Transactions of KSAE, Vol. 30, No. 10, pp.809-817 (October, 2022)

pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2022.30.10.809

자율주행을 위한 카메라 · 라이다 융합 기반 복셀-픽셀 매칭을 통한 포인트 클라우드 의미론적 분할 네트워크

송 하 민[1)] · 조 지 은[1)] · 하 진 수[1)] · 박 재 현[2)] · 조 기 춘[*1)]

건국대학교 스마트운행체공학과[1)] · 한양대학교 자동차공학과[2)]

Camera-LiDAR Fusion-Based Point Cloud Semantic Segmentation through Voxel-Pixel Matching for Autonomous Driving

Hamin Song[1)] · Jieun Cho[1)] · Jinsu Ha[1)] · Jaehyun Park[2)] · Kichun Jo[*1)]

[1)]Department of Smart Vehicle Engineering, Konkuk University, Seoul 05029, Korea
[2)]Department of Automotive Engineering, Hanyang University, Seoul 04763, Korea
(Received 4 July 2022 / Revised 4 August 2022 / Accepted 4 August 2022)

Abstract : To ensure safe autonomous driving, understanding the environment around the vehicle is essential. Point cloud semantic segmentation is an efficient task for understanding surrounding scenes. However, performing semantic segmentation with a single LiDAR sensor has limitations in terms of sparsity and absence of color/texture information. In order to address these limitations, the camera-LiDAR fusion-based network is actively evaluated. Existing sensor fusion-based networks project multi-sensor data onto each 2D plane. Regarding the epipolar geometry problem, the existing methods cannot have the 1:1 matching of pixel-point during feature map fusion. Therefore, in this paper, we apply a backbone network that utilizes 3D data as itself. We propose a fusion module based on voxel-pixel matching for accurate feature map fusion. For verification, we used the SemanticKITTI dataset, and the performance improved by 2.7 % compared to when a single LiDAR is used.

Key words : Autonomous driving(자율주행), Deep learning(딥러닝), Perception(인식), Sensor fusion(센서 퓨전), Semantic segmentation(의미론적 분할)

1. 서 론

안전한 자율주행을 위해서는 신뢰할 수 있는 인식 기술이 필요하다. 차량 주변의 객체 종류 및 위치를 파악하여야 상황에 적절한 주행 전략을 사용할 수 있기 때문이다.[1)] 차량의 주행환경을 인식하기 위해서는 카메라, 레이더, 라이다(Light Detection and Ranging, LiDAR)와 같은 센서들이 널리 사용되고 있다. 그 중 라이다 센서는 레이저를 사용하여 주변을 스캔하고 물체에 반사되어 돌아오는 시간을 측정함으로써 주변 환경을 높은 거리 정확도의 3차원 정보인 포인트 클라우드로 표현할 수 있다.

포인트 클라우드 데이터는 기하학적 정보(x, y, z)와 반사율(Intensity) 정보를 포함하지만, 의미 정보의 부재로 인해 차량 주변 환경 이해가 어렵다. 따라서 포인트 클라우드에 의미 정보를 부여하는 과정이 필요하며, 이를 위해 포인트 클라우드 의미론적 분할에 대한 연구가 활발히 이루어지고 있다. 포인트 클라우드 의미론적 분할이란 포인트 클라우드의 각 포인트가 속하는 클래스를 예측하는 작업이다. 이는 딥러닝의 발전에 따라 딥러닝 네트워크를 기반으로 이루어지며, 포인트 클라우드를 가장 세밀한 단위로 분류할 수 있으므로 장면 이해에 적합한 작업이다.

하지만 라이다 센서만을 사용하여 포인트 클라우드 의미론적 분할을 수행할 경우 몇 가지 한계점이 있다. 첫 번째로, 라이다 센서 구조의 특성상 객체가 센서로부터

*Corresponding author, E-mail: kichun@konkuk.ac.kr

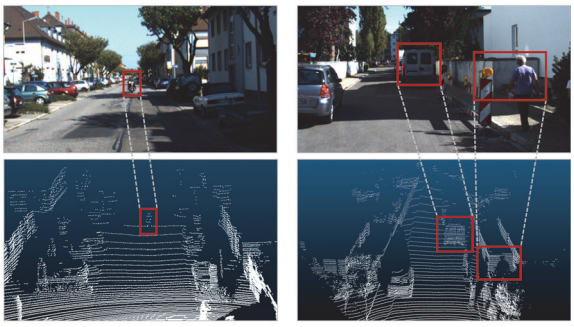

Fig. 1 Sparsity of point cloud

원거리에 위치하거나 객체의 크기가 작을수록 해당 객체에 반사되어 돌아오는 포인트의 개수가 감소한다. Fig. 1의 (a)은 원거리에 위치한 오토바이에 매우 적은 포인트가 맺힌 예시를 나타낸다. 이러한 경우 주변 포인트와의 관계에 기반하여 특징을 추출하는 딥러닝 네트워크에서 지역적인 특징 추출이 어렵기 때문에 오분류에 영향을 미칠 수 있다. 두 번째로, 분류에 큰 도움을 주는 색상 및 질감 정보를 제공받지 못한다. Fig. 1의 (b)와 같이 비슷한 형상을 갖는 객체를 위치 및 반사율 정보에만 기반하여 처리할 경우 오분류가 빈번히 발생할 수 있다.

앞서 언급한 두 한계점을 보완하기 위해서는 센서 융합이 필수적이다. 자율주행 차량의 다양한 인식 센서 중, 카메라는 포인트 클라우드 데이터에 비해 밀도 높은 고해상도의 픽셀 데이터를 제공한다. 또한 색상 및 표면 질감을 포함한 풍부한 정보를 제공할 수 있다. 따라서 카메라 센서와 라이다 센서를 융합하여 사용할 경우, 단일 라이다 사용 시 발생하는 두 한계점을 극복하고 각 센서의 장점을 극대화함으로써 더 신뢰할 수 있는 높은 정확도의 의미론적 분할 결과를 얻을 수 있을 것으로 기대된다.

최근 들어 카메라와 라이다 센서를 융합하여 포인트 클라우드 의미론적 분할을 수행하는 딥러닝 네트워크가 활발히 연구 중이다. 두 센서로부터 얻어지는 데이터를 융합하는 방식은 매우 중요하다. 가장 빈번하게 사용되는 방법은 Fig. 2와 (b)와 같이 포인트 클라우드를 구형 투영(Spherical projection)을 통해 거리 이미지(Range image)로 표현하고, 이미지 또한 구형 투영을 통해 거리 이미지로 표현하는 것이다. 이 방법은 두 센서로 얻은 데이터를 동일한 2차원 좌표계로 통일함으로써 딥러닝 네트워크에서 추출된 특징 지도를 융합하기 매우 용이하다는 장점이 있다. 하지만 이러한 전처리는 Fig. 2에서 확인할 수 있듯이 이미지의 해상도를 포인트 클라우드의 해상도로 낮추기 때문에, 카메라의 고해상도 특성을 활용하지 못하여 단일 라이다 사용 시 발생하는 데이터의 밀도가 낮은 문제를 해결할 수 없다. 따라서 이미지의 고해상도 장점을 활용하기 위해 전처리 없는 고해상도의 원본 이미지를 사용해야 한다.

현재 공개된 네트워크 중 원본 이미지를 사용하여 포인트 클라우드와 융합하는 네트워크는 대개 입력으로 원본 이미지와 구형 투영된 포인트 클라우드 거리 이미지를 활용하고 있다. 두 입력 데이터가 각 센서의 장점을 잃지 않는 형태이지만, 두 데이터 모두 2차원으로 표현된다는 점에서 중간 특징 지도 융합 시 픽셀-포인트의 기하학적으로 정확한 대응이 어렵다. 그 이유는 등극선 기하(Epipolar geometry)와 관련 지어 설명할 수 있다. Fig. 3은 여러 위치의 3차원 객체가 각 센서의 2차원 평면에 투영된 위치를 나타낸다. 본 예시에서 다양한 위치의 객체(A, B, C, D)가 카메라 2차원 평면에는 각기 다른 점으로 투영되지만, 라이다 2차원 평면에는 하나의 점으로 투영된다. 이때 카메라 2차원 평면의 여러 투영 점은 한 직선상에 위치하며 이를 등극선 직선(Epipolar line)이라 칭한

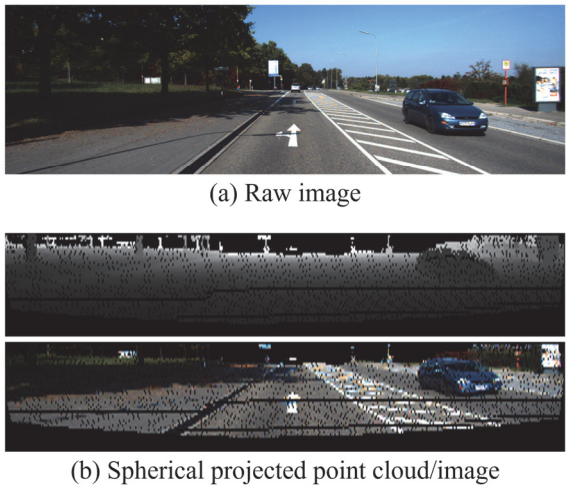

Fig. 2 Image resolution degradation due to spherical projection

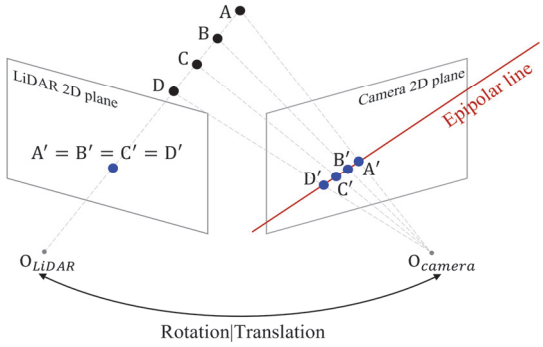

Fig. 3 Epipolar geometry

다. 이러한 경우에 객체 투영 결과는 각 센서 평면에서 점대점(1:1) 매칭이 아닌, 점대선(1:N) 매칭이라 볼 수 있다. 따라서 2차원으로 표현한 이종 센서 데이터를 기하학적으로 정확히 매칭시키기에 어려움이 있다. 이러한 어려움은 컨볼루션 네트워크 및 차원 축소 레이어를 거쳐 크기가 축소된 특징 지도를 매칭시킬 때 또한 발생하게 된다.

본 논문은 각 센서 데이터에서 추출된 특징 지도를 기하학적으로 정확하게 대응시킬 수 있는 센서퓨전 기반의 포인트 클라우드 의미론적 분할 네트워크 설계를 목적으로 한다. 정확한 특징 지도 매칭을 통해 각 센서로부터 추출된 특징을 정확한 위치에서 융합함으로써 단일 라이다 사용 모델보다 높은 성능을 달성하고자 한다.

이러한 목적을 달성하기 위해 본 논문에서는 복셀-픽셀 대응 관계를 이용한 특징 지도 융합 모듈을 새롭게 제안한다. 기존의 다른 모델들과 달리 포인트 클라우드에 복셀 기반의 백본 네트워크를 적용함으로써 3차원 정보를 유지하고, 캘리브레이션 파라미터를 사용하여 네트워크의 어느 위치에서나 정확한 특징 지도 매칭이 가능하도록 한다. 또한 특징 지도 매칭 시 이미지의 풍부한 정보를 추가적으로 활용하기 위해 매칭되는 픽셀 주변의 정보까지 고려한다. 제안하는 네트워크는 SemanticKITTI[2] 데이터셋으로 검증되었으며, 최신 성능의 센서퓨전 기반 포인트 클라우드 의미론적 분할 모델(PMF[3])보다 1.1 % 높은 성능(mIoU)을 달성한다.

2. 관련 연구

제안하는 네트워크와 관련한 이전 연구로 라이다 센서만을 사용한 포인트 클라우드 의미론적 분할 네트워크, 카메라-라이다 융합을 통한 포인트 클라우드 의미론적 분할 네트워크에 대하여 설명하고자 한다.

2.1 라이다 기반 포인트 클라우드 의미론적 분할

라이다 센서만을 사용한 포인트 클라우드 의미론적 분할 네트워크는 포인트 클라우드의 전처리 과정에 따라 크게 2차원 기반 방법, 3차원 기반 방법으로 분류할 수 있다.

2차원 기반 방법은 포인트 클라우드를 평면에 투영함으로써 2차원 이미지로 변환하여 사용한다. 따라서 이미지에 사용되는 CNN[4](Convolutional Neural Network) 모델을 적용하기 용이하고 데이터의 차원 축소로 인해 적은 메모리 부담이 있다는 장점이 있다. RangeNet++[5,6], SqueezeSeg[7], SalsaNext[8] 등의 모델은 구형 투영을 통해 거리 이미지를 생성한다. 이러한 구형 투영은 2차원 기반 방법에 속하는 대부분의 모델이 사용하는 방법이다. 다른 투영 방법으로는 PolarNet[9]이 사용한 조감도 생성 방법이 있다. 포인트 클라우드를 위에서 내려다 본 형태의 이미지로 변환하여 모델의 입력으로 사용한다. 이러한 다양한 투영 방법으로 생성된 2차원 이미지는 여러 층으로 구성된 컨볼루션 네트워크로 처리되어 의미론적 분할된 이미지를 예측한다. 예측된 2차원 이미지 결과를 3차원 포인트 클라우드에 입히는 방식으로 최종적인 의미론적 분할된 포인트 클라우드를 생성해낸다.

3차원 기반 방법은 포인트 클라우드의 차원 축소 없이 3차원 데이터 특성을 그대로 활용한다. 원본 포인트 클라우드를 전처리 없이 사용하는 방법은 PointNet[10]에서 처음 고안되었다. PointNet은 입력 포인트 클라우드에 다층 퍼셉트론(Multi-layer perceptron, MLP[11])을 적용하여 특징을 추출함으로써 불규칙적인 입력 형태에 영향을 받지 않게 하였다. 이후에 포인트 클라우드에 바로 적용 가능한 컨볼루션 연산인 포인트 컨볼루션이 고안됨으로써 KPConv[12], ConvPoint[13] 등의 모델이 등장하였다. 또한 3차원 데이터를 일정한 크기의 격자로 나누어 규격화하는 복셀화 기반 모델도 다양하게 연구되었다. PointGroup[14], 3D-MPA[15], PCSCNet[16] 등은 입력 포인트 클라우드를 복셀화한 후, 3차원 컨볼루션 네트워크를 적용한다. 3차원 기반 방법의 모델은 포인트 클라우드의 기하학적 특성을 보유한 채로 특징 추출을 수행하여 의미론적 분할된 포인트 클라우드를 생성한다.

추가로 최근 들어 2차원 기반 방법과 3차원 기반 방법을 모두 적용하는 모델도 등장하였다. RPVNet[17]은 다양한 전처리를 거친 포인트 클라우드를 모두 사용하여 높은 성능을 달성하였다.

2.2 센서 융합 기반 포인트 클라우드 의미론적 분할

카메라, 라이다 센서 데이터 융합 모델은 특징 지도 및 데이터의 융합 시점에 따라 크게 초기 융합(Early fusion), 후기 융합(Late fusion), 중기 융합(Middle fusion)으로 분류할 수 있다. 초기 융합은 네트워크 입력 전에 두 센서 데이터를 융합하는 방법으로, 융합된 데이터를 단일 네트워크로 처리한다. 따라서 각 센서 데이터의 특성을 고려하지 못한다는 단점이 있다. 후기 융합은 각 센서 데이터에 별도의 네트워크를 적용하여 예측 결과를 생성한다. 각 센서의 예측 결과를 융합하기 때문에 연산량 부담이 크다는 단점이 있다. 중기 융합은 각 센서 데이터에 별도의 특징 추출기를 적용하고, 원하는 시점에 특징 지도를 융합한다. 이는 각 센서 데이터의 특성을 고려할 수 있는 동시에 연산량 부담이 크지 않기 때문에 중기 융합

을 통한 포인트 클라우드 의미론적 분할 모델이 활발히 연구 중이다.

중기 융합을 통한 포인트 클라우드 의미론적 분할 모델은 앞서 언급한 2차원 기반 방법의 네트워크를 적용하는 것이 일반적이다.[18] VIASeg[19], RGBAL[20]는 포인트 클라우드와 이미지를 모두 구형 투영을 통해 거리 이미지로 변환한다. 이는 앞서 언급했듯이 카메라 센서의 고해상도 특징을 활용하지 못하는 단점이 있다. FuseSeg[21]는 원본 이미지와 구형 투영된 포인트 클라우드를 활용한다. 별도의 특징 추출기를 통해 특징 지도를 만들고 중간 특징 지도를 융합하는 구조를 가지고 있지만, 앞서 언급한 등극선 기하 문제로 인해 기하학적으로 정확한 특징 지도 융합에 어려움이 있다.

이 논문에서는 중기 융합 기반의 네트워크를 제안한다. 이때 등극선 기하와 관련한 특징 지도 대응이 어려운 문제를 해결하고자 3차원 복셀 기반의 모델인 PCSCNet을 기반으로 모델을 구성하였다.

3. 네트워크 구조

포인트 클라우드 의미론적 분할의 성능 향상을 위해 카메라와 라이다 특징 지도를 융합하는 것은 매우 중요하다. 기존의 많은 방법은 카메라의 고해상도 특성을 활용하지 못하거나, 특징 지도의 정확한 대응이 어려웠다. 따라서 이 논문에서는 3차원 기반 모델을 백본 네트워크로 사용하며, 특징 지도를 기하학적으로 정확히 대응시킬 수 있는 모듈을 소개한다.

제안하는 네트워크의 전체 구조는 Fig. 4에 나타나 있다. 네트워크는 포인트 클라우드와 이미지를 입력으로 받으며, 각 데이터에 별도의 인코더를 적용하여 특징을 추출한다. 인코더는 모두 4개의 블록으로 구성되어 있으며, 각 블록에서 추출된 특징 지도에 복셀-픽셀 매칭 기반의 특징 지도 융합 모듈을 적용한다. 융합된 특징 지도는 포인트 클라우드 인코더 블록으로 전달된다. 포인트 클라우드 인코더에서 추출 및 융합된 특징 지도는 단일 디코더로 전달되고 최종적으로 의미론적 분할된 포인트 클라우드를 생성한다. 제안하는 네트워크를 크게 백본 네트워크, 복셀-픽셀 매칭 기반 특징 지도 융합 모듈의 두 가지로 나누어 설명하고자 한다.

3.1 백본 네트워크

중기 융합을 통한 카메라-라이다 융합 네트워크를 설계하기 위하여 센서별 별도의 특징 추출기를 적용하고자 하였다. 따라서 센서별 별도의 백본 네트워크를 선정하여 적용하였다.

포인트 클라우드를 위한 백본 네트워크로는 복셀화 및 포인트 컨볼루션 기반의 포인트 클라우드 의미론적 분할 네트워크인 PCSCNet을 선정하였다. PCSCNet은 입력 포인트 클라우드를 사전에 설정한 격자 크기를 따라 복셀화하고, 비어있지 않은 복셀에 대한 복셀별 특징을 추출하기 위해 다층 퍼셉트론과 KPConv를 적용한다. 복셀별 특징이 추출되었다면, 차원을 축소해가며 특징을 추출하는 인코더와 원본 크기로 복원하며 포인트별 클래스를 예측하는 디코더 구조를 거친다. 이때 효율적인 연산을 수행하기 위하여 비어있지 않은 복셀만을 이용

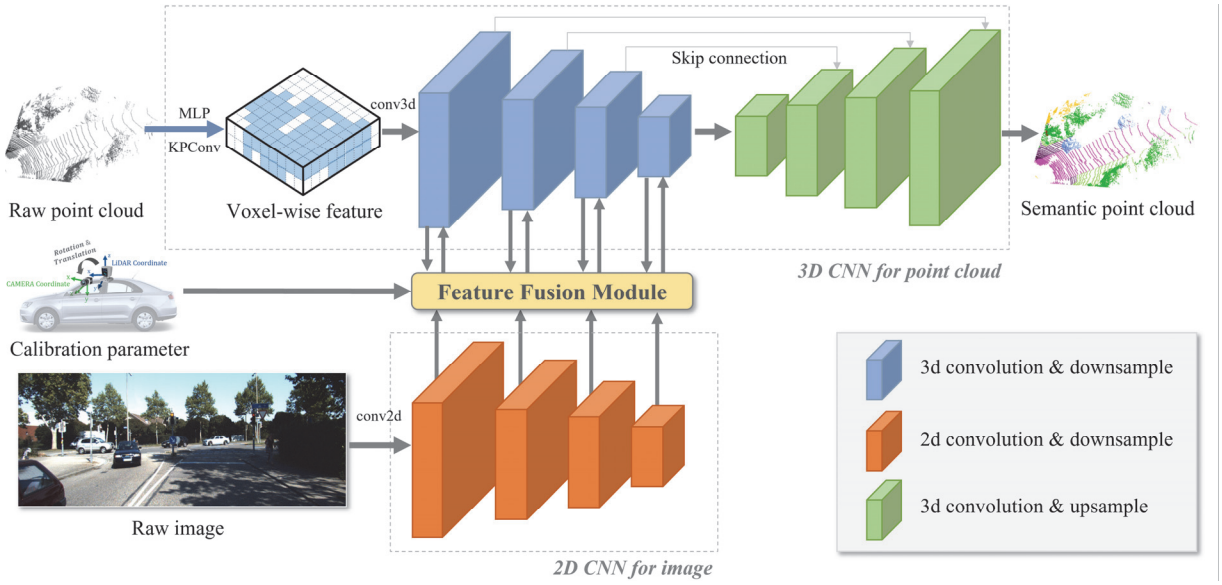

Fig. 4 Overall network architecture

하여 컨볼루션 연산을 수행하는 Sparse convolution[22]을 적용한다. 이는 비어있는 복셀에 대한 불필요한 연산을 줄임으로써 메모리 부담을 감소시킨다. PCSCNet은 최종적으로 포인트 별 클래스를 예측하여 의미론적 분할된 포인트 클라우드를 생성한다.

이미지의 특징 추출을 위한 백본 네트워크로는 ResNet[23]을 사용하였으며 모델 초기값으로 ImageNet[24] 데이터셋으로 사전학습된 가중치를 사용하였다. ResNet은 컨볼루션 층의 개수에 따라 다양한 모델을 제공하는데, 그 중 ResNet34를 적용하였다.

3.2 복셀-픽셀 매칭 기반 특징 지도 융합 모듈

PCSCNet의 인코더에서 추출된 특징 지도는 복셀별 특징으로 구성되어 있고, ResNet에서 추출된 특징 지도는 픽셀별 특징으로 구성되어 있다. 따라서 특징 지도 융합을 위해서 복셀과 픽셀을 정확하게 대응시킬 수 있어야 한다.

본 논문에서 제안하는 복셀-픽셀 매칭 기반 특징 지도 융합 모듈의 자세한 구조는 Fig. 5에서 확인할 수 있으며, 총 6가지 단계로 처리된다. 첫 번째 단계로, 복셀 내부에 존재하는 여러 포인트 중 임의의 하나의 포인트 선정한다. 이는 전처리 단계에서 해시 테이블을 생성하여 복셀 인덱스를 키(Key)로 삼아 포인트 좌표(x, y, z)를 얻을 수 있게 하였다. 두 번째로, 카메라의 초점 거리 및 주점, 라이다 센서와 카메라 센서 사이의 회전 및 이동 관계인 캘리브레이션 파라미터를 사용하여 포인트가 대응하는 이미지 특징 지도상의 좌표(u, v)를 도출한다. P_{rect}는 3차원 포인트를 왜곡이 없는 카메라 좌표계로 투영시키기 위한 투영 매트릭스이다. 이는 아래 식 (1)과 같이 구성되어 있으며 f는 초점 거리, c는 주점, b_x는 x축 방향의 Baseline을 의미한다.

$$P_{rect} = \begin{pmatrix} f_u & 0 & c_u & -f_u b_x \\ 0 & f_v & c_v & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{1}$$

추가로 라이다 좌표에서 카메라 좌표로 변환시키기 위한 회전 매트릭스와 이동 벡터가 사용된다. 이를 통해 복셀 내부 포인트가 대응하는 이미지 특징 지도상의 좌표를 구할 수 있고, 세 번째로 해당 픽셀을 포함한 주변 특징을 샘플링한다. 이는 이미지의 풍부한 정보를 추가적으로 활용하기 위함이며 대응되는 픽셀 주변 3×3 영역을 고려한다. 이렇게 얻어진 이미지의 특징은 $3\times3\times C_{pixel}$의 크기를 갖는다. 네 번째로, $3\times3\times C_{pixel}$ 크기의 특징을 $1\times1\times C_{pixel}$의 픽셀 특징으로 만들기 위하여 채널 별 Max pooling을 수행한다. 특징 지도에서 큰 값을 가질수록 더 강한 특징으로 더 큰 활성화 값을 가지게 되기 때문에 최대값을 유지하는 Max pool을 수행하였다. 다섯 번째 단계는 복셀 특징과 픽셀 특징을 융합하는 과정이다. 두 특징을 $1\times1\times(C_{pixel}+C_{voxel})$의 차원을 갖도록 결합한다. 그 후 채널 축소를 위한 1×1 컨볼루션 연산 및 배치 정규화, 활성화 함수를 거쳐 $1\times1\times C_{voxel}$크기의 새로운 복셀 특징을 생성한다. 마지막으로 생성된 복셀 특징을 업데이트한다.

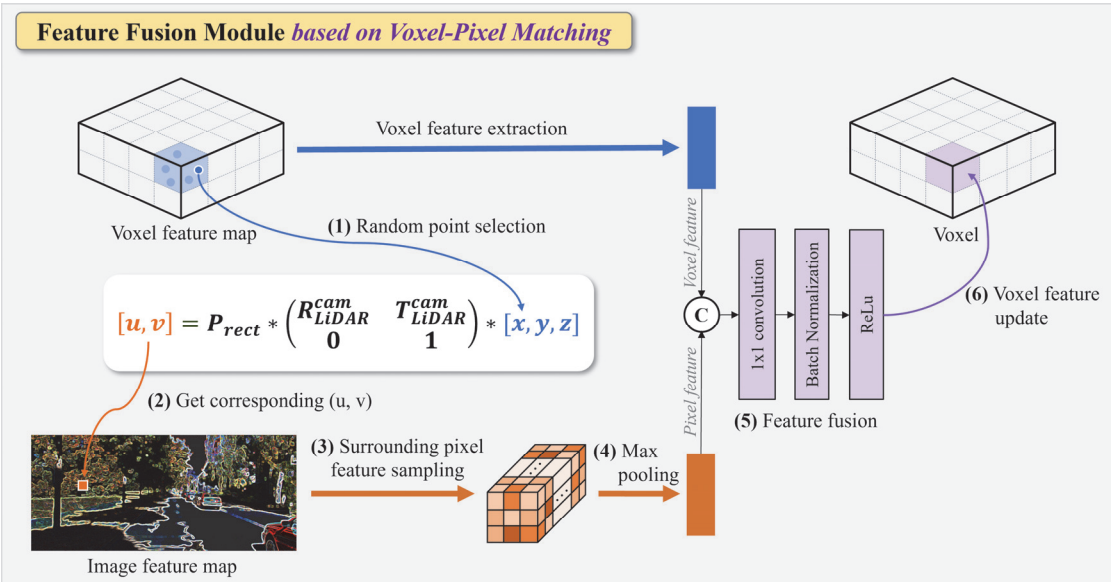

Fig. 5 Feature fusion module based on Voxel-Pixel matching

위의 설명은 단일 복셀 당 수행되는 내용이며, 비어있지 않은 모든 복셀에 대해 위의 과정들이 병렬적으로 수행된다. 또한 본 모듈은 엔코더 블록마다 적용되어 총 4번 수행되어, 입력되는 특징 지도의 크기에 관련 없이 사용 가능하다. 이를 통해 정확한 복셀-픽셀 매칭이 가능하여 포인트 클라우드의 특징 지도를 이미지 데이터를 사용하여 강화시키는 효과가 있다.

4. 실험 설계 및 결과

제안한 네트워크의 효과를 검증하기 위해 연구가 활발히 진행되고 있는 공개 데이터셋을 선정하고, 제안하는 모델의 학습 및 평가를 진행하였다. 사용한 데이터셋과 평가 지표에 대해 설명하고, 모델 구현에 사용한 세부적인 설정과 도출한 실험 결과에 대하여 분석하고자 한다.

4.1 데이터셋 및 평가 지표

SemanticKITTI는 포인트 클라우드의 의미론적 분할을 위해 활발히 사용되고 있는 데이터셋으로 총 19개로 분류되는 포인트 단위의 정답 레이블(Ground truth)을 제공한다. 이는 KITTI Odometry Benchmark[25]에 기반하고 있으며 다양한 주행 경로를 따라 취득한 43,000개의 라이다 스캔 데이터를 제공한다. 또한, 시간 동기화된 43,000개의 원본 이미지 데이터를 제공한다. 카메라는 차량 전방을 바라보는 1개만 설치되어 있기 때문에 모든 실험은 카메라 화각 내부의 포인트 클라우드에 대해서만 진행되었다. 또한 SemanticKITTI는 Test set 중 카메라 화각 내부 포인트에 대한 정답 레이블을 제공하지 않기 때문에 정량적인 성능 비교를 위해 Validation set을 사용하였다.

의미론적 분할된 포인트 클라우드의 성능을 평가하기 위해서는 Mean intersection-over-union(mIoU)가 가장 대표적으로 사용된다. mIoU는 클래스 별 IoU의 평균값이며 IoU는 모델의 예측 결과와 정답 레이블 간의 겹치는 비율을 나타낸다. n번째 클래스의 IoU는 아래의 식 (2)와 같이 구할 수 있다. N은 전체 클래스의 개수를 나타내며 모든 클래스 IoU의 평균값인 mIoU는 아래의 식 (3)을 통해 구할 수 있다.

$$IoU_n = \frac{Area\,of\,Intersection}{Area\,of\,Union} = \frac{TP_n}{TP_n + FP_n + FN_n} \tag{2}$$

$$mIoU = \frac{1}{N}\sum_{n=1}^{N} IoU_n \tag{3}$$

4.2 구현 세부 정보

제안하는 네트워크는 PyTorch로 구현하였으며, 모든 실험은 NVIDIA RTX 3090 GPU에서 수행되었다. 복셀 격자의 크기는 x, y 방향으로 0.1 m, z 방향으로 0.05 m로 설정하였다. 모델 학습 시 최적화 함수로는 Adam optimizer를 사용하였고 손실 함수로는 Cross entropy loss와 Lovasz-Sofrmax loss를 결합하여 사용하였다. 학습은 최

Table 1 Comparison on the SemanticKITTI validation set

Method	Input	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic sign	mIoU (%)
#Points	-	6384	44	52	101	471	127	129	5	21434	974	8149	67	6304	1691	20391	882	8125	317	64	-
RandLANet	LiDAR	92.0	8.0	12.8	74.8	46.7	52.3	46.0	0.0	93.4	32.7	73.4	0.1	84.0	43.5	83.7	57.3	73.1	48.0	27.3	50.0
RangeNet++	LiDAR	89.4	26.5	48.4	33.9	26.7	54.8	69.4	0.0	92.9	37.0	69.9	0.0	83.4	51.0	83.3	54.0	68.1	49.8	34.0	51.2
SequeezeSegV3	LiDAR	87.1	34.3	48.6	47.5	47.1	58.1	53.8	0.0	95.3	43.1	78.2	0.3	78.9	53.2	82.3	55.5	70.4	46.3	33.2	53.3
SalsaNext	LiDAR	90.5	44.6	49.6	86.3	54.6	74.0	81.4	0.0	93.4	40.6	69.1	0.0	84.6	53.0	83.6	64.3	64.2	54.4	39.8	59.4
SPVNAS	LiDAR	96.5	44.8	63.1	59.9	64.3	72.0	86.0	0.0	93.9	42.4	75.9	0.0	88.8	59.1	88.0	67.5	73.0	63.5	44.3	62.3
Cylinder3D	LiDAR	96.4	61.5	78.2	66.3	69.8	80.8	93.3	0.0	94.9	41.5	78.0	1.4	87.5	50.0	86.7	72.2	68.8	63.0	42.1	64.9
***PointPainting**	LiDAR+Camera	94.7	17.7	35.0	28.8	55.0	59.4	63.6	0.0	95.3	39.9	77.6	0.4	87.5	55.1	87.7	67.0	72.9	61.8	36.5	54.5
***RGBAL**	LiDAR+Camera	87.3	36.1	26.4	64.6	54.6	58.1	72.7	0.0	95.1	45.6	77.5	0.8	78.9	53.4	84.3	61.7	72.9	56.1	41.5	56.2
***PMF**	LiDAR+Camera	95.4	47.8	62.9	68.4	75.2	78.9	71.6	0.0	96.4	43.5	80.5	0.1	88.7	60.1	88.6	72.7	75.3	65.5	43.0	63.9
PCSCNet (backbone)	LiDAR	95.8	30.8	62.0	82.8	49.4	70.9	89.2	0.1	93.5	43.7	80.1	0.9	89.8	57.9	88.0	61.0	74.7	63.7	49.1	62.3
Ours	LiDAR+Camera	96.1	45.1	70.1	81.8	60.6	76.7	90.2	0.0	95.7	36.1	79.0	0.0	91.0	66.4	88.6	72.0	72.7	65.9	46.3	**65.0**

대 100 epoch까지 진행하였으며 초기 Learning rate는 0.001로 설정하였다. 데이터 증강 기법으로는 Flip, Scaling, Random jitter를 적용하였다.

4.3 실험 결과

제안하는 네트워크를 평가하기 위해 다양한 단일 라이다 기반 네트워크 및 카메라-라이다 융합 기반 네트워크와 성능을 비교하였다. 성능 비교는 Table 1에서 확인할 수 있으며, 별표 첨자가 붙은 네트워크의 성능은 PMF 논문이 도출한 결과를 인용한 것이다. 이때 각 클래스 별 평균 포인트 개수를 첫 번째 행에 명시하였다. Motorcycle의 경우 포인트 개수가 매우 희박하기 때문에 성능이 0에 수렴하는 것이 일반적이다. 네트워크의 전체 성능 비교는 백본 네트워크로 삼은 PCSCNet과의 비교, 다양한 공개 네트워크와의 비교로 분류할 수 있다.

먼저 백본 네트워크로 삼은 PCSCNet과 성능을 비교하면 mIoU가 2.7 % 향상하였다. 제안하는 모델의 기반이 되는 네트워크이기 때문에 카메라를 추가로 사용함으로써 얻을 수 있는 효과를 클래스별로 분석할 수 있다. 큰 성능 향상을 보이는 클래스는 Bicycle, Motorcycle, Person, Trunk로 각각 14.3 %, 8.1 %, 5.8 %, 9.0 %씩 증가하였다. 해당하는 클래스의 공통점은 크기가 작은 객체라는 것이다. 단일 라이다만 사용했을 때 작은 객체에 대해 포인트 희박성 문제가 있었고, 이로 인해 낮은 성능이 도출되었다. 하지만 카메라-라이다 융합으로 카메라에서 추출된 특징이 추가로 사용됨으로써 해당 문제가 보완되었음을 나타낸다. 작은 크기 객체에 대한 성능 향상은 Fig. 6에서 정성적으로 확인할 수 있다. 첫 번째 행은 SemanticKITTI가 제공하는 정답 데이터, 두 번째 행은 단일 라이다만 사용하여 도출한 예측 결과, 세 번째 행은 제안하는 센서융합 기반 네트워크로 도출한 예측 결과이다. 크기가 작고 원거리에 위치한 객체에 대해 센서융합 하였을 때 오분류 포인트가 감소하는 것을 확인할 수 있다.

다음으로, 다양한 공개 네트워크와의 성능을 비교해 보면 제안한 모델은 mIoU 65.0 %로 단일 라이다를 사용한 네트워크인 RandLANet[26], SqueezeSegV3[27], SPVNAS[28], Cylinder3D[29] 등 보다 높은 성능을 달성하였다. 또한 카메라-라이다 센서융합 기반의 네트워크인 PointPainiting[30], RGBAL, PMF보다 높은 성능을 달성하였다. 현재까지 SemanticKITTI 데이터셋에서 가장 높은 성능을 달성한 센서융합 기반의 네트워크는 PMF로 63.9 %의 성능을 보이고 있다. 하지만 제안하는 네트워크가 1.1 % 더 높은 성능을 도출함으로써, 센서융합 기반 네트워크 중에서는 최신 성능을 달성했다.

또한 제안하는 네트워크의 모듈 별 효과를 분석하기 위해 단계별 분석을 진행하였다. 정량적 분석 결과는 Table 2에서 확인 가능하다. Baseline, 즉 PCSCNet만 사용했을 때의 mIoU는 62.3 %이지만 카메라를 추가로 사용하였을 때 mIoU는 64.2 %로 1.9 %의 성능 향상을 보였

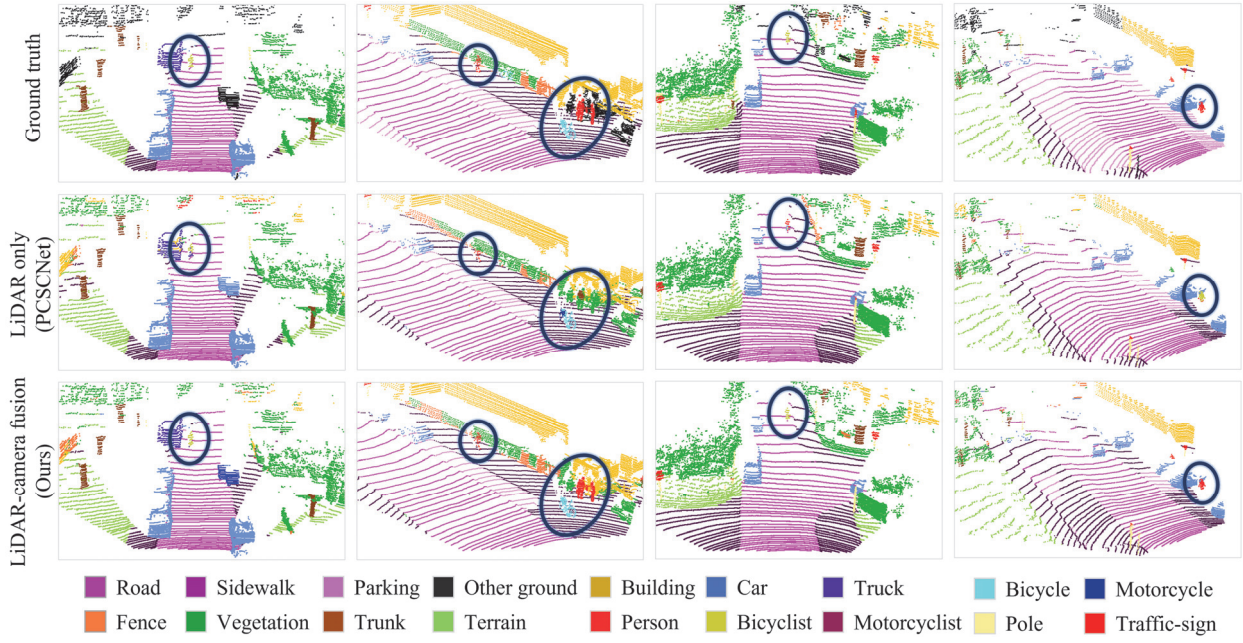

Fig. 6 Qualitative results on SemanticKITTI validation set

Table 2 Ablation study

	Baseline	LiDAR-camera fusion	Consider surrounding pixels	mIoU (%)
1	✓			62.3
2	✓	✓		64.2
3	✓	✓	✓	**65.0**

다. 이는 매칭되는 픽셀만을 사용하여 복셀과 픽셀을 1:1 대응시킨 결과이다. 이에 추가로 매칭되는 픽셀의 주변 3x3 영역까지 고려하였을 때 mIoU는 65.0 %로 Baseline 대비 2.7 %의 성능 향상을 보였다. 이를 통해 카메라 데이터 추가 사용의 효과와 융합 시 주변 픽셀 고려하는 방안의 효과를 검증하였다.

5. 결 론

본 논문은 복셀-픽셀 매칭 기반의 특징 지도 융합 모듈 포함하는 카메라-라이다 센서융합 기반 포인트 클라우드 의미론적 분할 모델을 제안하였다. 기존 센서융합 기반 네트워크의 정확한 특징 지도 매칭이 어려운 문제를 해결하고자, 포인트 클라우드 백본 네트워크로 복셀 기반의 모델인 PCSCNet을 적용하여 3차원 정보를 유지하였다. 이를 이용해 캘리브레이션 파라미터를 기반으로 정확한 특징 융합을 보장하는 특징 지도 융합 모듈을 제안하였다.

카메라를 추가로 사용함으로써 단일 라이다에서 취약하였던 Bicycle, Person 등의 작은 크기의 객체에 대한 성능을 크게 향상시켰으며, 본 네트워크의 우수성은 SemanticKITTI 데이터셋에서 검증되었다.

추가로, 본 네트워크에서 제안된 특징 융합 모듈의 활용성을 높이기 위해 카메라-라이다 센서융합 기반의 객체 검출, Instance 분할[31], Panoptic 분할[32] 등 다른 작업에도 적용할 예정이다.

후 기

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1C1C1007739). 또한, 이 성과는 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(P0020536, 2022년 산업혁신인재성장지원사업).

References

1) K. Jo, M. Lee, D. Kim, J. Kim, C. Jang, E. Kim, S. Kim, D. Lee, C. Kim, S. Kim, K. Huh and M. Sunwoo, “Overall Reviews of Autonomous Vehicle A1 – System Architecture and Algorithms,” The International Federation of Automatic Control (IFAC) International Autonomous Vehicles Symposium, Vol.46, No.10, pp.114-119, 2013.
2) J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss and J. Gall, “Semantickitti: A Dataset for Semantic Scene Understanding of Lidar Sequences,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.9297-9307, 2019.
3) Z. Zhuang, R. Li, K. Jia, Q. Wang, Y. Li and M. Tan, “Perception-Aware Multi-sensor Fusion for 3D Lidar Semantic Segmentation,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.16280-16290, 2021.
4) S. Albawi, T. Mohammed, A. T and S. AI-Zawi, “Understanding of a Convolutional Neural Network,” International Conference on Engineering and Technology (ICET), pp.1-6, 2017.
5) A. Milioto, I. Vizzo, J. Behley and C. Stachniss, “Rangenet++: Fast and Accurate Lidar Semantic Segmentation,” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.4213-4220, 2019.
6) J. Jeong, J. Park, C. Kim and K. Jo, “Modified RangeNet++ for Real-time Point Cloud Semantic Segmentation in Autonomous Driving Application,” KSAE Fall Conference Proceedings, p.1016, 2020.
7) B. Wu, A. Wan, X. Yue and K. Keutzer, “Squeezeseg: Convolutional Neural Nets with Recurrent Crf for Real-time Road-object Segmentation from 3D Lidar Point Cloud,” IEEE International Conference on Robotics and Automation (ICRA), pp.1887-1893, 2018.
8) T. Cortinhal, G. Tzelepis and E. Erdal Aksoy, “SalsaNext: Fast, uncertainty-aware Semantic Segmentation of LiDAR Point Clouds,” International Symposium on Visual Computing, pp.207-222, 2020.
9) Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong and H. Foroosh, “Polarnet: An Improved Grid Representation for Online Lidar Point Clouds Semantic Segmentation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9601-9610, 2020.
10) C. R. Qi, H. Su, K. Mo and L. J. Guibas, “Pointnet: Deep Learning on Point Sets for 3D Classification and Segmentation,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.652-660, 2017.

11) J. Tang, C. Deng and G. B. Huang, "Extreme Learning Machine for Multilayer Perceptron," IEEE Transactions on Neural Networks and Learning Systems, Vol.27, No.4, pp.809-821, 2015.
12) H. Thomas, C. Qi, J. E. Deschaud, B. Marcotegui, F. Goulette and LJ. Guibas, "Kpconv: Flexible and Deformable Convolution for Point Clouds," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.6411-6420, 2019.
13) A. Boulch, "ConvPoint: Continuous Convolutions for Point Cloud Processing," Computers & Graphics, Vol.88, pp.24-34, 2020.
14) L. Jiang, H. Zhao, S. Shi, S. Liu, CW. Fu and J. Jia, "Pointgroup: Dual-set Point Grouping for 3D Instance Segmentation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4867-4876, 2020.
15) F. Engelmann, M. Bokeloh, A. Fathi, B. Leibe and M. Nießner, "3D-mpa: Multi-proposal Aggregation for 3D Semantic Instance Segmentation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9031-9040, 2020.
16) J. Park, C. Kim and K. Jo, "PCSCNet: Fast 3D Semantic Segmentation of LiDAR Point Cloud for Autonomous Car using Point Convolution and Sparse Convolution Network," Expert Systems with Applications, 2022.
17) J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun and S. Pu, "Rpvnet: A Deep and Efficient Range-point-voxel Fusion Network for Lidar Point Cloud Segmentation," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.16024-16033, 2021.
18) H. Song, J. Cho, J. Ha and K. Jo, "Camera-LiDAR Fusion based Point Cloud Semantic Segmentation using Attention Network for Autonomous Vehicles," KSAE Fall Conference Proceedings, pp.568-568, 2021.
19) Z. Zhong, C. Zhang, Y. Liu and Y. Wu, "Viaseg: Visual Information Assisted Lightweight Point Cloud Segmentation," IEEE International Conference on Image Processing (ICIP), pp.1500-1504, 2019.
20) K. El Madawi, H. Rashed, A. EI Sallab, O. Nasr, H. Kamel and S. Yogamani, "Rgb and Lidar Fusion Based 3D Semantic Segmentation for Autonomous Driving," IEEE Intelligent Transportation Systems Conference (ITSC), pp.7-12, 2019.
21) G. Krispel, M. Opitz, G. Waltner, H. Possegger and H. Bischof, "Fuseseg: Lidar Point Cloud Segmentation Fusing Multi-modal Data," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp.1874-1883, 2020.
22) B. Liu, M. Wang, H. Foroosh, M. Tappen and M. Pensky, "Sparse Convolutional Neural Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.806-814, 2015.
23) K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.770-778, 2016.
24) A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, Vol.60, No.6, pp.84-90, 2012.
25) A. Geiger, P. Lenz and R. Urtasun, "Are We Ready for Autonomous Driving? the Kitti Vision Benchmark Suite," IEEE Conference on Computer Vision and Pattern Recognition, pp.3354-3361, 2012.
26) Q. Hu, B. Yand, L. Xie, S. Rosa, Y. Guo, Z. Wang and A. Markham, "Randla-net: Efficient Semantic Segmentation of Large-scale Point Clouds," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.11108-11117, 2020.
27) C. Xu, B. Wu, Z. Wang, W. Zhan, P. Vajda, K. Keutzer and M. Tomizuka, "Squeezesegv3: Spatially-adaptive Convolution for Efficient Point-cloud Segmentation," In European Conference on Computer Vision, pp.1-19, 2020.
28) H. Tand, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang and S. Han, "Searching Efficient 3D Architectures with Sparse Point-voxel Convolution," In European Conference on Computer Vision, pp.685-702, 2020.
29) H. Zhou, X. Zhu, X. Song, Y. Ma, Z. Wang, H. Li and D. Lin, "Cylinder3D: An Effective 3D Framework for Driving-scene Lidar Semantic Segmentation," arXiv Preprint arXiv:2008.01550, 2008.
30) S. Vora, A. H. Lang, B. Helou and O. Beijbom, "Pointpainting: Sequential Fusion for 3D Object Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4604-4612, 2020.
31) A. M. Hafiz and G. M. Bhat, "A Survey on Instance Segmentation: State of the Art," International Journal of Multimedia Information Retrieval, pp.171-189, 2020.
32) A. Kirillov, K. He, R. Girshick, C. Rother and P. Dollár, "Panoptic Segmentation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9404-9413, 2019.