불확실성기반 의미적 정보 강화를 통한 3차원 점유 그리드 예측 네트워크 성능 개선

강창원, 김예철, 최준원*

한양대학교

changwonkang@spa.hanyang.ac.kr, yckim@spa.hanyang.ac.kr, *junwchoi@hanyang.ac.kr

Uncertainty-aware semantic refinement for 3D occupancy grid prediction network

Changwon Kang, Yecheol Kim, Jun Won Choi*

Hanyang University

요 약

본 논문은 자율주행 환경에 사용되고 있는 모노 카메라 센서 사용 모델인 OccFormer 모델에서 점유 그리드 예측 성능을 개선하는 기술을 제안하였다. 우선 이미지 백본 네트워크를 통해 이미지 특징맵을 생성하고, 깊이 정보와 문맥정보를 추출한다. 이를 통해 생성된 3차원 점유 특징맵을 Swin-transfomer기반의 방법과 convolution layer를 통해 다중스케일의 3차원 점유 특징맵을 생성한다. 생성된 다중스케일의 3차원 점유 특징맵을 기반으로 불확실성 점수 헤드를 통해 불확실성 점수를 예측한다. 이후 불확실성 점수가 높은 영역을 선정하여 Deformable attention기법을 통해 이미지 특징맵의 정보를 기존에 생성한 3차원 점유 특징맵에 보충한 후 트랜스포머 기반의 디코더를 통해 점유 그리드 예측을 수행한다. 본 논문은 nuScenes 데이터셋을 이용하여 실험을 진행하였으며, 제안한 방식을 통해 기존 방법에 비해 높은 성능을 달성 하였다.

Ⅰ. 서 론

최근 딥러닝의 발전으로 자율주행 분야에 대한 연구가 활발히 진행되고 있다. 자율주행에 있어서 주변 환경을 인식하는 것은 중요하기 때문에, 3차원 동적 환경 인지를 위한 3차원 물체 검출 기법에 대한 많은 연구가 되었다[1, 2]. 하지만 동적 객체만을 검출하는 표현력의 한계로 인해, 주변 정적, 동적 환경의 종합적인 정보를 제공하지 못한다는 한계가 있다. 최근 테슬라에서 멀티뷰 모노 카메라로부터 3차원 점유 그리드를 추정하는 기법을 선보인 바가 있다. 이를 기반으로 3차원 점유 그리드 예측 기법에 관한 연구가 활발히 진행되기 시작했다[3,4,5]. 3차원 점유 그리드 예측은 3차원 공간을 균일한 그리드로 분할 후 그리드 별로 점유 예측 및 클래스를 추정하는 작업이다. 지금까지 제안된 3D 점유 예측 기법의 성능이 높지 않다는 한계가 있다. 따라서 안전한 자율 주행 차량 운행을 위해 모노 카메라 기반의 점유 그리드 예측에 관한 연구가 필요하다. 본 논문은 3차원에서 불확실 영역을 샘플링 하는 방법을 제안 한다. 이를 3차원 점유 그리드 예측 네트워크에 활용하여 점유 예측의 성능을 높이기 위해 인코더에서 의미적 정보 강화를 수행하는 효과적인 모듈을 제안함으로써, 3차원 점유 그리드 예측 모델의 성능을 높였다.

Ⅱ. 본론

2.1) 제안하는 3차원 점유 그리드 예측 네트워크

제안하는 3차원 점유 그리드 예측 네트워크는 주변 환경을 다루는 6대 카메라의 이미지를 입력으로 사용한다. 전체 구조는 그림 1에 첨부 하였다.

우선 모노카메라로 수집된 이미지를 입력으로 하여 이미지 백본 네트워크를 통해 이미지 특징맵을 추출한다. 추출된 이미지 특징맵을 활용하여 BEVDepth[9]와 동일하게 깊이정보와 문맥정보를 추출하고, Voxel Pooling을 통해 3D Feature Volume을 생성한다. 이렇게 생성된 3차원 점유 특징맵은 Swin Transformer[7]의 기법을 활용한 Dual-path transformer block과 convolution layer를 통과하여 다중 스케일 3차원 특징맵을 생성한다. 이를 불확실성 기반의 의미적 강화 모듈을 통과시켜 의미적 정보가 강화된 다중 스케일의 3차원 특징맵을 얻고, Deformable attention[8] 기법을 이용하여 다중 스케일의 특징맵을 정렬한다. 최종적으로 트랜스포머 기반의 디코더를 통해 점유 그리드 예측을 수행한다.

2.2) 불확실성 기반의 의미적 강화 모듈

제안하는 불확실성 기반의 의미적 강화모듈의 구조는 그림 2와 같다. 먼저 convolution layer로 이루어진 불확실성 점수 헤드를 통해 다중 스케일 3차원 특징맵의 불확실성 점수를 예측한다. 높은 점수로 예측된 영역을 선택하여 마스킹 한다. 이후 마스킹된 영역과 이미지 정보를 Deformable attention 기법을 통해 의미적 정보를 강화한다.

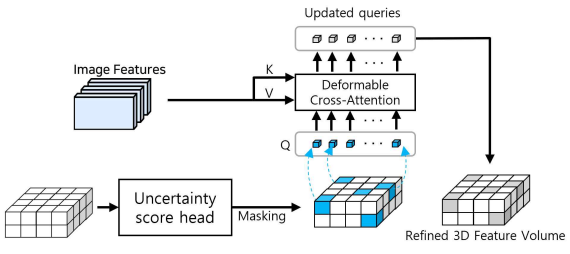

그림 2. 불확실성 기반의 의미적 정보 강화 모듈

2.3) 실험 결과

제안하는 네트워크를 학습 및 성능 평가를 위해 nuScenes 자율주행 공개 데이터셋의 1/7만 이용하였다. 제안하는 기법의 성능 평가 결과는 표1과 같다. 표1을 통해 nuScenes 3차원 라이다 세분화 데이터셋에서 베이스라인인 OccFormer[5]에 비해 높은 성능을 나타냈다. 이는 제안하는 불확실성 기반의 의미적 정보 강화 모듈이 효과적으로 의미적 정보를 강화하고 있음을 보인다.

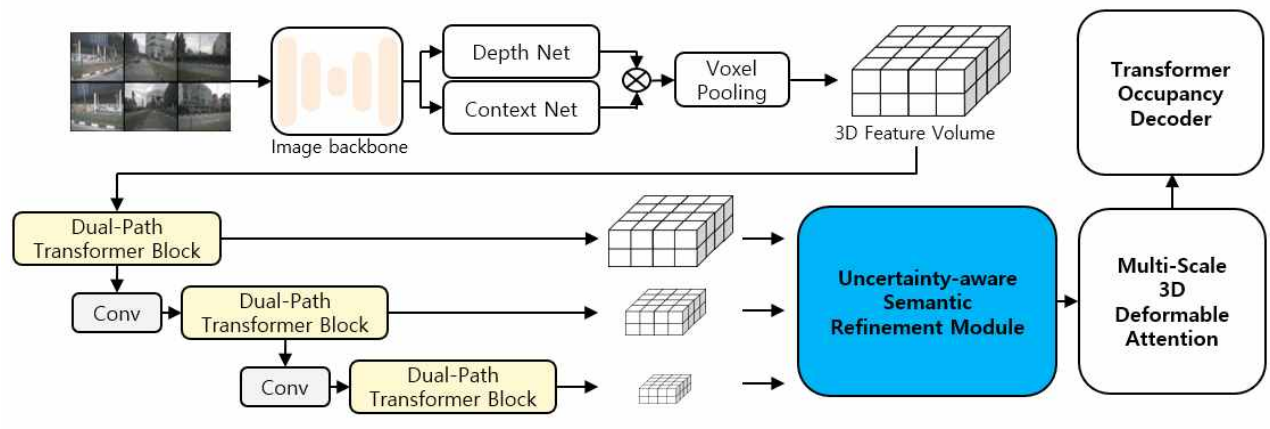

그림 1. 제안하는 점유 예측 네트워크 구성

Class \ method	OccFormer[5]	our method
barrier	**0.683**	0.671
bicycle	**0.358**	0.341
bus	0.837	**0.853**
car	0.796	**0.798**
construction vehicle	0.330	**0.359**
motorcycle	0.413	**0.531**
pedestrian	0.541	**0.547**
traffic cone	0.366	**0.377**
trailer	0.534	**0.536**
truck	**0.773**	0.754
driveable surface	0.915	**0.917**
other flat	**0.664**	0.661
sidewalk	0.636	**0.642**
terrain	0.670	**0.671**
manmade	0.783	**0.785**
vegetation	0.758	**0.760**
mIoU	0.629	**0.638**

표 1. 제안하는 방법의 성능

Ⅲ. 결론

본 논문은 효과적인 모노 카메라기반의 3차원 점유 그리드 예측을 위해 불확실성 기반의 의미적 강화 모듈 기술을 소개하였다. 이는 불확실성 점수를 기반으로 불확실한 영역을 선정한다. 선정된 영역은 이미지 특징맵과 deformable attention기법을 통해 다중스케일에서의 의미적 정보가 강화된다. 해당 알고리즘은 nuScenes 공개 데이터셋에서 학습과 성능 평가가 이루어졌으며, 베이스라인 대비 높은 성능을 보였다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-01373,인공지능대학원 지원(한양대학교))

참 고 문 헌

[1] Lang, Alex H., et al. "Pointpillars: Fast encoders for object detection from point clouds." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[2] Yin, Tianwei, Xingyi Zhou, and Philipp Krahenbuhl. "Center-based 3d object detection and tracking." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[3] Li, Yiming, et al. "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[4] Huang, Yuanhui, et al. "Tri-perspective view for vision-based 3d semantic occupancy prediction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[5] Zhang, Yunpeng, Zheng Zhu, and Dalong Du. "OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[6] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[7] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[8] Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." arXiv preprint arXiv:2010.04159 (2020).

[9] Li, Yinhao, et al. "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 2. 2023.