세 가지 범주로 나누어 분석하고자 한다.

2.1 Depth-free 방식

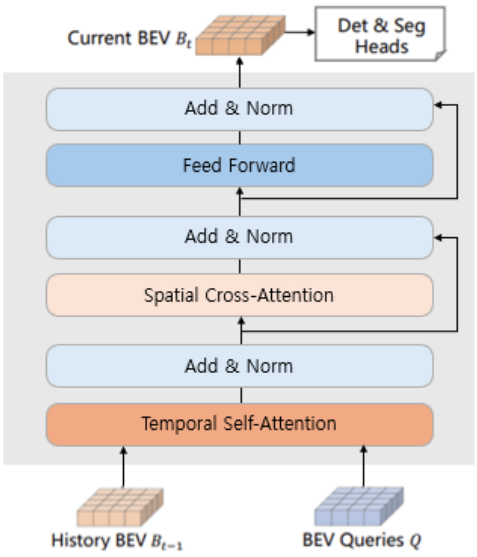

그림 1. Depth-free 방식의 대표 모델인 BEVFormer의 전체 구조

Depth-free 방식은 명시적인 depth map을 예측하지 않고, 이미지 feature에서 직접 BEV 표현을 추출하는 방식이다.

BEVFormer[4]는 대표적인 depth-free 모델로 BEV query와 cross-attention을 통해 이미지 feature에서 BEV representation을 생성한다. 이때 각 query는 자율주행 차량을 중심으로 설정된 BEV 평면의 격자 단위로 표현되며 cross-attention을 통해 특정 공간 영역에서 관련 정보를 집중적으로 추출한다. 또한 temporal self-attention을 도입해 연속된 시점의 BEV feature 간 관계를 학습하며 이를 통해 움직이는 객체의 위치 추정 및 가려짐 상황에서 인식 성능을 향상시킨다.

BEVFormer는 camera-only 기반 모델임에도 불구하고 LiDAR 방식의 모델에 근접하는 3D 객체 탐지 성능을 기록하며 이후 다양한 후속 depth-free 모델들의 기반이 되었다. 유사하게 SparseBEV[5] 또한 명시적인 depth supervision 없이 BEV 표현을 학습한다. BEV 평면 전체가 아닌 일부 중요한 위치에만 쿼리를 배치한 sparse query를 통해 이미지 feature를 샘플링한다. 이후 샘플링한 이미지 feature를 바탕으로 필요한 정보만 선택적으로 추출해 BEV 표현을 구성한다. 그러나 depth-free 방식은 정확한 depth alignment 없이 attention만으로 feature를 통합하기 때문에 기하학적 일관성이 떨어지고, 객체의 위치나 경계 표현에서 오류가 발생할 가능성이 크다.

2.2 Depth-supervised 방식

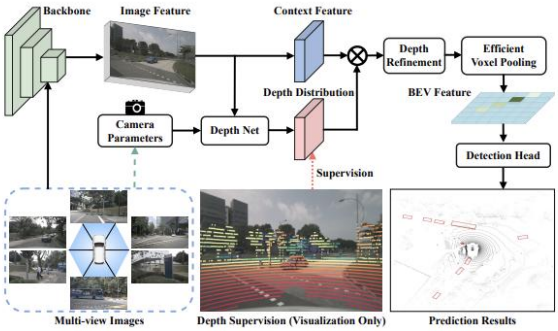

그림 2. Depth-supervision 방식의 대표 모델인 BEVDepth의 구조

Depth-supervised 방식은 학습 시 명시적인 depth supervision을 통해 이미지로부터 신뢰도 높은 depth 정보를 예측하고 이를 기반으로 3D 표현을 정렬하는 방식이다.

BEVDepth[6]는 대표적인 depth-supervised 모델로 Lift-Splat-Shoot(LSS) 구조를 기반으로 하여 이미지 feature를 frustum 공간으로 끌어올리고,

이를 BEV 평면에 정렬하여 3D feature map을 구성한다 [6]. 이때 사용되는 depth supervision은 LiDAR로부터 생성된 pseudo depth label이며, 이는 depth 추정의 안전성을 높이는 데 기여한다. 또한 BEVDepth는 view transformation 과정에서 발생하는 오류를 줄이기 위해 depth refine 모듈을 도입하여 부정확하게 투영된 feature의 위치를 보정한다.

BEVNeXt[7]는 BEVDepth를 기반으로 발전된 모델로 depth 추정의 정확도를 향상시키기 위해 CRF(Conditional Random Field)를 활용한다. CRF는 인접 픽셀 간의 관계를 고려하여 depth의 일관성을 높이며 고해상도 feature map에서 depth를 추정하여 라벨 손실을 줄인다. 또한 perspective refinement 모듈을 도입하여 객체 주변의 BEV feature를 backward projection 기반 attention으로 정교하게 보정한다.

이러한 방식은 depth 추정의 정밀도를 높이는 데 유리하지만 pseudo label 품질에 의존하며 연산량이 증가하는 단점이 있다. 그리고 LSS 방식을 기반으로 하기 때문에 pixel-to-depth 추정의 구조라는 특성상 원거리 객체에서는 depth 정확도가 감소할 수 있다는 한계도 존재한다.

Far3D[8]는 LSS기반 구조에서 벗어나 장거리 탐지 3D 객체 검출 성능 향상을 목표로 설계된 모델이다. 2D detector와 depth 분류 네트워크를 활용해 2D proposal과 pseudo-depth를 예측한 뒤, 객체가 실제로 존재할 가능성이 높은 위치를 추정해 3D query로 변환한다. 생성된 query는 기존 global query와 함께 transformer decoder에 입력되어 3D 객체를 예측하며 이 과정에서 depth supervision이 query의 초기 위치 정렬과 정확한 공간 추정에 기여한다.

2.3 Structure-regularized 방식

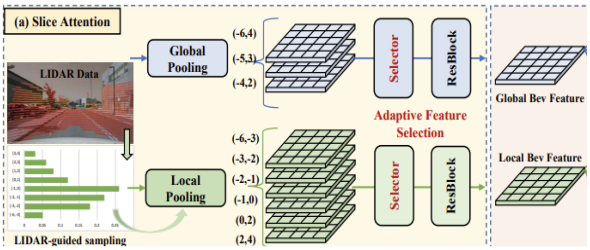

그림 3. Structure-regularized 방식의 대표 모델인 BEV-SAN의 slice attention 구조

Structure-regularized 방식은 depth 정보를 직접 예측하기보다는 공간 구조적 제약을 활용하여 BEV feature의 정밀도를 향상시키는 접근이다. 이는 객체의 높이, 위치, 주변 구조 등 local spatial 관계를 모델링하여 3D 표현을 보완한다.

3D-ASAP prior 기반의 접근은 LEGO[9]에서 제안되었으며, 이는 평면 구조에 대한 사전 지식을 정규화 제약을 사용하여 명시적으로 부여한다. 이 방식은 거리 기반의 3D 객체 탐지 문제에서 도로, 건물, 차량 지붕 등 대부분의 장면이 평면이라는 사실에 착안하여 모델이 불필요한 depth의 미세한 진동을 줄이고 실제 장면에 가까운 일관된 depth를 예측하도록 유도한다. LEGO는 3D 객체 탐지 모델은 아니지만, 이후 structure-regularized 방식에서 평면성, 기하 정합성 등의 개념적 기반으로 자주 인용되며 이후 3D 객체 탐지 연구에 큰영향을 미친다.

BEV-SAN[10]은 BEV 공간을 높이 기준으로 slicing하여 전역 및 지역 구조 정보를 통합한다. 전역 slicing은 높은 범위의 의미적 정보를 포착하고, 지역 slicing은 세밀한 높이 분포를 강조한다. 이후 slice feature는 SE-like attention을 통해 통합되며 transformer 기반의 양방향 attention 구조를 통해 구조 정보를 보강한다.

GeoBEV[11] 또한 structure-regularized 방식의 범주에 속하며 정확하고 조밀한 BEV 표현을 학습하기 위한 구조 정규화 기법들을 통합한다. Radial-Cartesian BEV Sampling을 적용하여 방사형으로 펼쳐진 feature를 bilinear sampling 방식으로 변환함으로써 BEV 평면 전체에 고르게 채워지도록 한다. 또한 In-Box Label을 도입하여 객체의 표면이 아닌 내부 구조 전체를 depth supervision 대상으로 설정함으로써, BEV 표현에 실제 장면의 기하 구조를 정밀하게 반영할 수 있도록 유도한다.

이와 같은 공간 구조 정규화 기반의 방식은 local geometry를 기준으로 feature를 분할 처리하기 때문에 동일 객체라도 서로 다른 구조 단위로 처리된다. 그 결과 BEV 표현의 기하학적 일관성을 강화하지만, 객체의 의미적 경계를 명확히 구분하거나 객체 단위의 일관된 공간 표현을 구성하는 데에는 한계가 있다.

2.4 데이터셋과 평가 지표

앞서 소개한 camera-only 기반 모델들은 자율주행 3D 객체 탐지 벤치마크 데이터셋인 nuScenes 데이터셋[12]을 사용하여 성능을 검증하였다. nuScenes는 총 1,000개의 시나리오(scene)로 구성되어 있으며, 각 시나리오는 약 20초 분량의 주행 데이터를 포함하고 있다. 데이터는 전방, 후방, 좌우 측면을 포함한 총 6대의 서라운드 뷰 카메라와 1개의 LiDAR, 5개의 레이더 센서로부터 수집된다. 대부분의 camera-only

방식	연구	입력 방식	mAP(%)↑	NDS(%)↑	데이터 셋
depth-free	BEVFormer [4]	RGB only	48.1	56.9	nuScenes
	SparseBEV [5]	RGB only	47.4	67.5	nuScenes
depth-supervised	BEVDepth [6]	RGB + pseudo-depth	52.0	60.9	nuScenes
	BEVNeXt [7]	RGB + pseudo-depth	44.4	51.3	nuScenes
	Far3D [8]	RGB + pseudo-depth	53.0	68.7	nuScenes
structure-regularized	BEV-SAN [9]	RGB only	35.1	48.2	nuScenes
	GeoBEV [10]	RGB only	57.9	66.2	nuScenes

표1. 제안된 자율주행 시나리오 하 Camera Only 3D Object Detection 모델의 성능 비교

방식은 이 중 RGB 카메라 이미지만을 입력으로 사용하며, 일부 모델에서는 LiDAR로부터 생성된 pseudo depth label을 학습 supervision으로 활용한다 [12].

각 연구에서 주요 평가 지표로는 mean Average Precision (mAP)과 nuScenes Detection Score (NDS)[12]를 사용하였다. mAP는 객체 클래스별로 precision-recall 곡선을 기반으로 평균 정밀도를 계산하는 지표로 객체의 탐지 정확도를 측정하는 데 사용된다. NDS는 mAP뿐만 아니라 위치 정확도, 크기, 방향, 속도, 속성 등 다양한 항목을 함께 고려한 nuScenes 전용 종합 평가 지표로 전반적인 3D 객체 탐지 성능을 평가할 수 있다.

NDS에서 반영하는 다섯 가지 평균 오류 지표에 대한 각각의 항목은 다음과 같다. mATE(mean Average Translation Error)는 예측된 객체의 중심 위치와 실제 위치 간의 평균 오차를 나타내며 값이 작을수록 정확한 위치 예측을 의미한다. mASE(mean Average Scale Error)는 객체 크기의 예측 오류이고, mAOE (mean Average Orientation Error)는 객체의 방향 예측 정확도를 나타내는 지표이다. mAVE (mean Average Velocity Error)는 객체 속도 예측의 평균 오차를 나타내고 mAAE (mean Average Attribute Error)는 객체의 정지 여부 등 속성 정보의 분류 정확도를 측정한다. 이 다섯 가지 항목은 NDS 계산식 내에서 TP(True Positive) 지표들의 집합으로 사용된다.

2.5 기존 연구 제안 모델 성능 비교

본 논문에서는 camera-only 기반 3D 객체 탐지 모델들을 depth 추정 방식에 따라 depth-free, depth-supervised, structure-regularized의 세 가지 범주로 나누어 분석했다.

Depth-free 방식에서는 BEVFormer[4]가 48.1% mAP, 56.9% NDS를 기록하였으며, SparseBEV5[5]는 mAP에서는 유사한 수준인 47.4%을 유지하면서도 67.5%로 더 높은 NDS를 달성하여 모션 표현 능력에서도 일정 수준 이상의 정밀도를 보였다. Depth-supervised 계열에서는 pseudo-depth supervision을 활용한 BEVDepth[6]가 52.0% mAP, 60.9% NDS를 기록하며 BEVFormer[4] 대비 높은 정확도를 보였고, Far3D[8]는 53.0% mAP와 68.7% NDS로 가장 높은 성능을 나타냈다.

Structure-regularized 방식의 GeoBEV[11]는 57.9% mAP, 66.2% NDS로 가장 정밀한 BEV 표현을 달성하였으며, 이는 Radial-Cartesian BEV Sampling과 In-Box Label을 통한 구조 정규화 효과로 볼 수 있다.

Ⅲ. 결론 및 향후 연구 방향

본 논문에서는 camera-only 기반의 3D 객체 탐지 모델들을 중심으로 각 모델의 depth 추정 방식과 BEV 표현 전략 등을 비교·분석하였다. 이 모델들의 성능은 depth 추정의 정확도에 크게 의존하며, 이를 위한 다양한 정규화 및 보정 기법들이 도입되어 왔다.

하지만 camera-only 방식은 본질적으로 RGB 이미지에서 직접적인 거리 정보를 얻을 수 없어, depth 추정의 불확실성, 기하학적 왜곡 등의 근본적인 한계를 지닌다. 특히 카메라로부터 멀리 떨어진 객체나 가려진 상황에서 정보가 부족해 예측이 불안정해지는 경향이 있다.

최근 연구에서는 camera와 LiDAR, radar 등의 센서를 결합하여 서로의 약점을 보완하는 multi-sensor fusion 모델이 주목받고 있다. 특히 camera 기반 BEV 구조에 LiDAR feature나 calibration-free alignment를 결합하는 BEVFusion[13], GraphBEV [14] 등의 연구는 단일 센서 기반 탐지의 한계를 극복하고자 한다. 따라서 depth 추정은 3D 객체 탐지의 핵심 과제이자 camera-only 방식의 성능 한계를 극복하기 위한 중심 연구 축으로 작용하고 있으며 향후에는 depth 추정 정밀도 향상뿐 아니라 연속적인 프레임 간의 시간·구조적 일관성까지 고려하는 연구가 필요하다.

Acknowledgements

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2021-II211341, 인공지능대학원지원(중앙대학교))을 받아 수행된 연구임.

참고문헌

[1] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang and Hongsheng Li, "3D Object Detection for Autonomous Driving: A Comprehensive Survey," in IJCV, 2023

[2] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, Oscar Beijbom, "PointPillars: Fast encoders for object detection from point clouds", In CVPR, 2019.

[3] Tianwei Yin, Xingyi Zhou, Philipp Krahenbuhl, "Center-based 3D object detection and tracking", In CVPR, 2021.

[4] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, Jifeng Dai, "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers", In ECCV, 2022.

[5] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, Limin Wang, "SparseBEV: High-Performance Sparse 3D Object Detection from Multi-Camera Videos", In ICCV, 2023.

[6] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, Zeming Li, "BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection", In AAAI,

[7] Zhenxin Li, Shiyi Lan, Jose M. Alvarez, Zuxuan Wu, "BEVNeXt: Reviving Dense BEV Frameworks for 3D Object Detection", In CVPR, 2024.

[8] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, Xiangyu Zhang, "Far3D: Expanding the Horizon for Surround-view 3D Object Detection", In AAAI, 2024.

[9] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "LEGO: Learning edge with geometry all at once by watching videos," In CVPR, 2018.

[10] Xiaowei Chi, Jiaming Liu, Ming Lu, Rongyu Zhang, Zhaoqing Wang, Yandong Guo, Shanghang Zhang, "BEV-SAN: Accurate BEV 3D Object Detection via Slice Attention Networks", In CVPR, 2023.

[11] Jinqing Zhang, Yanan Zhang, Yunlong Qi, Zehua Fu, Qingjie Liu, Yunhong Wang, "GeoBEV: Learning Geometric BEV Representation for Multi-view 3D Object Detection", In AAAI, 2025.

[12] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, Oscar Beijbom, "nuScenes: A multimodal dataset for autonomous driving", In CVPR, 2020.

[13] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," In CVPR, 2023.

[14] Z. Song, L. Yang, S. Xu, L. Liu, D. Xu, C. Jia, F. Jia, and L. Wang, "GraphBEV: Towards robust BEV feature alignment for multi-modal 3D object detection," In CVPR, 2024.