주파수 변이 인지 및 Mamba 아키텍처를 활용한 정밀 3 차원 공간 객체 탐지

이영생, 조인휘*
한양대학교, *한양대학교

twlight@hanyang.ac.kr, *iwjoe@hanyang.ac.kr

Precise 3D spatial object detection using frequency-variant recognition and Mamba architecture.

Lee Rong Sheng, Joe In Whee*
Hanyang Univ., * Hanyang Univ.

요 약

본 논문은 자율주행 및 로보틱스 분야에서 정밀한 인지를 가능케 하는 멀티모달 3D 객체탐지의 효율성을 극대화하고자 새로운 프레임워크를 제안하였다. 기존 융합 파이프라인이 직면한 멀티스케일 피처 융합의 한계와 스펙트럴 불일치 문제를 해결하기 위해, 멀티디렉셔널 선택적 스캐닝 기반의 혼합형 Mamba 아키텍처와 주파수 인지 융합을 위한 혼합형 주파수 모듈을 설계하였다. 벤치마크 데이터셋을 활용한 실험 결과, 제안 모델은 기존 소타(SOTA) 알고리즘을 상회하는 높은 탐지 스코어와 정밀도를 달성하였으며, 어블레이션 스터디를 통해 각 컴포넌트가 멀티모달 데이터 정렬 및 전체적인 성능 향상에 유효함을 입증하였다.

Ⅰ. 서 론

본 논문에서는 자율주행 및 로보틱스 시스템의 안전성을 보장하기 위해 카메라와 라이다 센서의 상호보완적 특성을 극대화하는 멀티모달 3D 객체탐지 기술을 다루고 있다. 카메라의 풍부한 텍스처 정보와 라이다의 정밀한 기하학적 데이터를 결합하는 과정에서, 입력 특성이 서로 다른 두 센서의 정보를 일관되게 정렬하고 통합하는 것이 핵심 과제로 제시된다. 특히 기존 융합 패러다임은 멀티스케일 특징 융합의 비효율성과 센서 간 스펙트럴 불일치로 인한 정렬 오류 문제를 겪고 있어, 실제 환경에서 안정적인 성능을 확보하는 데 제약이 존재한다 [1, 2].

이러한 한계를 극복하기 위해 본 논문에서는 선형 시간 복잡도를 보장하는 선택적 상태공간모델(SSM) 기반의 새로운 아키텍처를 제안하고 있다. 특히 이방성 롱-레인지 컨텍스트 전파를 위한 멀티디렉셔널 스캐닝 전략과 센서 간 주파수 특성을 조화시키는 적응형 주파수 대역 통합 메커니즘을 도입하여, 융합 과정에서 발생하는 불필요한 간섭을 줄이고 바운딩 박스 추정의 정밀도를 강화하고 있다[4, 5, 6, 7, 8, 9, 10].

최종적으로 nuScenes 벤치마크에서 기존 소타(SOTA) 알고리즘을 상회하는 성능을 달성함으로써, 실시간 자율주행 시스템에 적용 가능한 실용적 해결방법을 제시하고 있다. 더불어 실험 결과를 통해 제안 기법의 유효성을 정성·정량 관점에서 뒷받침하며, 단순한 성능 보고를 넘어 정확도와 효율의 균형을 어떻게 유지할 수 있는지에 대한 논의를 함께 포함한다.

또한 다양한 조건에서의 적용 가능성을 염두에 두고, 구성 요소별 역할과 전체 파이프라인에서의 동작 특성을 정리함으로써 향후 확장 방향에 대한 시사점을 제공한다. 즉, 현실적 제약을 고려한 설정에서도 안정적인 개선이 가능하다는 점을 강조하면서, 관련 멀티모달 3D 인지 시스템 설계에 참고가 될 수 있는 전반적 의미를 함께 정돈한다[3].

Ⅱ. 본론

본논문에서는 카메라- 라이다 멀티모달 3D 객체 검출 분야에서 초기/중간/후기 융합부터 BEV 기반 트랜스포머 계열까지 다양한 접근이 제안되어 왔음을 정리한다[4, 11, 12]. 이러한 흐름은 서로 다른 센서가 제공하는 단서를 하나의 프레임워크 안에서 통합하려는 지속적인 시도이며, 복잡한 주행 장면에서 안정적인 인지를 달성하기 위한 기반으로 이해될 수 있다. 특히 모달리티별 강점을 결합해 인지 범위를 확장하고, 다양한 상황에서의 강건성을 높이려는 방향으로 연구가 전개되어 왔다는 점을 함께 부각한다.

그러나 실제 환경의 지연 시간 제약을 고려하면 장거리 컨텍스트를 충분히 모델링하기 어렵고, 방향성 구조를 반영하는 데에도 한계가 남는다[13, 14, 15]. 또한 윈도우 기반 어텐션이 갖는 국소성 제한과 모달 간 스펙트럴 불일치 문제가 지속되어, 멀티스케일 특징을 취합하는 과정에서 정보가 편향되거나 불필요한 간섭이 누적될 수 있다. 이러한 배경에서 SSM·선형 어텐션 계열인 Mamba 의 선택적 스캐닝을 멀티스케일·방향

인지형 형태로 확장할 필요성이 정리되며, 계산 효율과 표현 일관성을 동시에 고려하는 설계가 요구됨을 강조한다. 요컨대 정확도 향상뿐 아니라 실제 배치 가능성을 위해 연산량과 메모리 부담을 관리하고, 융합 과정에서의 정보 손실을 최소화하려는 관점이 필요하다는 점을 정돈한다[17, 18, 19].

이를 해결하기 위해 이중 스트림(이미지/라이다)과 이중 공간 구조를 기반으로, 하이트 보존 복셀 인코딩, 양방향 투영, BEV 정렬을 결합한 MambaFusion 을 제안한다[7, 8]. 전반적인 구성은 모달별 특징을 독립적으로 보존하면서도 정렬 단계에서 공통 표현으로의 변환을 점진적으로 수행해 융합 안정성을 높이는 방향으로 설명되며, 서로 다른 신호 특성에서 발생할 수 있는 불일치가 후속 단계로 전이되지 않도록 흐름을 정돈한다. 핵심은 혼합형 Mamba 블록으로, Hilbert 곡선 기반 직렬화과 SS2D 선택적 스캐닝을 통해 선형 시간으로 글로벌 정보 교환을 수행하면서도 방향성 사전 정보를 확보한다[14, 15]. 이 과정은 장거리 의존성을 효율적으로 취합하는 동시에, 공간적 구조를 고려한 전파 경로를 유지함으로써 장면 내 문맥 연결을 보다 안정적으로 지원하는 취지를 갖는다. 또한 멀티 브랜치 CBR 스템과 픽셀/채널 어텐션을 결합한 모듈로 장거리 컨텍스트를 강화하고, depthwise conv 의 다중 수용영역과 MFCA 기반 주파수 밴드 집합을 포함한 모듈로 교차모달 스펙트럼 조화를 수행한다[16]. 각 모듈은 멀티스케일 계층을 따라 상호 보완적으로 작동하도록 구성되며, 전반적으로 교차모달 상호작용의 안정화를 목표로 하면서도 표현의 균형을 유지하는 방향을 지향한다.

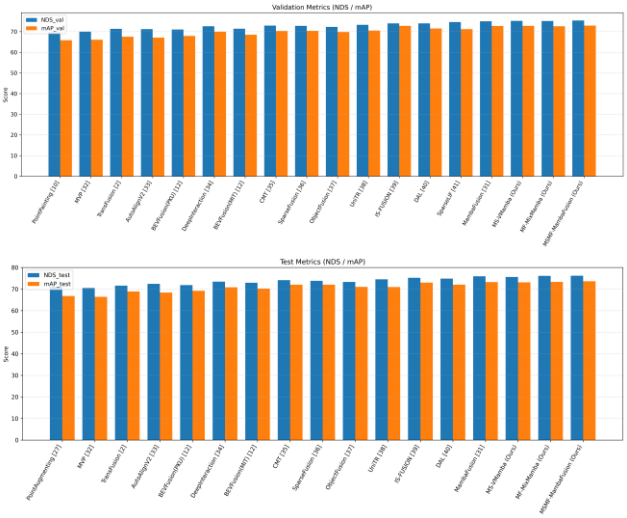

Table 1:nuScenes 검증/테스트 세트에서 주요 카메라-라이다 3D 객체 검출 방법들의 NDS 및 mAP 성능을 비교한 막대그래프.

실험은벤치마크에서 NDS 와 mAP 를 사용해 성능을 검증하는 방식으로 진행되며, SOTA 비교와 어블레이션을 통해 각 모듈의 기여를 체계적으로 분석한다. 특히 동일한 평가 프로토콜을 유지한 상태에서 구성 요소를 단계적으로 추가·제거하면서 성능 변화의 경향을 정리하고, 지연 시간을 고려한 설정에서도 개선 효과가 안정적으로 유지되는지를 함께 점검한다. 또한 정량 지표의 변화뿐 아니라 전반적인 검출 경향을 함께 확인함으로써 특정 상황에만 치우친 개선이 아닌지 해석하고, 모듈 간 결합이 전체 파이프라인에서 어떤 역할을 하는지에 대한 설명을 보완한다. 그 결과 롱-레인지 컨텍스트 모델링 측면에서, 그리고 바운더리 선명화 및 잡음 억제 측면에서 상대적으로 두드러진 이점을 제공함을 보이며, 최종적으로 풀 모델은 테스트 기준 NDS 76.2%와 mAP 73.6%를 달성하여 정확도-효율 절충 관점에서 의미 있는 개선을 확인한다.

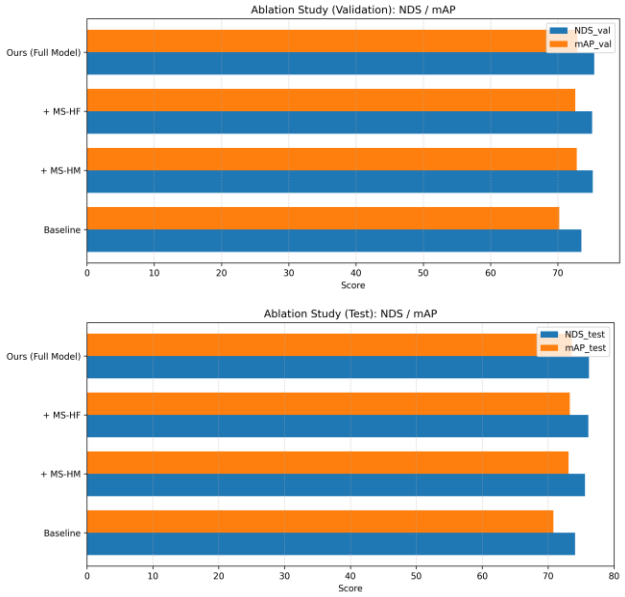

Table 2: nuScenes 검증/테스트 세트에서 구성요소별 NDS 및 mAP 변화를 보여주는 소거실험(ablations) 가로 막대그래프.

Ⅲ. 결론

본논문에서는 카메라-라이다 기반 3D 객체 탐지를 위한 멀티스케일·주파수 인지형 Mamba 프레임워크를 제안하고, 디렉셔널 선택적 스캐닝과 주파수 인지형 특징 집합을 통해 특징 피라미드 계층와 스펙트럼 정렬를 함께 강화하여 교차모달 잡음를 완화함을 보였다. 이를 통해 서로 상이한 특성을 갖는 두 모달리티의 정보를 보다 안정적으로 결합하고, 표현의 일관성을 높이는 방향으로 설계를 정리한다.

또한 제안 모듈들은 멀티스케일 특징을 효율적으로 취합하면서도, 주파수 관점에서의 정렬을 병행함으로써 세부 정보 보존과 전역 컨텍스트 활용을 동시에 지원하는 데 초점을 둔다. 하여, 작은 객체에서 요구되는 세밀한 단서와 원거리 장면에서 필요한 롱-레인지 컨텍스트를 한쪽으로 치우치지 않게 유지하도록 구성되어, 전반적인 융합 과정에서의 불필요한 간섭을 줄이는 효과를 기대할 수 있다.

마지막으로 nuScenes 실험과 어블레이션 스터디를 통해 실용적 지연 시간 조건에서의 성능과 각 컴포넌트의 기여를 확인함으로써, 적정한 연산 예산 내에서도 소형 객체 세부 정보과 롱-레인지 컨텍스트를 동시에 보존할 수 있음을 제시한다. 이는 실제 적용을 고려할 때도 과도한 연산 증가 없이 성능 향상을 도모할 수 있다는 점에서, 향후 관련 멀티모달 3D 인지 시스템 설계에 참고 가능한 시사점을 제공한다.

.

참 고 문 헌

[1] Y. Zhang, et al., "Perception and sensing for autonomous vehicles under adverse weather: A survey," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 196, pp. 146- 177, 2023.

[2] J. Choe, H. Choi, D. Lee, et al., "Performance verification of autonomous driving lidar sensors under fog and rain," Sensors, vol. 24, no. 1, p. 14, 2023.

[3] H. Caesar, V. Bankiti, A. H. Lang, et al., "nuScenes: A multimodal dataset for autonomous driving," in Proc. CVPR, 2020.

[4] X. Bai, Z. Hu, X. Zhu, et al., "TransFusion: Robust lidar-camera fusion for 3D object detection with transformers," in Proc. CVPR, 2022.

[5] S. Vora, A. H. Lang, B. Helou, et al., "PointPainting: Sequential fusion for 3D object detection," in Proc. CVPR, 2020.

[6] C. Wang, C. Ma, M. Zhu, et al., "PointAugmenting: Cross-modal augmentation for 3D object detection," in Proc. CVPR, 2021.

[7] Z. Liu, H. Tang, A. Amini, et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," arXiv preprint arXiv:2205.13542, 2022.

[8] Z. Li, W. Wang, H. Li, et al., "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in Proc. ECCV, 2022.

[9] H. Wang, H. Tang, S. Shi, et al., "UniTR: A unified and efficient multi-modal transformer for bird's-eye-view representation," in Proc. ICCV, pp. 6792- 6802, 2023.

[10] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. ICCV, 2021.

[11] Q. Cai, Y. Pan, T. Yao, et al., "ObjectFusion: Multi-modal 3D object detection with object-centric fusion," in Proc. ICCV, pp. 18067- 18076, 2023.

[12] Z. Zhou and S. Tulsiani, "SparseFusion: Distilling view-conditioned diffusion for 3D reconstruction," in Proc. CVPR, pp. 12588- 12597, 2023.

[13] S. Wang, B. Z. Li, M. Khabsa, et al., "Linformer: Self-attention with linear complexity," arXiv preprint arXiv:2006.04768, 2020.

[14] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint arXiv:2312.00752, 2024.

[15] H. Liu, L. Zhuang, W. Zhou, et al., "Vision Mamba: Efficient visual representation learning with state space model," arXiv preprint arXiv:2401.09417, 2024.

[16] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in Proc. ICCV, 2021.

[17] J. Chen, W. Ma, X. Shi, et al., "IS-Fusion: Instance-augmented lidar-camera fusion for BEV 3D object detection," in Proc. CVPR, 2024.

[18] W. Hu, P. Li, Y. Han, et al., "Rethinking lidar-camera fusion in 3D object detection: Detecting as labeling (DAL)," in Proc. ECCV, 2024.

[19] H. Zhang, L. Liang, P. Zeng, et al., "SparseLiF: High-performance sparse lidar-camera fusion for 3D object detection," in Proc. ECCV, pp. 109- 128, 2024.