# 3차원 포인트 클라우드의 의미적 분할을 위한 멀티-모달 교차 주의집중

배혜림, 김인철
경기대학교 컴퓨터과학과
thvk654@kyonggi.ac.kr, kic@kyonggi.ac.kr

# Multi-Modal Cross Attention for 3D Point Cloud Semantic Segmentation

HyeLim Bae, Incheol Kim
Department of Computer Science, Kyonggi University

요 약

3차원 포인트 클라우드의 의미적 분할은 환경을 구성하는 물체 단위로 포인트 클라우드를 분할하는 작업으로서, 환경의 3차원적 구성을 이해하고 환경과 상호작용에 필수적인 시각 지능을 요구한다. 본 논문에서는 포인트 클라우드에서 추출하는 3차원 기하학적 특징과 함께 멀티-뷰 영상에서 추출하는 2차원 시각적 특징들도 활용하는 새로운 3차원 포인트 클라우드 의미적 분할 모델 MFNet을 제안한다. 제안 모델은 서로 이질적인 2차원 시각적 특징과 3차원 기하학적 특징의 효과적인 융합을 위해, 새로운 중기 융합 전략과 멀티-모달 교차 주의집중을 이용한다. 본 논문에서는 ScanNetV2 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해, 제안 모델 MFNet의 우수성을 입증한다.

## 1. 서론

포인트 클라우드(point cloud)는 포인트들의 집합으로서 3차원 물체나 환경을 표현하는 가장 기본적인 데이터 표현법이다. 최근 들어 자율 주행(autonomous driving), 서비스 로봇(service robot), 증강 현실(augmented reality)과 같이 3차원 환경에서 동작하는 체화 인공지능(embodied AI)이 발전함에 따라, 포인트 클라우드를 대상으로 하는 3차원 물체 탐지(3D object detection)와 3차원 의미적 분할(3D semantic segmentation) 기술들이 주목받고 있다. 특히 이 중에서 3차원 포인트 클라우드의 의미적 분할은 환경을 구성하는 물체 단위로 포인트 클라우드를 분할하는 작업으로서, 이 작업을 위해서는 포인트 클라우드의 각 포인트마다 이 포인트가 속한 물체를 예측해서 분류 레이블(class label)을 할당해야 한다. 따라서 이 작업은 지능형 에이전트에게 자신이 활동해야 할 환경의 3차원적 구성을 이해하고, 환경과 상호작용하기 위해 필수적인 시각 지능을 요구한다.

3차원 포인트 클라우드의 의미적 분할에 관한 기존 연구들은 크게 포인트 클라우드의 3차원 기하학적 특징만을 이용하는 방식과, 포인트 클라우드의 기하학적 특징 외에 멀티-뷰 RGB-D 입력 영상들에서 추출하는 2차원 시각적 특징을 함께 이용하는 방식들로 크게 나눌 수 있다[1]. PointNet++[1], SPH3D-GCN[1], PTV2[2]와 같이 포인트 클라우드의 3차원 기하학적 특징만을 이용한 분할 방식은 각 포인트의 xyz 위치 정보로부터 해당 포인트 주변의 지역적인 기하학적 특징을 추출한 후, 이를 바탕으로 각 포인트의 분류 레이블을 결정한다. 하지만 정규 합성곱 신경망(convolutional neural network, CNN)으로 특징 추출 학습이 용이했던 2차원 영상(image)이나 1차원 자연어 텍스트(text)와는 달리, 포인트 클라우드는 구성 포인트들이 불규칙적으로 분포하고 있어 정규 합성 곱을 그대로 적용해서는 특징 추출이 어렵다는 문제점이 있다. 이러한 문제점을 해결하기 위해 기존 연구들에서는 새롭게 설계한 포인트 합성곱 신경망들(point CNN)들을 이용하는 방식, 근거리 이웃 포인트들과의 관계에 기초한 그래프 신경망(graph neural network, GNN)을 이용하는 방식 등을 제안하였다. 하지만 3차원 기하학적 특징만을 이용하는 포인트 클라우드 의미적 분할 방식들은 포인트들의 희소성(sparsity) 문제와 함께 컬러 영상의 풍부한 시각적 특징들을 활용하지 못하는 한계점이 공통적으로 존재한다.

* 본 연구는 정보통신기획평가원의 재원으로 정보통신방송 기술개발사업의 지원을 받아 수행한 연구 과제(클라우드에 연결된 개별 로봇 및 로봇그룹의 작업 계획 기술 개발, 2020-0-00096)입니다.

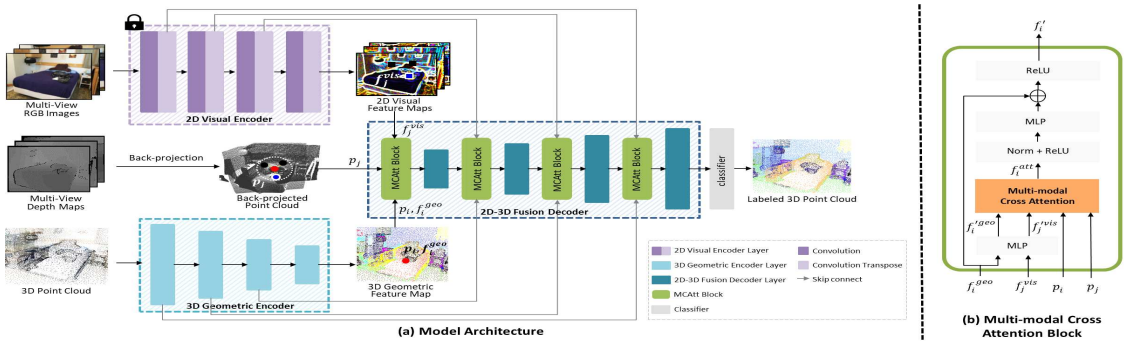
반면에 포인트 클라우드의 3차원 기하학적 특징과 멀티-뷰 RGB-D 영상들에서 추출하는 2차원 시각적 특징을 함께 이용하는 MVPNet[1], SAFNet[3], BPNet[4] 등의 의미적 분할 방식은 포인트들의 희소성 문제를 극복하고 분할에 더 많은 풍부한 멀티-모달 특징정보들을 이용할 수 있어 높은 분할 성능을 기대할 수 있지만, 서로 이질적인 3차원 기하학적 특징과 2차원 시각적 특징들을 어떻게 융합해야 시너지 효과를 얻을 수 있느냐 하는 새로운 도전 과제가 있다. MVPNet[1]의 연구에서는 두 특징에 대한 초기 융합(early fusion)을 제안하였으나, 양쪽 각각의 고유한 특징을 충분히 추출하기 어렵다는 문제점이 있다. 이를 극복하기 위해 SAFNet[3]의 연구에서는 3차원 기하학적 특징과 2차원 시각적 특징들을 서로 독립적으로 추출한 다음 후기 융합(late fusion)을 수행하였으나, 이 경우에는 양쪽 특징 간의 내밀한 결합이 어렵다는 단점이 존재한다. 이러한 점들을 고려하여 BPNet[4]의 연구에서는 3차원 기하학적 특징과 2차원 시각적 특징들 간의 중기 융합(intermediate fusion)을 시도하였다. 하지만 두 특징의 단순 결합(concatenate) 후 1x1 합성곱(1x1 convolution)을 통해 두 특징을 융합하기 때문에, 양쪽 특징 간의 연관성을 충분히 반영해 융합이 이루어질 수 없는 한계점이 있다.

본 논문에서는 이러한 기존 방식들의 한계성을 극복하고자, 새로운 3차원 포인트 클라우드의 의미적 분할 모델을 제안한다. 제안 모델에서는 포인트 클라우드의 3차원 기하학적 특징과 멀티-뷰 RGB-D 영상들의 2차원 시각적 특징들을 위해 새로운 중기 융합 전략과 멀티-모달 교차 주의집중(multi-modal cross attention) 기반의 특징 융합 방식을 제시한다. 또한 제안 모델은 새로운 그룹 벡터 어텐션(grouped vector attention)과 위치 인코딩(position encoding)을 이용하는 Transformer 기반의 포인트 특징 추출기인 PTV2[2]를 채용함으로써, 포인트 클라우드의 더 정확한 3차원 기하학적 특징을 이용할 수 있다. 본 논문에서는 벤치마크 데이터 집합 ScanNetV2를 이용한 정량적 평가 실험을 통해 제인 모델의 우수성을 입증한다.

## 2. 3차원 의미적 분할 모델

### 2.1 모델 개요

본 논문에서는 새로운 멀티-모달 특징을 기반으로 하는 3차원 포인트 클라우드 의미적 분할 모델인 MFNet (Multi-Modal Fusion Network)을 제안한다. 제안 모델의 전체 구조는 (그림 1)의 (a)와 같이, 멀티-뷰 영상에서 시각적 특징을 추출하는 2차원 시각적 인코더(2D Visual




(그림 1) 제안모델의 구성

Encoder), 포인트 클라우드에서 기하학적 특징을 추출하는 3차원 기하학적 인코더(3D Geometric Encoder), 2차원 시각적 특징과 3차원 시각적 특징을 융합하는 2차원-3차원 융합 디코더(2D-3D Fusion Decoder)로 구성된다. 특히 2차원-3차원 융합 디코더는 2차원 인코더 계층과 3차원 인코더 계층 각각의 특징정보 간의 연관성을 반영하는 (그림 1)의 (b)와 같은 멀티-모달 교차 주의집중 블록(Mulit-Modal Cross Attention Block, MCAtt Block)들과 디코더 계층들로 구성되어 있다.

제안 모델의 2차원 시각적 인코더는 멀티-뷰 RGB 영상들로부터 각 계층별로 특화된 시각적 특징 지도들을 추출한다. 이를 위해 ImageNet 데이터 집합으로 사전 학습된 ResNet34의 합성곱 계층을 백본(backbone)으로 채용한다. 한편, 멀티-뷰 RGB 영상들에서 계층적으로 추출되는 2차원 시각적 특징 지도들은 나중에 포인트 클라우드에서 추출되는 3차원 기하학적 특징들과의 융합을 위해, 인코더의 각 계층마다 추출되는 시각적 특징 지도들의 해상도를 입력 RGB-D 영상들의 해상도와 동일하게 유지해야 한다. 따라서 인코더의 각 계층 블록마다 합성곱(convolution) 계층외에 합성곱 전치(convolution transpose) 계층을 추가로 삽입하여 시각적 특징 지도의 해상도를 일정하게 유지한다.

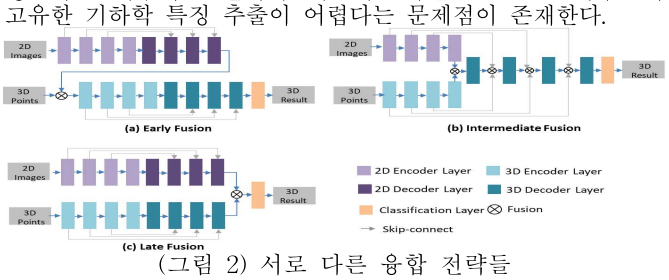
반면, 제안 모델의 3차원 기하학적 인코더는 입력 포인트 클라우드로부터 각 계층 블록마다 다운 샘플링(down sampling)된 포인트 클라우드들과 이에 대응하는 3차원 기하학적 특징 지도들을 획득한다. 이를 위해 3차원 기하학적 인코더의 각 계층 블록들은 다운 풀링(down pooling) 계층과 포인트 특징 추출 계층으로 구성된다. 제안 모델의 3차원 기하학적 인코더에서는 효율성이 높은 그룹 벡터 어텐션(grouped vector attention)과 새로운 위치 인코딩(position encoding)을 이용하는 Transformer 기반의 PTV2[2] 인코더를 포인트 특징 추출기로 채용하였다.

한편, 제안 모델에서는 고해상도의 멀티-뷰 2차원 RGB 영상들과 저밀도의 3차원 입력 포인트 클라우드 간의 원활한 대응과 융합을 위해, 고해상도의 멀티-뷰 깊이 지도(depth map)들로부터 3차원 공간으로의 역투영(back-projection) 과정을 통해 고밀도 포인트 클라우드를 추가로 생성한다. 이렇게 역투영으로 생성된 고밀도 포인트 클라우드는 2차원-3차원 융합 디코더에서 MCAtt Block을 통해 다운 샘플링된 저밀도 포인트 클라우드의 각 포인트를 중심으로 고밀도 이웃 포인트들(dense neighboring points)의 2차원 시각적 특징들을 집계(aggregation)하는데 이용된다.

2차원-3차원 융합 디코더는 멀티-모달 교차 주의집중 블록(MCAtt Block), 업 풀링(up pooling) 계층, 그리고 PTV2[2] 디코더 계층으로 구성된 다수의 계층 블록들로 이루어져 있다. 각 계층 블록의 MCAtt Block은 연관성에 따라 고밀도 이웃 포인트들의 2차원 시각적 특징들을 집계한 후 이것을 3차원 기하학적 특징과 결합함으로써, 해당 계층의 저밀도 포인트 클라우드의 각 포인트를 위한 멀티-모달 특징을 계산한다. 이와 같이 1차 융합된 각 포인트의 멀티-모달 특징은 PTV2[2] 디코더 계층을 거치면서 다시 한번 심층 융합된다. 2D-3D 융합 디코더의 마지막 계층은 일종의 분류기(classifier) 역할을 수행함으로써, 각 포인트의 의미적 레이블(semantic label)을 결정한다.

## 2.2 융합 전략

시각적 특징과 기하학적 특징의 융합 전략은 크게 초기 융합(early fusion), 중기 융합(intermediate fusion), 후기 융합(late fusion)으로 나눌 수 있다. 먼저 초기 융합 전략은 (그림 2)의 (a)와 같이 3차원 포인트 클라우드가 3차원 인코더-디코더을 거치기 전, 멀티-뷰 영상들로부터 2차원 인코더-디코더를 통해 추출된 시각적 특징 지도를 포인트 클라우드에 먼저 융합하는 방식이다. 초기 융합 전략은 다른 융합 전략보다 시각적 특징과 기하학적 특징 간의 결합 강도가 강하다는 장점이 있다. 하지만 초기 융합은 포인트 클라우드의 고유한 위치 정보가 초기부터 손실되기 때문에 3차원 포인트 클라우드의 고유한 기하학 특징 추출이 어렵다는 문제점이 존재한다.




(그림 2) 서로 다른 융합 전략들

반면, 중기 융합 전략은 (그림 2)의 (b)와 같이, 2차원 인코더 계층마다의 시각적 특징과 3차원 인코더 계층마다의 기하학적 특징을 각 계층별 융합 후, 이를 스킵 연결을 통해 3차원 디코더에 계층별로 결합하는 전략이다. 중기 융합 전략은 두 인코더의 계층별 특징 결합을 통해 다양한 차원의 시각적 특징과 기하학적 특징을 반영할 수 있다. 하지만 두 인코더 계층마다 추출되는 특징 지도의 크기가 다르기 때문에 두 특징 간의 매핑이 어렵다는 문제가 존재한다. 따라서 이를 해결하기 위한 정교한 설계가 요구된다.

마지막 후기 융합 전략 (그림 2)의 (c)와 같이, 포인트 클라우드가 3차원 인코더-디코더를 모두 통과한 기하학적 특징과 2차원 인코더-디코더를 통해 추출된 시각적 특징을 마지막에 융합하는 전략이다. 후기 융합 전략은 2차원과 3차원 각각의 도메인에서 독립적으로 특징을 추출함으로써, 각 도메인의 고유한 특징 추출이 가능하다. 하지만 다른 융합 전략에 비해 시각적 특징과 기하학적 특징의 충분한 융화 과정이 이루어지지 않아 두 특징 간의 결합 강도가 약하다는 문제점이 있다. 본 논문의 제안 모델에서는 서로 다른 모달의 입력 데이터로부터 추출되는 2차원 시각적 특징과 3차원 기하학적 특징의 고유성을 최대한 보장하면서도 서로 밀접하게 융합할 수 있는 중기 융합 전략을 채택하였다.

## 2.3 멀티-모달 교차 주의집중 블록

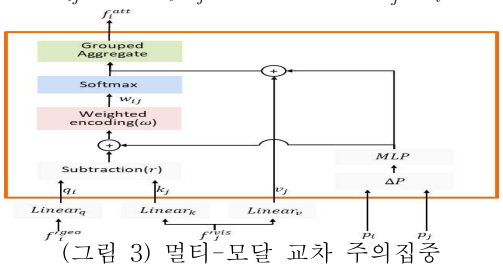
(그림 1)의 (b)와 같이, 제안 모델의 멀티-모달 교차 주의집중 블록(MCAtt Block)은 3차원 기하학적 인코더의 각 계층에서 추출되는 저밀도 포인트 클라우드의 기하학적 특징 $f_i^{geo}$을 중심으로, 고밀도 포인트 클라우드 내 고밀도 이웃 포인트 $p_j$들의 시각적 특징 $f_j^{vis}$들과의 연관성을 반영하여 이들을 집계한 후 3차원 기하학적 특징과 결합함으로써, 저밀도 포인트 클라우드의 각 포인트 $p_i$의 멀티-모달 특징 $f_i'$을 생성한다. 이때, 저밀도 포인트 클라우드의 기하학적 특징 $f_i^{geo}$을 중심으로, 고밀도 이웃 포인트 $p_j$들의 시각적 특징 $f_j^{vis}$들을 집계하는 방식은 (그림 3)과 같은 멀티-모달 교차 주의집중(Multi-modal Cross Attention)을 이용한다. 먼저 (식 1)과 같이, 다층 퍼셉트론을 거친 $f_i'^{geo}$으로부터 주의집중의 중심이 되는 쿼리 $q_i$를, $f_j'^{vis}$으로부터 쿼리와의 연관성을 계산하기 위한 키 $k_j$와 이에 대해 대응되는 벨류 $v_j$을 구한다.

$$q_i = Linear_q(f_i'^{geo}), k_j = Linear_k(f_i'^{vis}), v_j = Linear_v(f_i'^{vis}) \quad \text{(식 1)}$$

이후 (식 2)와 같이, 저밀도 포인트 $p_i$와 고밀도 이웃 포인

트 $p_j$들로부터 구한 위치 인코딩(position encoding) $\delta$는 쿼리 $q_i$와 키 $k_j$ 간의 차를 통해 획득한 관계 특징 $\gamma(q_i, k_j)$과 더한 후 그룹화된 가중치 벡터 $w_{ij}$들을 구한다. 이때 그룹화된 가중치 벡터 $w_{ij}$는 $c$개의 채널을 갖는 관계 특징 $\gamma(q_i, k_j)$의 채널을 $g$개의 그룹으로 나눈 후, 각 그룹마다의 주의집중 가중치 값을 갖는 벡터이다.

$$w_{ij} = \omega(r(q_i, k_j) + \delta),\ \delta = MLP(p_j - p_i) \qquad (식\ 2)$$




(그림 3) 멀티-모달 교차 주의집중

다음 (식 3)과 같이, 그룹화된 가중치 벡터 $w_{ij}$는 활성함수인 $Softmax$를 통해 0-1 사이의 확률값으로 변환하고, 이를 위치 인코딩 $\delta$을 반영한 밸류 $v_j$에 반영 및 집계하여 최종적인 주의집중된 2차원 시각적 특징 $f_i^{att}$을 구한다. 이때, 밸류 $v_j$의 c개의 채널 중 같은 그룹에 속한 채널은 동일한 주의집중 가중치 값을 반영한다.

$$f_i^{att} = \sum_{p_j}^{M(p_i)} \sum_{n=1}^{g} \sum_{m=1}^{c/g} Softmax(w_{ij})_n (v_j + d)^{nc/g+m} \qquad (식\ 3)$$

## 3. 구현 및 실험

본 논문에서의 제안 모델은 Pytorch 라이브러리를 통해 구현되었으며, Ubuntu 20.04 LTS 운영체제와 2개의 GeForce RTX 3090 GPU가 탑재된 하드웨어에서 학습 및 실험을 진행하였다. 제안 모델의 학습과 성능 평가를 위해, 2차원과 3차원 데이터를 모두 포함하는 실내 장면의 RGB-D 데이터 세트인 ScanNetV2 벤치마크 데이터를 사용하였고, 이 중 1201개의 장면 데이터는 모델의 학습에, 312개의 장면 데이터는 모델의 평가에 사용하였다. 이때 사전 학습된 2차원 시각적 인코더를 통해 시각적 특징을 추출하였고, 최적화 알고리즘 Adam과 손실함수 크로스엔트로피(CrossEntropy)를 사용하여 전체 모델에 대한 종단 간 학습(end-to-end)을 진행하였다.

첫 번째 실험은 제안 모델의 중기 융합 전략의 타당성을 입증하기 위한 실험이다. 위 실험에서는 제안 모델에서 사용된 백본과 멀티-모달 교차 주의집중 방식은 동일하게 유지하되, 융합 전략은 (그림 2)의 (a), (b), (c)와 같이 초기 융합(early fusion), 중기 융합(intermediate fusion), 후기 융합(late fusion) 전략을 각각 적용한 성능을 비교한다.

<표 1> 융합 전략에 따른 분할 성능 비교

| Strategy | Visual Feature | Geometry Feature | oAcc (%) | mIoU (%) |
|---|---|---|---|---|
| (a) Early Fusion | en-decoding | xyz | 88.48 | 68.47 |
| (b) Intermediate Fusion | layer-wise encoding | layer-wise encoding | **89.70** | **71.87** |
| (c) Late Fusion | en-decoding | en-decoding | 88.89 | 69.54 |

<표 1>의 실험 결과를 살펴보면 제안 모델 MFNet과 같이, (b)중기 융합 전략이 (a) 초기 융합 전략과 (b) 후기 융합 전략보다 성능 척도 oAcc, mIoU 면에서 모두 가장 높은 성능을 보인 것을 확인할 수 있다. (b) 중기 융합 전략을 이용한 모델이 (a) 초기 융합 전략을 이용한 모델보다 성능 척도 oAcc, mIoU 면에서 약 1.37%, 4.96%의 성능 향상을, (c) 후기 융합 전략을 이용한 모델보다 약 0.91%, 3.35%의 성능 향상을 보였다. 위 실험을 통해 제안 모델에서 채택한 중기 융합 전략의 유효성을 확인할 수 있었다.

두 번째 실험은 2차원 시각적 특징과 3차원 기하학적 특징 간의 융합을 위한 제안 모델의 멀티-모달 교차 주의집중 방식의 우수성을 입증하기 위한 실험이다. 이 실험에서는 MCAtt Block을 제외한 나머지는 제안 모델과 동일하게 구성하되, 두 특징이 융합되는 MCAtt Block 부분만을 (a) 단순 결합방식을 사용한 경우(concatenate), (b) 단순 결합 후 선형 변환 방식을 이용한 경우(concatenate+linear transformation), (c) 벡터 어텐션 방식을 적용한 경우(vector attention), 그리고 제안 모델과 같이 (d) 그룹 벡터 어텐션 방식을 이용하는 경우(grouped vector attention)에 대해 각각 분할 성능을 비교한다.

<표 2> 멀티-모달 특징 융합 방식에 따른 분할 성능 비교

| 2D-3D Feature Fusion | oAcc(%) | mIoU(%) |
|---|---|---|
| (a) concatenate | 89.53 | 71.04 |
| (b) concatenate + linear transformation | 89.34 | 71.06 |
| (c) vector attention | 89.52 | 71.10 |
| (d) grouped vector attention | **89.70** | **71.87** |

<표 2>의 실험 결과를 살펴보면, 제안 모델과 같이 (d) 그룹 벡터 어텐션 방식을 이용한 경우가 다른 방식을 이용한 경우들에 비해 가장 높은 성능을 보이는 것을 알 수 있다. 이와 같은 실험 결과를 통해, 2차원 시각적 특징과 3차원 기하학적 특징 융합 시, 그룹 벡터 어텐션 기반의 멀티-모달 특징 융합 방식이 성능 향상에 긍정적인 영향을 미치는 것을 확인할 수 있었다.

세 번째 실험은 본 논문에서 제안한 MFNet 모델의 우수성을 입증하기 위해, 기존의 대표적인 3차원 의미적 분할 모델들과 성능을 비교한 실험이다. 이 실험에서는 3차원 컬러 포인트 클라우드의 기하학적 특징만을 이용하는 모델들(Colored Point Cloud [rgb+xyz])[1,2]. 멀티-뷰 RGB-D 영상들의 시각적 특징과 포인트 클라우드의 기하학적 특징을 함께 이용하는 모델들(Multi-View RGB-D Images+Point Cloud[xyz])[1,3,4]을 비교한다. 이때 PTV2[2]*와 BPNet[4]* 모델은 앞서 언급한 제안 모델의 구현 환경과 동일한 환경에서 평가한 성능을 나타낸다.

<표 3> 3차원 의미적 분할 모델들 간의 성능 비교

| Model | Input | mIoU(%) |
|---|---|---|
| PointNet++[1] | Colored Point Cloud[rgb+xyz] | 33.9 |
| SPH3D-GCN[1] | | 61.0 |
| PTV2[2]* | | 64.36 |
| MVPNet[1] | Multi-View RGB-D Images + Point Cloud[xyz] | 64.1 |
| SAFNet[3] | | 65.4 |
| BPNet[4]* | | 70.17 |
| **MFNet(Ours)** | | **71.87** |

<표3>의 실험 결과를 살펴보면 제안 모델 MFNet이 다른 분할 모델들보다 가장 우수한 성능을 보이는 것을 확인할 수 있다. 실험 결과를 자세히 살펴보면, 컬러 포인트 클라우드만을 이용한 모델들 중에서는 PTV2[2]의 모델이 우수하며, 멀티-뷰 RGB-D 영상들과 포인트 클라우드의 기하학적 특징을 함께 이용하는 모델들 중에서는 제안 모델이 가장 높은 성능을 보였다. 또한, 포인트 특징 추출기인 PTV2[2]만을 이용한 기존 분할 모델보다 PTV2[2]를 백본의 일부로 사용하지만 2차원 시각적 특징을 포함한 멀티-모달 특징과 더불어 특징 융합을 위해 MCAtt Block 등을 추가로 도입한 제안 모델 MFNet이 mIoU 측면에서 약 11.67%의 성능 향상을 보였다. 이를 통해 제안 모델 MFNet의 우수성을 재확인할 수 있었다.

## 4. 결론

본 논문에서는 새로운 3차원 포인트 클라우드의 의미적 분할 모델 MFNet을 제안하였다. 이 제안 모델은 멀티-뷰 영상에서 추출하는 2차원 시각적 특징들과 포인트 클라우드에서 추출하는 3차원 기하학적 특징 간의 중기 융합 전략과 멀티-모달 교차 주의집중 블록(MCAtt Block)을 적용함으로써, 의미적 분할 성능을 향상시켰다. 또한, 본 논문에서는 ScanNetV2 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해, 제안 모델의 우수성을 입증하였다.

## 참고문헌

[1] Y. He, H. Yu, and X. Liu, et. al, "Deep Learning based 3D Segmentation: A Survey," *arXiv preprint arXiv:2103.05423,* 2021.

[2] X. Wu, Y. Lao, and L. Jiang, et al., "Point Transformer V2: Grouped Vector Attention and Partition-based Pooling," *arXiv preprint arXiv:2210.05666,* 2022.

[3] L. Zhao, J. Lu, and J. Zhou, "Similarity-Aware Fusion Network for 3D Semantic Segmentation," *Proc. of IROS,* 2021, pp.1585-1592.

[4] W. Hu, H. Zhao, and L. Jian, et al., "Bidirectional Projection Network for Cross Dimension Scene Understanding," *Proc. of CPVR,* 2021, pp.14373-14382.