KIPS Trans. Softw. and Data Eng.
Vol.12, No.12 pp.505~518
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.12.505

Effective Multi-Modal Feature Fusion for 3D Semantic Segmentation with Multi-View Images

Hye-Lim Bae† · Incheol Kim††

ABSTRACT

3D point cloud semantic segmentation is a computer vision task that involves dividing the point cloud into different objects and regions by predicting the class label of each point. Existing 3D semantic segmentation models have some limitations in performing sufficient fusion of multi-modal features while ensuring both characteristics of 2D visual features extracted from RGB images and 3D geometric features extracted from point cloud. Therefore, in this paper, we propose MMCA-Net, a novel 3D semantic segmentation model using 2D-3D multi-modal features. The proposed model effectively fuses two heterogeneous 2D visual features and 3D geometric features by using an intermediate fusion strategy and a multi-modal cross attention-based fusion operation. Also, the proposed model extracts context-rich 3D geometric features from input point cloud consisting of irregularly distributed points by adopting PTv2 as 3D geometric encoder. In this paper, we conducted both quantitative and qualitative experiments with the benchmark dataset, ScanNetv2 in order to analyze the performance of the proposed model. In terms of the metric mIoU, the proposed model showed a 9.2% performance improvement over the PTv2 model using only 3D geometric features, and a 12.12% performance improvement over the MVPNet model using 2D-3D multi-modal features. As a result, we proved the effectiveness and usefulness of the proposed model.

Keywords : 3D Semantic Segmentation, Point Cloud, Multi-View RGB-D Images, 2D-3D Feature Fusion

멀티-뷰 영상들을 활용하는 3차원 의미적 분할을 위한 효과적인 멀티-모달 특징 융합

배 혜 림† · 김 인 철††

요 약

3차원 포인트 클라우드 의미적 분할은 각 포인트별로 해당 포인트가 속한 물체나 영역의 분류 레이블을 예측함으로써, 포인트 클라우드를 서로 다른 물체들이나 영역들로 나누는 컴퓨터 비전 작업이다. 기존의 3차원 의미적 분할 모델들은 RGB 영상들에서 추출하는 2차원 시각적 특징과 포인트 클라우드에서 추출하는 3차원 기하학적 특징의 특성을 충분히 고려한 특징 융합을 수행하지 못한다는 한계가 있다. 따라서, 본 논문에서는 2차원-3차원 멀티-모달 특징을 이용하는 새로운 3차원 의미적 분할 모델 MMCA-Net을 제안한다. 제안 모델은 중기 융합 전략과 멀티-모달 교차 주의집중 기반의 융합 연산을 적용함으로써, 이질적인 2차원 시각적 특징과 3차원 기하학적 특징을 효과적으로 융합한다. 또한 3차원 기하학적 인코더로 PTv2를 채용함으로써, 포인트들이 비-정규적으로 분포한 입력 포인트 클라우드로부터 맥락정보가 풍부한 3차원 기하학적 특징을 추출해낸다. 본 논문에서는 제안 모델의 성능을 분석하기 위해 벤치마크 데이터 집합인 ScanNetv2을 이용한 다양한 정량 및 정성 실험들을 진행하였다. 성능 척도 mIoU 측면에서 제안 모델은 3차원 기하학적 특징만을 이용하는 PTv2 모델에 비해 9.2%의 성능 향상을, 2차원-3차원 멀티-모달 특징을 사용하는 MVPNet 모델에 비해 12.12%의 성능 향상을 보였다. 이를 통해 본 논문에서 제안한 모델의 효과와 유용성을 입증하였다.

키워드 : 3차원 의미적 분할, 포인트 클라우드, 멀티-뷰 RGB-D 영상들, 2차원-3차원 특징 융합

1. 서 론

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 클라우드(point cloud)를 이용한 3차원 물체 탐지(3D object detection), 3차원 의미적 분할(3D semantic segmentation, 3DSS), 3차원 개체 분할(3D instance segmentation, 3DIS), 3차원 장면 그래프 생성(3D scene graph generation, 3DSGG) 등과 같은 3차원 장면 이해(3D scene understanding) 기술들이 주목받고 있다. 이 중에서 3차원 의미적 분할(3DSS)은 포인트 클라우드(point cloud)를

※ 본 연구는 정보통신기획평가원의 재원으로 정보통신방송 기술개발사업의 지원을 받아 수행한 연구 과제(No. 2020-0-00096 클라우드에 연결된 개별 로봇 및 로봇그룹의 작업 계획 기술 개발)입니다.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 종신회원 : 경기대학교 컴퓨터공학부 교수
Manuscript Received : August 16, 2023
First Revision : September 19, 2023
Accepted : October 7, 2023
* Corresponding Author : Incheol Kim(kic@kyonggi.ac.kr)

구성하는 각 포인트별로 해당 물체의 분류 레이블(class label)을 예측하는 시각 인식 작업이다. 이러한 3차원 의미적 분할 능력은 한 에이전트가 자신이 활동해야 할 환경의 3차원 장면 구성을 심층적으로 이해하고, 해당 환경과 효과적으로 상호작용하기 위해서는 필수적으로 요구되는 시각 지능이다.

포인트 클라우드 기반의 3차원 의미적 분할 연구에 앞서, 입력 영상을 구성하는 픽셀(pixel) 단위로 분류 레이들을 할당하는 영상 기반의 2차원 의미적 분할(image-based 2D semantic segmentation)에 관한 다양한 연구들이 있었다[1-4]. 일반적인 영상 기반의 2차원 의미적 분할 모델은 Fig. 1A와 같이, 입력 RGB 영상으로부터 2차원 시각적 인코딩을 통해 시각적 특징(visual feature)을 추출하고 이를 바탕으로 다시 2차원 시각적 디코딩 과정을 통해 픽셀 단위의 분할을 수행한다. 초기의 영상 기반 2차원 의미적 분할 연구들에서는 입력 영상에 관한 시각적 인코딩 혹은 디코딩을 위해 주로 합성곱 신경망(convolution neural network, CNN) 구조를 이용하였다[1,2]. 하지만 그 후 연구들에서는 자연어 처리 분야에서 큰 효과를 보여준 트랜스포머(Transformer) 신경망 구조를 이용한 2차원 의미적 분할 모델들[3,4]이 활발히 소개되었다.

한편, 포인트 클라우드 기반의 3차원 의미적 분할을 위한 기존 모델들은 대부분 Fig. 1B와 같이, 입력 포인트 클라우드에 대한 3차원 기하학적 인코딩 과정을 통해 3차원 기하학적 특징(3D geometric feature)을 추출한 후, 이를 바탕으로 3차원 기하학적 디코딩 과정을 거쳐 각 포인트마다 분류 레이블 예측하는 방식을 적용하였다[5-12]. 하지만 이러한 3차원 기하학적 특징 기반의 분할 방식은 (1) RGB-D 영상을 구성하는 픽셀들에 비해 포인트 클라우드를 구성하는 포인트들이 대부분 불규칙적으로 분포할 뿐만 아니라 상대적으로 더 희소

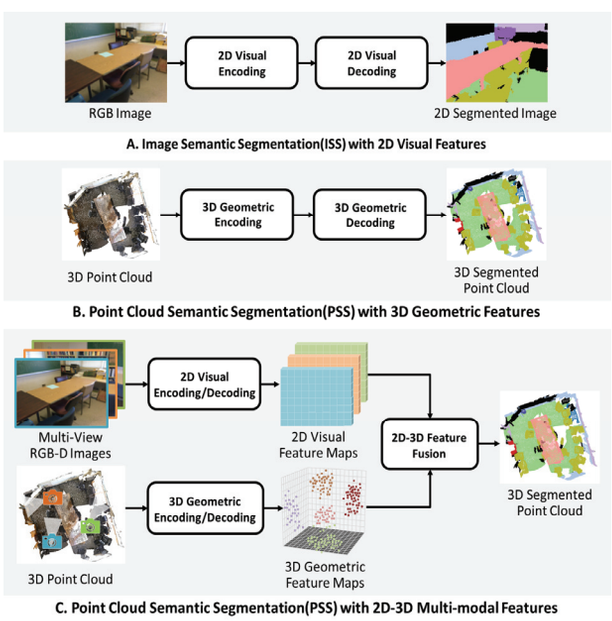

Fig. 1. Different Types of Semantic Segmentations

(sparse)하여, 물체별 혹은 물체 부품별로 세밀한 기하학적 특징을 얻기 어렵다는 문제점과 (2) RGB 컬러 영상이 갖는 색상 또는 텍스쳐와 같은 풍부한 시각적 특징들을 충분히 활용하지 못한다는 한계점이 존재한다. 이러한 점들을 고려하여, Fig. 1C와 같이 포인트 클라우드로부터 추출하는 3차원 기하학적 특징뿐만 아니라 해당 장면에 관한 2차원 RGB-D 영상들에서 추출하는 시각적 특징(visual feature)들도 함께 활용해보려는 2차원-3차원 멀티-모달 특징 기반의 분할 모델들도 등장하였다[13-17].

일반적으로 2차원-3차원 멀티-모달 특징 기반의 3차원 의미적 분할 모델을 설계하기 위해서는 (1) 포인트 클라우드를 이용한 3차원 기하학적 특징 인코딩 방식, (2) 2차원 시각적 특징과 3차원 기하학적 특징 간의 효율적인 융합 전략, (3) 효과적인 벡터 수준의 특징 융합 연산을 결정해야 하는 등의 중요한 설계 이슈들이 있다.

먼저, 첫 번째 설계 이슈는 불규칙하고 희소하게 분포한 포인트들의 집합인 포인트 클라우드로부터 효과적으로 기하학적 특징을 추출할 수 있도록 3차원 기하학적 인코딩 방식을 결정하는 일이다. 1차원 단어 시퀀스인 자연어 텍스트(text), 2차원 픽셀들의 배열인 영상(image)과는 달리, 포인트 클라우드는 3차원 공간에 불규칙하게 분포한 포인트들로 이루어져 있기 때문에 특징 추출을 위해 정규 합성곱을 적용하기 어렵다. 이 문제를 극복하기 위해, 기존 연구들에서는 새롭게 설계한 포인트 합성곱(point convolution), 근거리 이웃 포인트들에 기초한 그래프 합성곱(graph convolution based) 등 다양한 포인트 특징 추출 방식들을 제안하였다. 하지만 포인트 주변의 지역적 특징 추출 시 이웃 포인트들 간의 연관성에 따른 차등적인 특징 반영이 어렵고, 네트워크가 깊어질수록 파라미터가 급증함에 따라 많은 연산 비용 및 과적합 문제가 발생하였다.

멀티-모달 특징 기반의 3차원 의미적 분할 모델의 두 번째 설계 이슈는 2차원 시각적 특징과 3차원 기하학적 특징 간의 융합 전략(fusion strategy)의 하나로서, 각각 다계층 인코딩 과정과 디코딩 과정을 통해 추출되는 두 가지 이질적인 특징들을 어떤 시점에서 융합하는 것이 가장 효과적인지를 결정하는 일이다. MVPNet[13]의 연구에서는 초기 융합(early fusion)을, SAFNet[16]의 연구에서는 후기 융합(late fusion)을 각각 제안하였으나, 두 특징의 결합 강도가 너무 강해 고유 특징을 잃어버리거나 결합 강도가 약해 충분한 융합이 이루어지지 않아 두 특징 간의 상호보완적 시너지 효과를 얻기 힘들다는 한계를 보였다.

마지막으로 세 번째 설계 이슈는 서로 이질적인 2차원 시각적 특징과 3차원 기하학적 특징 간의 벡터 수준의 구체적인 융합 연산(fusion operation) 방식을 결정하는 일이다. [13,14]의 연구에서는 3차원 기하학적 특징을 인코딩하기 전, 깊이 영상으로부터 생성한 고밀도 포인트 클라우드를 활용하여 분할 대상인 포인트 클라우드의 각 포인트를 중심으로 고밀도 이웃

포인트들 간의 위치 정보로 상대적 거리 특징을 계산 후, 이에 대응되는 시각적 특징 벡터와 단순 결합(concatenate) 및 다층 퍼셉트론(multi-perceptron)을 거치는 방식을 적용하였다. 반면에 [17]의 연구에서는 2차원 시각적 특징 벡터와 3차원 기하학적 특징 벡터의 단순 결합 후 1x1 합성곱(1x1 convolution)을 치는 방식을 통해, 두 특징 벡터를 융합하였다. 하지만, 이와 같은 기존의 특징 융합 연산 방식들은 모두 픽셀 중심의 2차원 시각적 특징과 포인트 중심의 3차원 기하학적 특징 간의 연관성을 충분히 반영하였다고 보기 어려운 한계점이 있다.

이러한 기존 방식들의 한계성을 극복하고자, 본 논문에서는 멀티-모달 특징 기반의 새로운 포인트 클라우드 의미적 분할 모델인 MMCA-Net을 제안한다. 제안 모델 MMCA-Net은 (1) 입력 포인트 클라우드로부터 맥락정보가 풍부한 3차원 기하학적 특징 인코딩을 위해 Transformer 기반의 PTv2[12]를 채용한다. 또한, (2) 이질적인 2차원 시각적 특징과 3차원 기하학적 특징 간의 효과적인 융합을 위한 중기 융합 전략을 채택하고, (3) 벡터 수준의 융합 연산을 수행하는 멀티-모달 교차 주의집중(multi-modal cross attention)을 적용한다. (4) 본 논문에서는 벤치마크 데이터 집합 ScanNetv2[18]를 이용한 정량 및 정성적 평가 실험들을 통해 제안 모델의 우수성을 입증한다.

본 논문의 2장에서는 3차원 포인트 클라우드 의미적 분할과 관련한 선행 연구들에 대해 살펴보고, 3장에서는 제안 모델의 구체적인 설계에 관해 설명한다. 이어서 4장에서는 제안 모델의 구현과 다양한 실험 결과들에 관해 소개를, 5장에서는 결론 및 향후 연구에 대해 정리한다.

2. 관련 연구

2.1 3차원 의미적 분할

3차원 의미적 분할 작업을 위한 접근 방식은 크게 투영 영상 기반의 분할 방식(projected images based segmentation), 복셀 기반의 분할 방식(voxel based segmentation), 포인트 기반의 분할 방식(point based segmentation), 멀티-모달 분할 방식(multi-modal based segmentation)으로 나눌 수 있다. 과거 연구에서는 2차원 영상 인식 기술을 활용하여 3차원 의미적 분할을 수행하는 투영 영상 기반의 분할 방식들을 제안하였다[19-22]. 투영 영상 기반의 분할 방식은 3차원 장면을 2차원 평면으로 투영한 RGB 영상들로부터 2차원 합성곱 신경망(2D CNN)을 통해 2차원 시각적 특징을 추출한 후, 이를 기반으로 각 픽셀마다 레이블을 예측한 2차원 분할 결과를 다시 3차원으로의 역투영을 통해 3차원 레이블을 결정하는 분할 방식이다. 이 방식은 오직 투영된 RGB 영상으로부터의 분할 결과에만 의존하기 때문에 풍부한 컬러와 텍스쳐 정보를 사용할 수 있으나, 포인트 클라우드의 3차원 위치 정보는 전혀 활용하지 못한다는 한계가 있다.

복셀 기반의 분할 방식은 포인트 클라우드를 그래도 사용하는 것이 아닌, 복셀화된 3차원 표현을 입력받아 3차원 합성곱 신경망(3D CNN)을 통해 3차원 의미적 분할을 수행하는 분할 방식이다[23,24]. 포인트 클라우드는 각 포인트들이 불규칙적으로 흩어져있기 때문에 정규 합성곱을 그대로 적용할 수 없다. 따라서 과거 연구에는 이를 극복하기 위해 포인트 클라우드의 복셀화 과정을 통해 3차원 격자 형태의 규칙적인 분포를 갖는 복셀(voxel)로 표현한 후, 정규 합성곱을 적용하는 방법들을 제안하였다. 하지만 복셀화 과정은 포인트 클라우드의 각 포인트의 고유한 위치 정보를 훼손시킨다는 문제점이 있다.

포인트 기반의 분할 방식은 포인트 클라우드 또는 컬러 포인트 클라우드를 입력받아 각 포인트의 위치 정보(xyz) 또는 위치 정보와 컬러 정보(xyz+rgb)로부터 추출한 3차원 기하학적 특징만을 활용하여 분할하는 방식이다[5-12]. 각 포인트로부터 3차원 기하학적 특징을 어떻게 추출하느냐는 곧 분할 성능으로 직결되기 때문에 지금까지도 효과적인 3차원 기하학적 특징 추출을 위한 다양한 연구가 진행되고 있다. 하지만 포인트 클라우드 자체의 희소성 문제로 인해 RGB 영상과 같은 밀도 있는 컬러 정보와 텍스쳐 정보를 사용할 수 없다는 단점이 있다.

2차원-3차원 멀티-모달 특징 기반의 분할 방식은 포인트 클라우드뿐만 아니라 RGB-D 영상들을 입력받아, 3차원 기하학적 특징과 2차원 시각적 특징을 함께 사용하는 분할 방식이다[13-17]. 3차원 기하학적 특징은 동일한 색상을 갖는 물체들의 서로 다른 구조 정보를 통해 구별이 가능하지만, 유사한 구조를 갖는 물체들은 구별하는 데 어려움이 있다. 반면, 2차원 시각적 특징은 유사한 구조를 갖는 물체이더라도 서로 다른 텍스처를 통해 구별 가능하지만, 유사한 텍스쳐와 빛의 밝기에 취약하다는 단점을 갖는다. 따라서 두 특징을 함께 사용하면 각각의 장단점을 상호보완함으로써 분할 성능에 도움을 줄 수 있다.

2.2 3차원 기하학적 특징 인코딩

포인트 클라우드의 포인트들로부터 3차원 기하학적 특징 인코딩 방식에 따라 다층퍼셉트론 기반의 방식(multi-layer perceptron based), 그래프 합성곱 기반의 방식(graph convolution based), 포인트 합성곱 기반의 방식(point convolution based), 그리고 트랜스포머 기반의 방식(transformer based)으로 나눌 수 있다. 먼저 다층퍼셉트론 기반의 방식은 포인트 클라우드의 각 포인트마다 다층 퍼셉트론(mlp)를 거쳐 각 포인트별 기하학적 특징을 추출하는 방식이다[5,6]. PointNet[5]와 PointNet++[6]의 연구에서는 개별 포인트마다 갖는 3차원 위치 좌표(xyz)로부터 동일한 공유 다층 퍼셉트론(shared mlp)과 최대 풀링(max pooling)을 통해 추출된 포인트 클라우드의 전역적 특징(global feature)을 추출하였다. 하

지만, 이러한 전역적 특징은 세밀한 지역적 구조를 포착하기 어려울 뿐 아니라, 이웃 포인트들의 관계 정보를 표현하기 어렵다는 한계가 존재하였다.

이에 따라 이웃한 포인트들 간의 관계 정보를 반영한 포인트 특징을 수행하기 위해 그래프 합성곱 기반의 방식이 등장하였다[7,8]. DGCNN[7]의 연구에서는 각 포인트는 노드(node)를, 포인트 간의 관계는 간선(edge)을 나타내는 그래프를 생성한 후, 이웃 포인트들과의 간선에 간선 합성곱(EdgeConv)을 적용하여 이웃 포인트들의 관계를 반영한 포인트 특징 추출이 가능하게 하였다. 한편 SPH3D-GCN[8]의 연구에서는 각 포인트를 중심으로 특정 반지름 내 존재하는 이웃 포인트들에 대해 한 번의 초기 그래프 생성 후, 이 초기 그래프를 바탕으로 계층마다 구면 합성곱(spherical convolution)과 샘플링, 최대 풀링을 반복적으로 적용하며 지역적 포인트 특징을 추출하였다. 하지만, 이러한 그래프 합성곱 기반의 방식들은 계층마다 새로운 그래프 생성 및 처리 과정으로 인해 많은 연산과 기억 공간을 요구하는 문제가 있다.

한편, 2차원 합성곱과 같이 불규칙한 포인트 클라우드에 적용할 수 있는 합성곱 커널을 설계하고 이를 통해 포인트 특징을 추출하는 포인트 합성곱 기반의 분할 방식에 관한 연구가 등장하였다[9,10]. PointConv[9]의 연구에서는 이웃 포인트들 간의 상대적 위치 정보를 토대로 합성곱 커널을 각각 예측한 후, 이를 이용해 합성곱 연산을 수행하는 방법을 제안하였다. 반면에 PAConv[10]의 연구에서는 각 포인트의 지역적 특성에 맞게 미리 정의된 가중치 행렬들을 동적으로 조합함으로써, 효율적으로 합성곱 커널을 생성하는 새로운 위치 적응형 포인트 특징 추출기를 제안하였다. 하지만 포인트 합성곱 기반의 분할 방식들은 이웃 포인트들 마다 서로 다른 합성곱 커널을 이용해야 하기 때문에 과도한 연산 비용과 기억 공간을 요구하는 문제가 존재한다.

최근 언어 및 비전 트랜스포머에 관한 연구가 활발해지면서, 트랜스포머(transformer)를 기반의 포인트 추출기를 통해 포인트 클라우드의 의미적 분할을 수행하는 연구가 주목받고 있다[11,12]. Point Transformer[11]는 각 포인트를 중심으로 이웃 포인트들 간의 벡터 주의집중(vector attention)을 적용함으로써 벡터 수준에서의 연관성을 반영한 특징 추출이 가능하게 하였다. 그러나, 이웃한 포인트 특징 벡터들의 모든 채널에 대해 계산하기 때문에 학습해야 할 파라미터 수가 많다는 단점이 있다. 이를 해결하기 위해 PTv2[12]은 각 포인트 특징 벡터의 채널을 그룹화하여 그룹 단위의 벡터 주의집중을 수행하는 그룹 벡터 주의집중(grouped vector attention)과 새로운 위치 인코딩(position encoding)을 제안하였다. PTv2[12]의 그룹 벡터 주의집중은 학습 파라미터 수와 연산 비용 문제를 개선 시켰으며, 이웃 포인트들의 차등적인 주의집중을 효과적으로 반영하면서 다른 분할 모델보다 효율적인 포인트 특징 추출이 가능하다는 장점이 있다.

2.3 멀티-모달 특징 융합 전략

2차원 시각적 특징과 3차원 기하학적 특징 간의 멀티-모달 융합은 2차원 시각적 특징과 3차원 기하학적 특징의 융합 전략에 따라 대표적으로 초기 융합(early fusion), 후기 융합(late fusion), 중기 융합(intermediate fusion) 전략으로 나눌 수 있다.

초기 융합은 3차원 포인트 클라우드가 3차원 포인트 추출 모델을 거치기 전, 멀티-뷰 영상들로부터 추출된 시각적 특징 지도를 포인트 클라우드에 먼저 융합하는 전략이다. [13,14]의 연구에서는 사전 학습된 2차원 네트워크를 통해 시각적 특징을 추출한 후, 초기 융합을 통해 분할 대상의 각 포인트를 중심으로 근접 거리의 픽셀들의 시각적 특징을 집계함으로써 보강된 3차원 포인트 특징 지도를 먼저 획득하고, 이를 토대로 3차원 의미적 분할을 수행하였다. 하지만 초기 융합 과정에서 3차원 포인트들의 고유한 위치 정보가 훼손되는 문제가 발생하였고, 사전 학습된 2차원 네트워크를 통해 시각적 특징을 추출하기 때문에 잘못된 시각적 특징이 추출되었을 경우 전체 분할 모델의 성능에 많은 영향을 미친다는 단점이 존재하였다.

이에 따라, 고유한 기하학적 특징을 보존할 수 있도록 후기 융합 전략을 통해 두 특징을 융합하는 후기 융합 전략이 등장하였다[15,16]. 후기 융합 전략은 포인트 클라우드가 3차원 포인트 추출 모델을 모두 거친 후, 분류 직전의 마지막에 시각적 특징과 융합하는 방식이다. 마지막에 두 특징이 융합되기 전까지 2차원과 3차원의 각 도메인에 독립적인 네트워크를 적용하기 때문에 초기 융합의 문제였던 3차원 포인트들의 위치 정보가 훼손되는 것을 방지하고 도메인의 특성에 맞게 고유한 특징을 추출할 수 있도록 하였다. 하지만 초기 융합 전략에 비해 2차원 시각적 특징과 3차원 기하학적 특징 간의 융합 과정이 짧아 2차원 3차원 정보의 충분한 융화가 이루어지는데 어려움을 보였다. 또한 2차원 네트워크 혹은 3차원 네트워크의 특징 추출이 올바르지 않을 때 전체 분할 성능에 미치는 부정적 효과가 크다는 문제가 있었다.

이러한 문제점을 해결하고자, 3차원 포인트 추출 모델을 거치는 과정 중에서 시각적 특징과 융합하는 중기 융합 전략이 등장하였다[17]. 중기 융합 전략은 3차원 포인트 추출 모델을 거치는 과정 중에서 시각적 특징과 결합하는 방식이다. [17]의 연구에서는 2차원 디코더 계층마다 추출되는 시각적 특징과 3차원 디코더 계층 블록마다 추출되는 기하학적 특징들을 각 계층 블록마다 융합하는 중기 융합을 채택하였다. 이를 통해 2차원 시각적 특징과 3차원 기하학적 특징 간의 융합 과정을 충분히 거치면서도 다양한 차원의 두 특징을 활용할 수 있도록 하였다. 하지만 이때 2차원 디코더 계층과 3차원 디코더 계층마다 2차원 시각적 특징을 3차원 기하학적 특징에, 3차원 기하학적 특징을 2차원 시각적 특징에 결합하는 양방향 결합을 수행하게 된다. 이에 따라 각 디코더 계층을 거칠

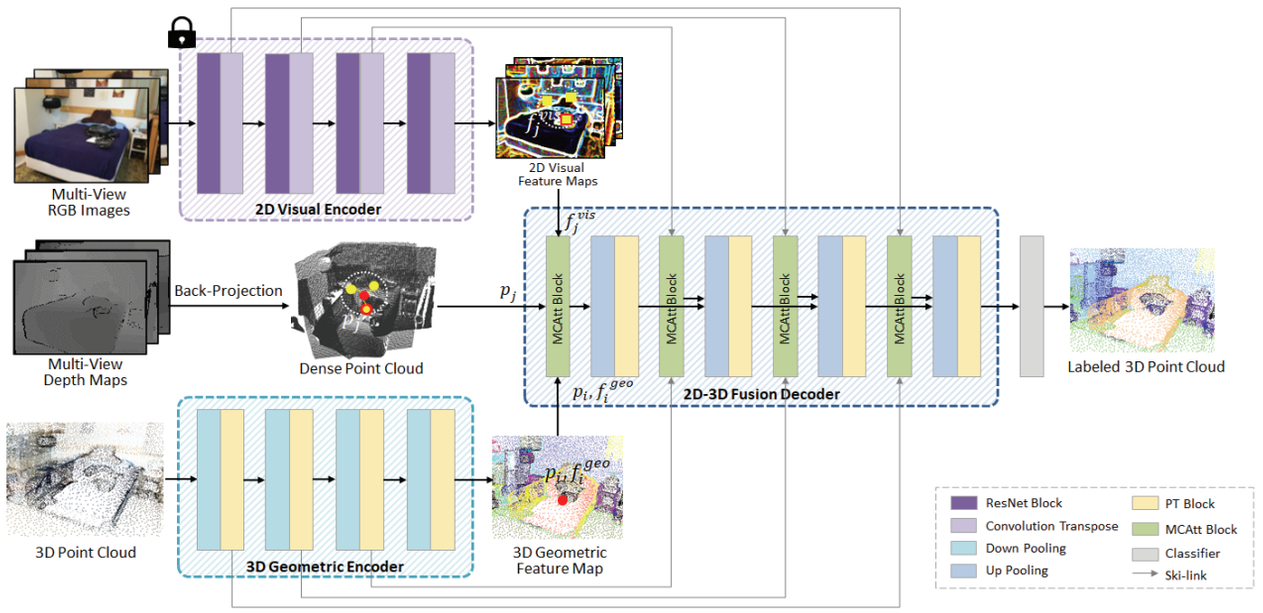

Fig. 2. Architecture of the Proposed Model

수록 2차원 시각적 특징과 3차원의 기하학적 특징의 고유한 정보가 빠르게 손실되어 간다는 문제가 있었다.

2.4 멀티-모달 특징 융합 연산

벡터 수준에서 2차원 시각적 특징과 3차원 기하학적 특징을 효과적으로 융합하기 위해 기존 연구에서는 다양한 융합 연산들을 제안하였다[13,16,17]. 먼저, BPNet[17]은 2차원 시각적 특징 벡터와 3차원 기하학적 특징 벡터를 단순 결합(concatenate) 후 1x1 합성곱(1x1 convolution)을 거치는 융합 연산을 수행하였다. 한편, MVPNet[13]은 멀티-뷰 깊이 영상을 역투영하여 고밀도 포인트 클라우드를 활용한 융합 연산을 수행하였다. 분할 대상인 저밀도 포인트 클라우드의 각 중심 포인트와 근접한 고밀도 이웃 포인트들과의 상대적 거리 특징을 생성 후 대응되는 시각적 특징을 단순 결합 및 다층 퍼셉트론(mlp)를 거치는 방식을 적용하였다. 하지만 두 융합 방식 모두 기하학적 특징과 시각적 특징 간의 직접적인 연관성에 대한 계산 없이 단순 결합하기 때문에 두 특징 정보가 적응적으로 융합하기 어렵다는 문제가 존재한다.

이러한 문제를 해결하기 위해 SAFNet[16]에서는 2차원 시각적 특징과 3차원 기하학적 특징 간의 적응적인 융합을 위해 분할 대상인 저밀도 포인트 클라우드와 멀티-뷰 깊이 영상들로부터 생성한 고밀도 포인트 클라우드 간의 유사도를 먼저 계산한 후, 기하학적 특징과 앞서 계산된 유사도를 적용한 시각적 특징을 결합하는 연산 방식을 제안하였다. 하지만, 제안된 유사도 계산은 시각적 특징과 기하학적 특징이 아닌, 고밀도 포인트 클라우드와 저밀도 포인트 클라우드 내 포인트들 간의 위치 정보를 기반하여 유사성을 계산하므로, 두 특징 간의 연관성을 직접적으로 반영하여 융합하였다고 보기 어렵다.

3. 3차원 의미적 분할 모델

3.1 모델 개요

본 논문에서는 새로운 2차원-3차원 멀티-모달 특징을 이용하는 포인트 클라우드 기반의 3차원 의미적 분할 모델 MMCA-Net을 제안한다. 제안 모델은 크게 Fig. 2와 같이, 환경 장면(indoor scene)에 대한 멀티-뷰 RGB 영상들로부터 2차원 시각적 특징 지도들을 추출하는 2차원 시각적 인코더(2D Visual Encoder), 동일 장면에 대한 3차원 포인트
클라우드로부터 3차원 기하학적 특징 지도들을 추출하는 3차원 기하학적 인코더(3D Geometric Encoder), 인코더들을 통해 추출된 2차원 시각적 특징 지도들과 3차원 기하학적 특징 지도들을 융합하는 2차원-3차원 융합 디코더(2D-3D Fusion Decoder)로 구성된다. 그리고 2차원-3차원 융합 디코더의 각 계층 블록은 다시 2차원 시각적 인코더의 계층별 2차원 시각적 특징 지도와 3차원 기하학적 인코더의 계층별 3차원 기하학적 특징 지도를 서로 융합하는 멀티-모달 교차 주의집중 블록(Multi-Modal Cross Attention Block, MCAtt Block), 그리고 이렇게 융합된 2차원-3차원 멀티-모달 특징을 디코딩하는 디코딩 블록의 쌍으로 이루어져 있다.

한편, 제안 모델의 2차원 시각적 인코더에서는 멀티-뷰 RGB 영상들을 입력받아, 각 계층 블록별로 특화된 2차원 시각적 특징 지도들을 추출한다. 이때 2차원 시각적 인코더는 RGB 영상들을 통해 사전 학습된(pre-trained) 인코더를 사용한다. 한편, 멀티-뷰 RGB 영상들로부터 계층적으로 추출되는 2차원 시각적 특징 지도들은 후속에 있을 포인트 클라우드로부터 추출되는 3차원 기하학적 특징들과의 융합을 위해, 인코더의 각 계층 블록마다 추출되는 2차원 시각적 특징 지도는 입력 RGB-D 영상

들의 해상도와 동일한 해상도로 유지된다.

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. 이때, 3차원 기하학적 인코더에서는 효율성이 높은 그룹 벡터 주의집중 (grouped vector attention)과 새로운 위치 인코딩(position encoding)을 이용하는 Transformer 기반의 PTv2[12]를 채용하였다.

한편, RGB 영상들과 함께 입력된 멀티-뷰 깊이 지도들은 역투영(back-projection) 과정을 통해 고밀도 포인트 클라우드(dense point cloud)를 생성한다. 이렇게 생성된 고밀도 포인트 클라우드는 2차원-3차원 융합 디코더의 MCAtt Block에서 분할 대상인 입력 포인트 클라우드가 다운 샘플링된 저밀도 포인트 클라우드의 각 포인트를 중심으로 고밀도 이웃 포인트들(dense neighboring points)의 2차원 시각적 특징들을 집계(aggregation)하는데 이용된다.

마지막으로, 제안 모델의 2차원-3차원 멀티-모달 융합 디코더는 2차원 시각적 특징과 3차원 기하학적 특징을 융합하는 과정을 수행한다. 각 계층의 MCAtt Block은 스킵 연결을 통해 2차원 시각적 인코더와 3차원 기하학적 인코더의 각 계층 블록마다의 2차원 시각적 특징 지도들과 저밀도 포인트 클라우드의 3차원 기하학적 특징 지도들, 그리고 고밀도 포인트 클라우드를 입력받아, 해당 계층의 저밀도 포인트 클라우드의 각 포인트에 대한 2차원-3차원 멀티-모달 특징을 구한다. 이렇게 1차로 융합된 멀티-모달 특징은 디코더 계층 블록을 거치면서 다시 한번 심층 융합되고, 2차원-3차원 융합 디코더를 모두 거친 후 가장 마지막에 분류기(classifier) 계층을 거쳐 각 포인트의 의미적 레이블(semantic label)을 결정한다. 후속 절들에서는 제안 모델의 2차원 시각적 인코딩(2D visual encoding)과 3차원 기하학적 인코딩(3D geometric encoding), 그리고 멀티-모달 융합 전략(multi-modal fusion strategy)과 멀티-모달 교차 주의집중 블록(MCAtt Block)에 대해 상세히 설명한다.

3.2 2차원 시각적 인코딩

제안 모델의 2차원 시각적 인코더는 Fig. 3과 같이, 입력된 멀티-뷰 RGB 영상들로부터 각 계층 블록마다 인코딩된 2차원 시각적 특징 지도들을 추출한다. 2차원 시각적 인코더의 각각의 계층 블록은 잔차 블록 (Residual Block, ResNet Block)과 합성곱 전치(convolution transpose) 계층으로 구성되어 있으며, 각 계층 블록마다의 ResNet Block은 ImageNet [25] 데이터 집합으로 사전 학습된 ResNet34[26]의 ResNet Block을 백본(backbone)으로 채용한다. 이때, ResNet Block을 거치면서 줄어든 2차원 시각적 특징 지도들의 해상도를 입력 멀티-뷰 RGB 영상들과의 해상도와 동일하게 유지하기 위해 합성곱 전치 계층을 추가로 삽입한다.

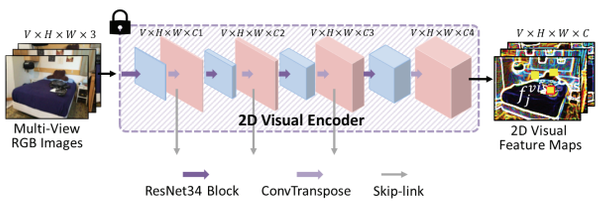

Fig. 3. 2D Visual Encoder

Equation(1)의 F^{2D}은 2차원 시각적 인코더의 l번째 계층 블록에서 추출되는 2차원 시각적 특징 지도들을 나타내며, $H\times W$의 해상도를 갖는 V개 멀티-뷰 영상들을 이루는 각 픽셀마다의 시각적 특징 벡터 f_j^{2D}들의 집합으로 구성된다.

$$F^{2D}=\left\{f_j^{2D}\ \middle|\ j\in V\times H\times W\right\} \tag{1}$$

이러한 2차원 시각적 인코더는 2차원 시각적 디코더와 함께 ScanNetv2[18] 데이터 집합을 이용하여 영상의 각 픽셀마다 분류 레이블을 할당하는 2차원 의미적 분할 작업을 통해 사전 학습된다.

3.3 3차원 기하학적 인코딩

제안 모델의 3차원 기하학적 인코더는 Fig. 4와 같이, 분할 대상인 입력 포인트 클라우드로부터 각 계층 블록별로 다운 샘플링된 저밀도 포인트 클라우드와 3차원 기하학적 특징 지도를 추출한다. 이때 각 계층 블록은 다운 풀링(down pooling) 계층과 포인트 트랜스포머 블록(Point Transformer Block, PT Block)으로 구성되어 있으며, 이때 PT Block은 3차원 기하학적 인코더의 효율적인 3차원 기하학적 특징 지도를 추출하기 위해 트랜스포머 기반의 PTv2[12]의 인코더를 백본으로 채용한다.

다운 풀링 계층은 Fig. 4의 첫 번째 계층과 같이, 입력된 전체 포인트 클라우드에 대해 3차원 격자(grid)를 통해 포인트들의 그룹을 생성한 후, 3차원 격자의 각 윈도우(window) 마다 대표적인 포인트를 샘플링 하는 그리드 기반의 풀링을 수행한다. 이후 PT Block은 PTv2의 그룹 벡터 주의집중과 새로운 위치 인코딩 기법을 통해 다운 샘플링된 모든 포인트 클라우드에 대해 각 포인트와 이웃한 포인트들 간의 자기-주의집중(self-attention)을 수행하여 각 포인트의 3차원 기하학적 특징을 업데이트한다. Equation (2)는 3차원 기하학적 인코더의 l

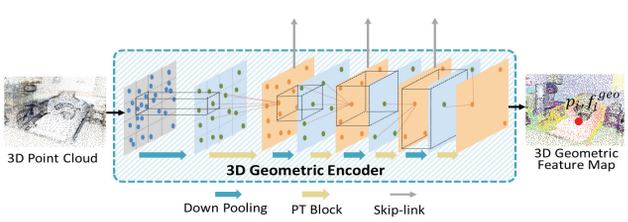

Fig. 4. 3D Geometric Encoder

번째 계층 블록에서 추출되는 저밀도 포인트 클라우드의 3차원 기하학적 특징 지도 F^{3D}를 나타내며, 다운 샘플링된 M개의 포인트에 대한 위치 정보 p_i와 기하학적 특징 벡터 f_i^{3D}들의 집합으로 구성된다.

$$F^{3D} = \left\{ (p_i, f_i^{3D}) \mid i \in M \right\} \tag{2}$$

3.4 멀티-모달 융합 전략

2차원 시각적 특징 지도와 3차원 기하학적 특징 지도의 융합 전략을 결정하는 것은 3차원 의미적 분할 모델 전체의 구조를 결정하는 일로 분할 성능에 중요한 영향을 미친다. 융합 전략은 Fig. 5와 같이 크게 초기 융합(early fusion), 후기 융합(late fusion). 중기 융합(intermediate fusion)으로 나눌 수 있다.

초기 융합 전략은 [13,14]의 연구에서 사용된 융합 전략으로, Fig. 5A와 같이 2차원 인코딩-디코딩 과정을 거쳐 얻은 2차원 시각적 특징들을 포인트 클라우드에 대한 3차원 인코딩 과정을 거치기 전에, 각 포인트에 미리 융합하는 전략이다. 반면에 후기 융합 전략은 [16]의 연구에서 사용된 융합 전략으로, Fig. 5B와 같이 각기 2차원 인코더-디코더와 3차원 인코더-디코더를 독립적으로 거쳐 얻는 2차원 시각적 특징과 3차원 기하학적 특징을 마지막에 융합하는 전략이다. 초기 및 후기 융합 전략 모두 계층별로 상이한 2차원 시각적 특징들과 3차원 기하학적 특징들을 효과적으로 융합하기 어렵다는 문제가 있다. 이에 반해 중기 융합 전략은 Fig. 5C와 같이, 2차원 인코더 계층마다의 추출되는 시각적 특징과 3차원 인코더 계층마다의 추출되는 기하학적 특징을 각 계층별로 융합한 후, 이를 스킵 연결(skip-link)을 통해 3차원 디코더에 다시 계층별로 융합하는 전략이다. 중기 융합 전략은 2차원과 3차원 인코더 계층별 스킵 연결을 통해 다양한 차원의 고유한 두 특징 간의 융합이 가능하다. 대신 두 인코더 계층마다 추출되는 각 특징 지도의 크기가 달라 픽셀-포인트 매핑이 어려워지는 문제가 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 Fig. 5C와 같은 중기 융합 전략을 채택하였다. 또한 앞서 언급한 것과 같이, 2차원 인코더의 계층마다 추출되는 시각적 특징 지도의 크기를 동일하게 유지하고 시각적 특징 지도에 대응되는 깊이 영상의 고밀도 포인트 클라우드를 함께 활용함으로써 픽셀-포인트 매핑 문제를 해결한다.

중기 융합 전략을 채택한 제안 모델의 2차원-3차원 멀티-모달 디코더는 Fig. 2와 같이, 각 계층 블록별로 두 특징 간의 연관성을 고려하여 융합을 수행하는 멀티-모달 교차 주의집중 블록(MCAtt Block), 그리고 3차원 디코딩을 수행하는 업풀링(up pooling) 계층 및 포인트 트랜스포머 블록(PT Block)으로 구성된다. 계층별 멀티-모달 주의집중 블록은 스킵 연결을 통해 3차원 기하학적 인코더의 각 계층 블록으로부터 저밀도 포인트 클라우드의 3차원 기하학적 특징 지도 F^{3D}를, 2차원 시각적 인코더의 각 계층 블록으로부터 2차원 시각적 특징 지도들 F^{2D}과 이에 대응되는 깊이 영상으로부터 생성한 고밀도 포인트 클라우드를 입력받는다. 3차원 기하학적 특징 지도 F^{3D}를 이루는 각 포인트 p_i를 중심으로 고밀도 포인트 클라우드 내의 이웃한 k개의 고밀도 이웃 포인트 p_j들을 찾은 후, 중심 포인트 p_i의 3차원 기하학적 특징 벡터 f_i^{3D}를 중심으로 고밀도 이웃 포인트 p_j들에 대응되는 2차원 시각적 특징 벡터 f_j^{2D}들의 연관성을 반영하여 집계한다. 이후 이것을 다시 3차원 기하학적 특징과 결합함으로써, 해당 계층의 저밀도 포인트 클라우드의 각 포인트 p_i에 대한 새로운 2차원-3차원 멀티-모달 특징 $f_i^{2D\&3D}$을 계산한다.

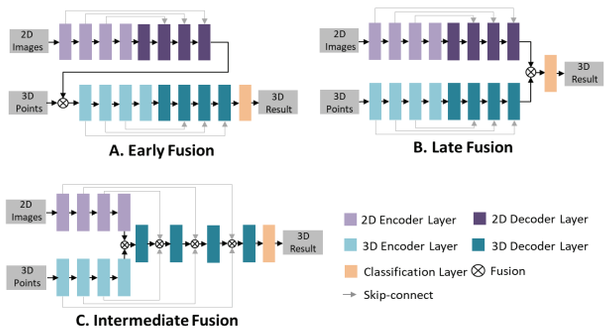

Fig. 5. Three Different Fusion Strategies

이렇게 계층 블록별로 융합된 2차원-3차원 멀티-모달 특징 지도는 이전 디코더 계층 블록을 거친 특징 지도와 함께 업풀링 계층을 통해 3차원 그리드 기반의 보간(interpolation)을 수행한 후, 보간된 2차원-3차원 멀티-모달 특징 지도는 PT Block을 통해 다시 한번 심층 융합을 진행한다. 2차원-3차원 융합 디코더를 모두 거쳐 입력된 3차원 포인트 클라우드의 크기와 동일해진 멀티-모달 특징 지도는 분류기를 통해 각 포인트마다 분류 레이블이 할당됨으로써 포인트 클라우드의 최종적인 의미적 분할이 완료된다.

3.5 멀티-모달 교차 주의집중 블록

제안 모델의 멀티-모달 교차 주의집중 블록(MCAtt Block)은 3차원 기하학적 특징 벡터와 2차원 시각적 특징 벡터 간의 주의집중 정도를 계산 후 이를 바탕으로 벡터 수준에서의 융합을 수행한다. Fig. 6A와 같이, 3차원 기하학적 인코더의 각 계층 블록에서 추출되는 저밀도 포인트 클라우드의 각 포인트 p_i의 3차원 기하학적 특징 벡터 f_i^{3D}을 중심으로, 고밀도 이웃 포인트 p_j들의 2차원 시각적 특징 f_j^{2D}들과의 연관성을 충분히 반영하여 이들을 집계한 후 3차원 기하학적 특징 f_i^{3D}와 다

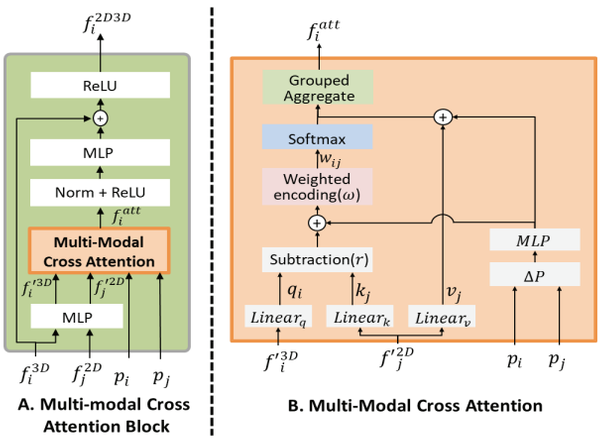

Fig. 6. Multi-Modal Cross Attention Block

시 결합함으로써, 포인트 p_i의 2차원-3차원 멀티-모달 특징 f_i^{2D3D}을 생성한다. 이때, 3차원 기하학적 특징 벡터 f_i^{3D}를 중심으로, 2차원 시각적 특징 벡터 f_j^{2D}들을 집계하는 방식은 Fig. 6B와 같은 멀티-모달 교차 주의집중(Multi-modal Cross Attention)을 이용한다.

먼저 3차원 기하학적 벡터 f_i^{3D}와 2차원 시각적 벡터 f_j^{2D}는 다층 퍼셉트론(mlp)를 거쳐 $f_i{}'^{3D}$와 $f_j{}'^{2D}$를 구한 후 Equation (3)과 같이, 선형 함수(linear)를 통해 $f_i{}'^{3D}$으로부터 주의집중의 중심이 되는 쿼리 q_i를, $f_j{}'^{2D}$으로부터 쿼리와의 연관성을 계산하기 위한 키 k_j와 이에 대해 대응되는 밸류 v_j을 각각 구한다.

$$\begin{aligned} q_i &= Linear_q(f_i{}'^{3D}), \\ k_j &= Linear_k(f_i{}'^{2D}), \\ v_j &= Linear_v(f_i{}'^{2D}) \end{aligned} \tag{3}$$

이후 Equation (4)와 같이, 저밀도 포인트 p_i와 고밀도 이웃 포인트 p_j들로부터 2개의 위치 인코딩(position encoding) δ을 구하고 하나의 위치 인코딩 δ는 쿼리 q_i와 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 c개의 채널을 갖는 관계 특징 $\gamma(q_i,k_j)$의 채널을 g개의 그룹으로 나눈 후, 각 그룹마다의 주의집중 가중치 값을 갖는 벡터이다.

$$w_{ij} = \omega(r(q_i,k_j)+\delta), \quad \delta = MLP(p_j - p_i) \tag{4}$$

다음 Equation (5)과 같이, 그룹화된 가중치 벡터 w_{ij}는 활성함수인 $Softmax$를 통해 0-1 사이의 확률값으로 변환하고, 이를 앞서 획득한 위치 인코딩 δ을 반영한 밸류 v_j에 적용 후 집계하여 최종적으로 주의 집중된 2차원 시각적 특징 f_i^{att}을 구한다. 이때, $M(p_i)$는 포인트 p_i와 이웃한 k개의 고밀도 포인트 p_j들, c는 채널 수, g는 채널을 나눈 그룹 수를 의미한다. 그룹화된 가중치 벡터 w_{ij}를 그룹별로 밸류 v_j에 적용하기 위해 마찬가지로 밸류 v_j의 c개의 채널을 g개의 그룹으로 나눈 후, 그룹 n에 속한 밸류 v_j의 채널 m에 대해, 동일한 그룹 n에 속한 가중치 벡터 w_{ij}의 주의집중 가중치 값을 적용한다.

$$f_i^{att} = \sum_{p_j}^{M(p_i)} \sum_{n=1}^{g} \sum_{m=1}^{c/g} Softmax(w_{ij})_n (v_j + d)^{nc/g+m} \tag{5}$$

4. 구현 및 실험

4.1 모델 구현과 학습

본 논문의 제안 모델 MMCA-Net은 Python 딥러닝 라이브러리인 Pytorch를 이용해 구현하였으며, 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 장면의 RGB-D 데이터 집합인 ScanNetv2[18] 벤치마크 데이터를 사용하였다. ScanNetv2는 침실, 부엌, 서재 등 총 1,500개 이상의 장면 데이터와 250만 뷰의 영상들을 포함한다. 각 장면마다 2차원 RGB-D 비디오 프레임과 프레임별 카메라 포즈, 3차원 포인트 클라우드 등이 포함되어 있다. 이 중 1,201개의 장면 데이터는 모델의 학습에, 312개의 장면 데이터는 모델의 평가에 사용하였다.

제안 모델에서 2차원 시각적 특징을 추출하는 2차원 시각적 인코더는 사전 학습된 2차원 의미적 분할 모델의 인코더 부분을 사용한다. 이때 2차원 의미적 분할 모델은 학습을 위해 손실함수로는 크로스엔트로피(cross entropy)를, 최적화 알고리즘은 확률적 경사 하강법(stochastic gradient descent)를 사용하여 사전 학습하였다. 사전 학습된 2차원 시각적 인코더를 채용한 제안 모델에 대해서는 학습을 위한 손실함수로 크로스엔트로피와 최적화 알고리즘으로 Adam을 사용하여 전체 모델에 대해 종단 간 학습(end-to-end)을 진행하였다.

4.2 실험 및 평가

제안 모델 MMCA-Net의 성능을 분석하기 위해 (1) 서로 다른 3차원 기하학적 추출기에 따른 성능 비교 실험, (2) 융합 전략에 따른 성능 비교 실험, (3) 특징 벡터 융합 방식에 따른 성능 비교 실험, (4) 기존 모델들과의 성능 비교 실험으로 총 4가지의 정량적 평가 실험들을, 마지막으로 (5) 제안 모델의 분할 결과를 정성적으로 비교 분석하기 위한 정성적 평가 실험을 진행하였다. 모든 정량적 평가 실험들에서는 성능 평가

Table 1. Performance Comparison with Different 3D Geometric Encoders

3D Geometric Encoder	oAcc(%)	mIoU(%)
PointNet++[6]	81.82	54.23
PAConv[10]	86.76	65.56
PointTransformer[11]	86.56	63.89
PTv2[12]	**89.70**	**71.87**

척도로 전체 정확도(overall accuracy, oAcc)와 평균 교차 영역 비율(mean intersection over union, mIoU)를 이용한다. oAcc는 평가를 진행한 모든 포인트 클라우드의 전체 포인트들 중에서 얼마나 많은 포인트들이 올바른 레이블로 분류되었는지를 나타내며, mIoU는 각 물체마다 정답 레이블 포인트들과 예측된 레이블 포인트들의 일치 정도를 계산하고 이를 전체 물체에 대해 평균화한 수치를 나타낸다.

첫 번째 실험은 제안 모델 MMCA-Net에서 채용한 그룹 벡터 주의집중 기반의 PTv2[12]의 포인트 특징 추출기의 우수성을 입증하기 위한 실험이다. 이 실험에서는 제안 모델을 구성하는 3차원 기하학적 인코더를 제외한 나머지 부분은 제안 모델과 동일하게 구성하되, 3차원 기하학적 인코더 부분은 제안 모델과 같은 PTv2[12], 그리고 기존의 PointNet++[6], PAConv[10], Point Transformer[11]를 각각 사용했을 때의 분할 성능을 서로 비교해본다. PointNet++은 다층퍼셉트론 기반의 방식, PAConv은 포인트 합성곱 기반의 방식을, 그리고 PointTransformer와 PTv2는 트랜스포머 기반의 방식을 통해 기하학적 특징을 추출한다.

Table 1의 실험 결과를 살펴보면, 본 연구에서 제안한 PTv2를 채용한 분할 모델이 다른 모델을 채용한 분할 모델보다 분할 성능이 가장 높은 것을 확인할 수 있다. PTv2를 채용한 분할 모델이 PointNet++과 PAConv보다 oAcc 면에서 각각 약 9.63%, 3.39% 더 높은 성능을 보였으며, mIoU 면에서는 각각 약 32.53%, 9.62%의 성능 향상을 보였다. 또한, 트랜스포머 기반의 분할 모델들과 비교해보자면, 그룹 벡터 주의집중을 사용한 PTv2가 벡터 주의집중을 사용한 PointTransformer보다 oAcc, mIoU 면에서 약 3.63%, 12.49% 향상된 성능을 보였다. 이는 PointTransformer의 벡터 주의집중 방식보다 PTv2의 그룹 주의집중 방식이 이웃 포인트 간의 차등적인 주의집중을 효과적으로 수행함으로써, 더 정확한 3차원 기하학적 특징 추출에 도움이 된 것으로 추측한다. 이와 같은 실험 결과를 종합하였을 때, 제안 모델에 채용한 그룹 벡터 주의집중 기반의 PTv2의 인코더가 전체 분할 성능 향상에 긍정적인 영향을 미치는 것을 확인할 수 있다.

두 번째 실험은 제안 모델에 적용된 중기 융합 전략의 타당성을 입증하기 위한 실험이다. 위 실험에서는 제안 모델에서 사용된 백본들과 멀티-모달 교차 주의집중 방식은 동일하게 유지하되, 융합 전략은 초기 융합(early fusion), 후기 융합(late fusion), 중기 융합(intermediate fusion) 전략을 각각 적용한 성능을 비교한다. (a) 초기 융합 전략은 Fig. 5A와 같이 2차원 인코더-디코더를 거쳐 추출된 시각적 특징과 3차원 인코더-디코더를 거치지 않은 포인트 클라우드의 위치 정보(xyz) 그대로의 기하학적 특징을 융합하는 방식이다. (b) 후기 융합 전략은 Fig. 5B와 같이 2차원 인코더-디코더를 거친 시각적 특징과 3차원 인코더-디코더를 거친 기하학적 특징을 결합하는 방식을, 마지막 제안 모델의 (c) 중기 융합 전략은 Fig. 5C와 같이, 2차원 인코더 계층별 시각적 특징과 3차원 인코더 계층별 기하학적 특징을 융합하는 전략이다. 이때 초기 융합 전략은 [13,14]의 선행 연구에서 적용된 초기 융합 전략을, 후기 융합 전략은 [16]의 선행 연구에서 적용된 후기 융합 전략을 각각 나타낸다.

Table 2. Performance Comparison with Different Fusion Strategies

Strategy	Visual Feature	Geometry Feature	oAcc (%)	mIoU (%)
(a) Early Fusion [13,14]	2D en-decoder	None (xyz)	88.48	68.47
(b) Late Fusion[16]	2D en-decoder	3D en-decoder	88.89	69.54
(c) Intermediate Fusion	2D encoder layer-wise	3D encoder layer-wise	**89.70**	**71.87**

Table 2의 실험 결과를 살펴보면, 제안 모델 MMCA-Net과 같이, (c) 인코더 계층별 중기 융합 전략이 나머지 다른 융합 전략보다 성능 척도 oAcc, mIoU 면에서 가장 높은 성능을 보였다. 구체적으로 살펴보면, (b) 후기 융합 전략은 (a) 초기 융합 전략보다 oAcc, mIoU 면에서 0.46%, 1.56% 향상된 성능을 보인다. 이는 초기 융합 전략은 빠른 융합으로 인해 포인트 클라우드의 가장 중요한 위치 정보가 빠르게 손실되지만, 후기 융합 전략은 시각적 특징뿐만 아니라 고유한 기하학적 특징이 독립적으로 추출되어 포인트 클라우드의 고유한 구조적 정보가 충분히 반영 융합되었기 때문으로 추측된다.

한편, (c) 중기 융합 전략은 (a) 초기 융합과 (b) 후기 융합 전략보다 oAcc면에서 각각 약 1.37%, 0.91%의 성능 향상률을, mIoU면에서 각각 약 4.96%, 3.35%의 성능 향상률을 보였다. 중기 융합 전략은 2차원과 3차원 인코더 계층별로 독립적으로 특징을 추출함으로써 초기 융합에 비해 각 특징의 고유성을 보장하면서, 각 인코더 계층별 융합을 통해 후기 융합 전략보다 충분한 융합 과정이 수행되어 더 높은 성능을 보인 것으로 판단된다. 위 실험을 통해, 제안 모델에서 채택한 중기 융합 전략의 유효성을 확인할 수 있다.

세 번째 실험은 2차원 시각적 특징과 3차원 기하학적 특징 간의 특징 벡터의 융합을 위한 제안 모델의 멀티-모달 교차 주의집중 방식의 우수성을 입증하기 위한 실험이다. 이 실험에서는 MCAtt Block을 제외한 나머지는 제안 모델과 동일하

Table 3. Performance Comparison with 2D-3D Feature Fusion Operations

Fusion Operation	oAcc(%)	mIoU(%)
(a) Concatenate	89.53	71.04
(b) Concatenate + Linear transformation[13]	89.34	71.06
(c) Vector attention	89.52	71.10
(d) Grouped vector attention	**89.70**	**71.87**

게 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 주의집중 방식을 적용한 경우(vector attention), 그리고 제안 모델의 멀티-모달 교차 주의집중 방식과 같이 (d) 그룹 벡터 주의집중 방식을 이용하는 경우(grouped vector attention)에 대해 각각의 분할 성능을 비교한다. 특히 (b) 단순 결합 후 선형 변환 방식은 [13]의 선행 연구에서 적용한 융합 연산 방식과 동일하다.

Table 3의 실험 결과를 살펴보면, 제안 모델과 같이 (d) 그룹 벡터 주의집중 방식을 이용한 경우가 다른 방식을 이용한 경우들에 비해 가장 높은 성능을 보였다. (a) 단순 결합 방식은 (b) 단순 결합 후 선형 변환 방식보다 oAcc 면에서 더 향상된 성능을 보였으나 mIoU 측면에서 낮은 성능을 보였다. 반면, (c) 벡터 주의집중 방식은 (a) 단순 결합방식보다 oAcc 면에서 더 낮은 성능을 보였으나 mIoU 면에서 더 높은 성능을 확인할 수 있다.

한편, (d) 그룹 벡터 주의집중 방식은 (a) 단순 결합방식과 (c) 벡터 주의집중 방식보다 oAcc 면에서 약 0.18%, 0.20%로 미미하지만 향상된 분할 성능을 보였으며, mIoU 면에서 약 1.16%, 1.08%의 성능 향상률을 보였다. 이와 같은 실험 결과를 통해, 제안 모델과 같은 (d) 그룹 벡터 주의집중 방식은 2차원 시각적 특징과 3차원 기하학적 특징 간의 연관성이 반영되어 융합되기 때문에 (a) 단순 결합방식보다 성능 향상에 도움이 된다는 것을 알 수 있다. 또한 두 특징 벡터 간의 연관성을 바탕으로 주의집중 정도를 계산할 때 (c) 벡터 주의집중 방식보다 (d) 그룹 벡터 주의집중 방식이 성능 향상에 더 효과적임을 알 수 있다.

마지막 네 번째 실험은 기존의 대표적인 3차원 포인트 클라우드 의미적 분할 모델들과의 비교를 통해, 제안 모델 MMCA-Net의 우수성을 입증하기 위한 실험이다. 이 실험에서는 기존 모델 중에서 RGB-D 영상들로부터의 2차원 시각적 특징만 이용하는 모델들(RGB-D Video Frames)[27], 포인트들의 위치 정보와 색상 정보가 포함된 컬러 포인트 클라우드의 혼합 특징을 이용하는 모델들(Colored Point Cloud [xyz + RGB]) [6,8,10-12], 그리고 제안 모델과 같이 멀티-뷰 RGB-D 영상들로부터 추출한 2차원 시각적 특징과 포인트 클라우드에서 추출한 3차원 기하학적 특징을 같이 이용하는 2차원-3차원 멀티-모달 특징 기반의 모델들(Multi-view RGB-D Images + Point Cloud[xyz]) [13, 15-17]을 모두 비교한다. 이때 PAConv[10]*, PointTransformer[11]*, PTv2[12]*, BPNet[17]* 모델은 앞서 언급한 제안 모델의 구현 환경과 동일한 환경에서 학습 및 평가한 성능을 나타낸다.

Table 4의 실험 결과를 살펴보면, 제안 모델이 전체 비교 모델들 중 분할 물체 클래스에 대해서 문(door), 테이블(table), 바닥(floor) 등 20개의 전체 물체 클래스 중 절반이 성능 척도 IoU 면에서 가장 높은 성능을 보였으며, 모든 클래스에 대한 IoU를 평균화한 mIou 면에서 가장 높은 성능을 보였다. 자세히 살펴보면, 컬러 포인트 클라우드만을 이용한 모델들 중에서는 그룹 벡터 주의집중 기반의 PTv2가 다층퍼셉트론 기반의 PointNet++와 그래프 합성곱 기반의 SPH3D-GCN, 포인트 합성곱 기반의 PAConv보다 각각 mIoU 측면에서 약 94.1%, 7.87%, 15.44% 더 높은 성능을 보였다. 또한 벡터 주의집중 기반의 Point Transformer보다 11.7% 더 향상된 성능을 보였다. 한편, 멀티-뷰 RGB-D 영상들과 포인트 클라우드의 기하학적

Table 4. Performance Comparison with Other 3D Semantic Segmentation Models

Model	Input	mIoU	bath	bed	bkshf	cab	chair	cntr	curt	desk	door	floor	other	pic	fridge	shower	sink	sofa	table	toilet	wall	window
OnlineSegFusion[27]	RGB-D Video Frames	51.5	60.7	64.4	57.9	43.4	63.0	35.3	62.8	44.0	41.0	76.2	30.7	16.7	52.0	40.3	51.6	56.5	44.7	67.8	70.1	51.4
PointNet++[6]	Colored Point Cloud [xyz + RGB]	33.9	58.4	47.8	45.8	25.6	36.0	25.0	24.7	27.8	26.1	67.7	18.3	11.7	21.2	14.5	36.4	34.6	23.2	54.8	52.3	25.2
SPH3D-GCN[8]		61.0	85.8	77.2	48.9	53.2	79.2	40.4	64.3	57.0	50.7	93.7	41.4	4.6	41.0	70.2	60.2	70.5	54.9	85.9	77.3	53.4
PAConv[10]*		57.04	68.7	64.4	63.8	51.6	80.5	51.0	46.4	54.1	45.5	92.7	28.5	12.1	37.9	48.7	52.8	67.1	66.8	82.3	75.2	49.0
Point Transformer[11]*		58.94	67.6	68.9	66.9	55.4	84.4	48.5	53.8	56.0	45.0	89.4	37.8	17.6	42.9	42.0	52.6	77.3	70.0	76.6	76.4	49.6
PTv2[12]*		65.81	84.5	77.9	54.6	58.6	88.0	66.1	51.7	58.2	60.3	93.8	49.2	24.9	46.3	58.8	66.0	79.4	69.3	89.1	80.6	58.8
3DMV[15]	Multi-View RGB-D Images + Point Cloud [xyz]	48.4	48.4	53.8	64.3	42.4	60.6	31.0	57.4	43.3	37.8	79.6	30.1	21.4	53.7	20.8	47.2	50.7	41.3	69.3	60.2	53.9
MVPNet[13]		64.1	83.1	71.5	67.1	59.0	78.1	39.4	67.9	64.2	55.3	93.7	46.2	25.6	64.9	40.6	62.6	69.1	66.6	87.7	79.2	60.8
SAFNet[16]		65.4	75.2	73.4	66.4	58.3	81.5	39.9	75.4	63.9	53.5	94.2	47.0	30.9	66.5	53.9	65.0	70.8	63.5	85.7	79.3	64.2
BPNet[17]*		70.17	85.7	80.7	80.0	67.5	90.1	56.1	62.1	69.8	60.3	94.6	57.6	24.5	48.0	65.7	68.1	80.5	77.1	90.7	83.3	60.8
MMCA-Net(ours)		71.87	88.7	79.2	78.9	65.6	88.2	57.8	71.6	63.1	67.6	95.4	56.9	33.1	66.7	67.5	64.4	75.9	73.3	94.2	85.0	64.2

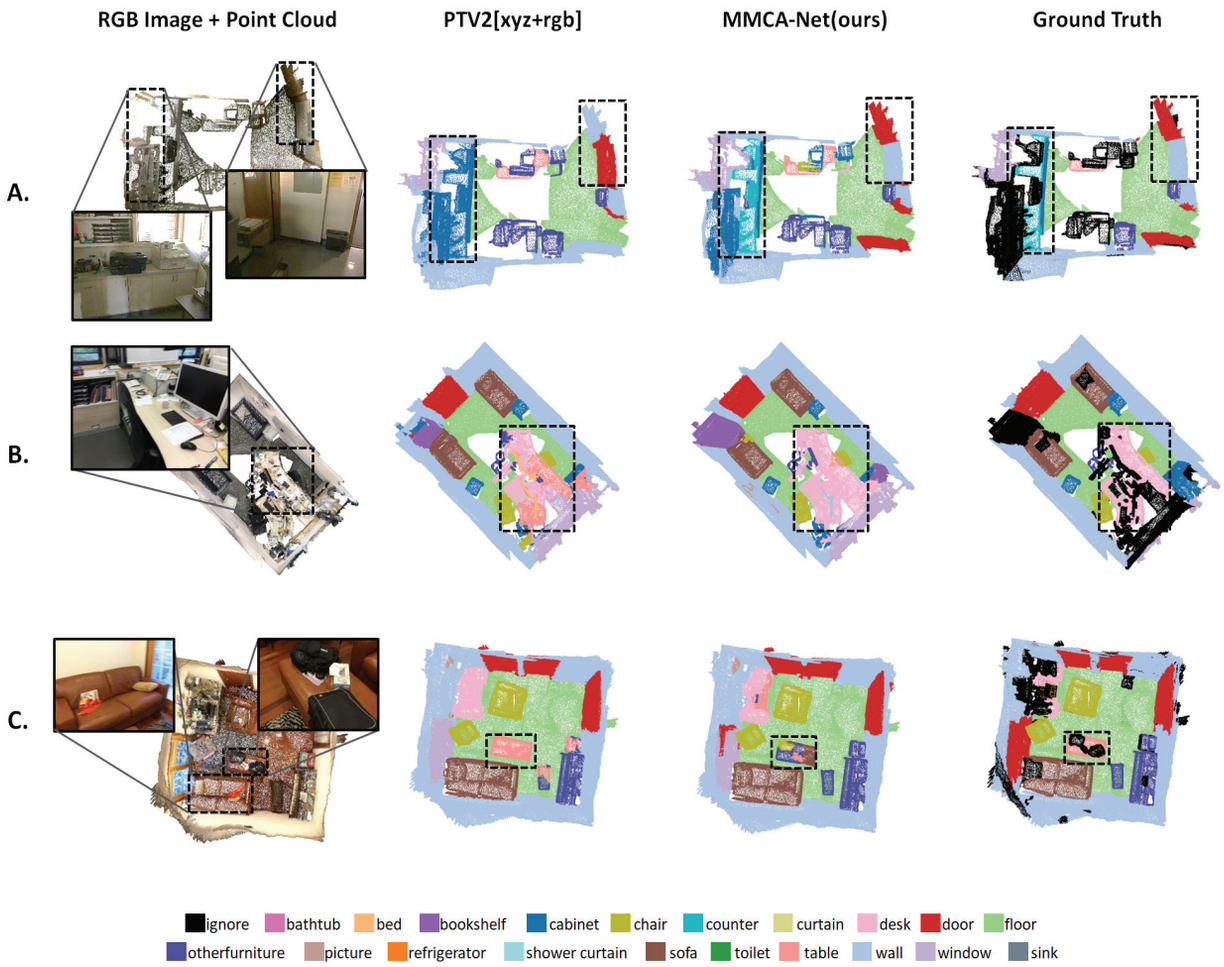

Fig. 7. Qualitative Evaluation of 3D Semantic Segmentation Results

특징을 함께 이용하는 모델들 중에서는 제안 모델이 초기 융합 전략을 적용한 MVPNet보다 mIoU 면에서 약 12.12%를, 후기 융합 전략을 적용한 3DMV, SAFNet보다 각각 약 48.49%, 9.92% 더 높은 성능을 보였다. 제안 모델과는 다른 중기 융합 전략을 채택한 BPNet보다 의자(chair), 소파(sofa), 책상(table) 등 일부 클래스에 대해서 낮은 성능을 보였으나, 전체 클래스에 대한 mIoU 면에서 약 2.42% 향상된 성능을 보였다.

사용된 입력 정보에 따라 분할 모델들의 성능을 비교해보면, RGB-D 비디오 프레임들을 입력으로 사용하는 OnlineSegFusion이 컬러 포인트 클라우드의 3차원 기하학적 특징만을 활용한 PTv2 보다 책꽂이(bookshelf), 커튼(curtain)과 냉장고(refridgerator)에 대해 IoU 면에서 더 높은 성능을 보였으나, 나머지 물체 클래스에 대해서는 PTv2가 더 높은 성능을 보였고 mIoU 면에서도 약 27.8%의 성능 향상을 보였다. 또한, 포인트 특징만을 추출하는 그룹 벡터 기반의 PTv2가 멀티-뷰 RGB-D 영상들 및 포인트 클라우드로부터 2차원-3차원 멀티-모달 특징을 활용함과 동시에 PTv2를 백본 일부로 채용한 제안 모델 MMCA-Net보다 카운터(counter), 싱크대(sink), 소파(sofa)에 대해 더 높은 성능을 보였다. 하지만, 나머지 물체 클래스에 대해서는 제안 모델 MMCA-Net이 더 높은 성능을 보였고 mIoU 면에서도 약 11.6%의 성능 향상률을 보였다. 이와 같은 실험을 통해 본 논문에서 제안한 MMCA-Net의 우수성을 확인할 수 있었다. 또한 PTv2는 컬러 포인트 클라우드를 사용한 다른 분할 모델들에 비해 가장 뛰어난 성능을 보였고, 이를 일부 백본으로 채용한 제안 모델의 분할 성능에 긍정적인 영향을 미친 것으로 판단된다.

마지막 실험은 몇 가지 분할 결과의 사례들을 통해 제안 모델 MMCA-Net의 분할 성능을 정성적으로 분석하는 실험이다. Fig. 7은 정성적 평가를 위해 나타낸 그림이며, 위 실험에서는 3차원 컬러 포인트 클라우드만을 입력으로 받아 의미적 분할을 수행하는 PTv2[xyz+rgb] 모델과 RGB-D 영상들과 포인트 클라우드를 함께 입력으로 받아 의미적 분할을 수행하는 제안 모델 MMCA-Net과 분할 결과를 비교한다. Fig. 7A, Fig. 7B는 제안 모델 MMCA-Net의 성공적인 분할 사례를, Fig. 7C는 제안 모델의 실패 사례를 보여준다.

먼저 Fig. 7A에서는 RGB 영상들로부터 물체마다 갖는 서

로 다른 색상 및 텍스처 정보의 도움받아 올바른 분할을 수행한 사례를 보여준다. PTv2 모델의 분할 결과를 살펴보면, 카운터(counter)를 이루는 포인트들이 사물함(cabinet)으로, 문(door)을 이루는 포인트들은 벽(wall), 벽을 이루는 포인트들은 문으로 잘못된 분류가 이루어진 것을 확인할 수 있다. 이는 카운터를 이루는 포인트들을 직육면체 형태의 사물함을 이루는 하나의 면으로 인식되어 잘못된 기하학적 특징이 추출된 것으로 추측한다. 또한 문과 벽은 평면적인 구조로 이루어져 있다는 유사한 구조적 특징을 가지고 있어 분할이 잘못 이루어진 것으로 판단한다. 반면 RGB 영상들을 함께 사용하는 제안 모델은 카운터, 문, 그리고 벽을 이루는 포인트들 모두 올바르게 분류된 것을 볼 수 있다. 이는 다른 물체를 이루는 포인트들이 구조적으로 서로 유사한 특징을 가질 경우, RGB 영상 내에서의 색상 및 텍스처 정보의 차이를 통해 서로 다른 물체를 구분하는 데 도움을 준 것으로 추측한다. 이를 통해 RGB 영상의 풍부한 2차원 시각적 특징을 통해 포인트 클라우드의 기하학적 특징이 갖는 한계점을 극복한 것을 확인할 수 있다.

한편, Fig. 7B는 RGB 영상들로부터 동일한 물체에 대한 균일한 색상 및 텍스처 정보의 도움을 받아 정확한 분할을 수행하는 사례를 나타낸다. PTv2의 분할 결과를 살펴보면, 책상(desk)를 이루는 일부 포인트들에 대해 테이블(table)로 잘못된 분할이 이루어져 있다. 이는 책상과 테이블은 서로 유사한 구조로 되어 있어서 잘못 추출된 3차원 기하학적 특징으로 인해 같은 물체를 이루는 포인트들에 대해 일부는 책상으로, 일부는 테이블로 분류되었다고 판단된다. 반면, 제안 모델 RGB 영상들을 함께 활용한 MMCA-Net의 분할 결과에서는 책상을 이루는 모든 포인트들이 올바르게 분류가 이루어진 것을 확인할 수 있는데, 이는 RGB 영상들에서 책상에 해당하는 픽셀들이 모두 균일한 색상 정보를 가지고 있어 더 정확한 분할을 수행한 것으로 추측한다. 위 사례를 통해 RGB 영상들의 시각적 특징을 활용함으로써 정확한 분할을 수행하는데 있어 긍정적 효과를 미치는 것을 확인할 수 있다.

반면, Fig. 7C는 올바른 분할 작업을 수행한 PTv2와 달리, 제안 모델이 분할 작업에서 실패한 사례를 나타낸다. 테이블(table)을 이루는 포인트들에 대해 PTv2 모델은 올바른 분할을 수행한 반면, 제안 모델은 해당 물체를 이루는 일부 포인트들을 소파(sofa), 책상(chair) 등으로 잘못된 분할이 이루어졌다. 이러한 오분할의 원인으로는 해당 환경에서 소파와 테이블은 서로 유사한 색상과 텍스처를 가지고 있기 때문에 분별력이 약한 2차원 시각적 특징 추출이 이루어졌고, 제안 모델의 특징 융합 방식에서는 RGB 영상에 기초한 2차원 시각적 특징이 포인트 클라우드의 3차원 기하학적 특징보다 포인트 분할에 상대적으로 더 큰 영향을 주었기 때문인 것으로 추측한다. 위와 같은 제안 모델의 문제점을 해결하기 위해서는 분별력이 더 높은 2차원 시각적 인코더 채용과 더불어, 입력 데이터의 특성에 따라 2차원 특징과 3차원 특징의 가중치를 좀 더 세밀하게 조절할 수 있는 적응적 특징 융합 기법으로 확장하는 추가적인 연구가 필요한 것으로 판단된다.

5. 결 론

본 논문에서는 포인트 클라우드의 3차원 기하학적 특징 외에도 멀티-뷰 RGB-D 입력 영상에서 추출한 2차원 시각적 특징을 함께 활용하는 새로운 2차원-3차원 멀티-모달 특징 기반의 3차원 포인트 클라우드의 의미적 분할 모델 MMCA-Net을 제안하였다. 제안 모델은 입력 포인트 클라우드로부터 맥락 정보가 풍부한 3차원 기하학적 특징 추출을 위해 Transformer 기반의 PTv2[12]를 채용하였고, 2차원 시각적 특징과 3차원 기하학적 특징의 효과적인 융합을 위해 새로운 중기 융합 전략과 멀티-모달 교차 주의집중 블록(MCAtt Block)을 적용하였다. 본 논문에서는 제안 모델의 성능 분석을 위해 ScanNetv2 [18] 벤치마크 데이터 집합을 이용한 다양한 정량적, 정성적 실험들을 진행하였고, 이를 통해 제안 모델의 효과와 유용성을 확인하였다. 구체적으로는 성능 척도 mIoU 측면에서 제안 모델은 3차원 기하학적 특징만을 이용하는 PTv2 모델에 비해 9.2%의 성능 향상을, 2차원-3차원 멀티-모달 특징을 사용하는 MVPNet 모델에 비해 12.12%의 성능 향상을 보였다.

한편, 앞선 정성적 평가 실험에서 언급한 것과 같이, 유사한 구조를 가졌으나 서로 다른 물체의 포인트들에 대해서는 제안 모델이 멀티-뷰 영상의 2차원 시각적 특징을 함께 이용함으로써, 분할의 정확도를 향상시키는데 긍정적인 효과를 보였다. 하지만 일부 유사한 색상을 가진 물체들의 포인트들에 대해서는 3차원 기하학적 특징만을 이용하는 기존 모델보다 오히려 낮은 분할 정확도를 나타내 보였다. 이러한 현재 제안 모델의 한계점을 극복하기 위해서는 추가 학습 혹은 인코더 구조 개선을 통해 보다 분별력이 향상된 2차원 시각적 인코더를 채용하는 것과 입력 데이터의 특성에 따라 2차원 특징과 3차원 특징의 가중치를 좀 더 세밀하게 조절할 수 있는 적응적 특징 융합 기법에 대한 향후 연구가 필요하다고 판단한다. 또한, 본 논문에서 이용한 ScanNetv2 이외에 또 다른 대표 벤치마크 데이터 집합인 S3DIS를 이용해 모델의 확장 적용 가능성과 성능을 분석해보는 추가 연구도 필요하다고 판단한다.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp.234-241, 2015.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp.2881-2890.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.10012-10022, 2021.

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, Vol.34, pp.12077-12090, 2021.

[5] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.652-660, 2017.

[6] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems (NeurIPS)*, Vol.30, pp.5099-5108, 2017.

[7] Y. Wang, Y. Sun, Z. Liu, and S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *Journal of ACM Transactions on Graphics*, Vol.38, No.5, pp.1-12, 2019.

[8] H. Lei, N. Akhtar, and A. Mian, "Spherical kernel for efficient graph convolution on 3d point clouds," *Journal of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.43, No.10, pp.3664-3680, 2020.

[9] W. Wu, Z. Qi, and L. Fuxin, "PointConv: Deep convolutional networks on 3d point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.9621-9630, 2019.

[10] M. Xu, R. Ding, H. Zhao, and X. Qi, "PAConv: Position adaptive convolution with dynamic kernel assembling on point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3173-3182, 2021.

[11] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.16259-16268., 2021.

[12] X. Wu, Y. Lao, L. Jiang, .X. Liu, and H. Zhao, "Point transformer V2: Grouped vector attention and partition-based pooling," *arXiv preprint arXiv:2210.05666*, 2022.

[13] M. Jaritz, J. Gu, and H. Su, "Multi-view PointNet for 3d scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.3995-4003, 2019.

[14] C. Du, M. A. Vega Torres, Y. Pan, and A. Borrmann, "MV-KPConv: Multi-view KPConv for enhanced 3d point cloud semantic segmentation using multi-modal fusion with 2d images," in *Proceedings of the European Conference on Product and Process Modeling*, 2022.

[15] A. Dai, and M. Niessner, "3DMV: Joint 3d multi-view prediction for 3d semantic scene segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.452-468, 2018.

[16] L. Zhao, J. Lu, and J. Zhou, "Similarity-aware fusion network for 3d semantic segmentation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.1585-1592, 2021.

[17] W. Hu, H. Zhao, L. Jian, J. Jia, and T. T. Wong, "Bidirectional projection network for cross dimension scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CPVR)*, pp.14373-14382, 2021.

[18] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5828-5839, 2017.

[19] A. Boulch, B. L. Saux, and N. Audebert, "Unstructured point cloud semantic labeling using deep segmentation networks," *3dor@ eurographics*, Vol.3, pp.17-24, 2017.

[20] A. Boulch, J. Guerry, B. L. Saux, and N. Audebert, "SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks," *Computers & Graphics*, Vol.71, pp.189-198, 2018.

[21] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," *arXiv preprint arXiv:1602.07360*, 2016.

[22] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss. "RangeNet++: Fast and accurate LiDAR semantic segmentation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.4213-4220, 2019.

[23] J. Huang and S. You, "Point cloud labeling using 3d convolutional neural network," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp.2670-2675, 2016.

[24] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, M. Nießner, "ScanComplete: Large-scale scene completion and semantic segmentation for 3D scans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4578-4587, 2018.

[25] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. F. Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.248-255, 2009.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016.

[27] D. Menini, S. Kumar, M. R. Oswald, E. Sandstrom, C. Sminchisescu, and L. V. Gool, "A real-time online learning framework for joint 3d reconstruction and semantic segmentation of indoor scenes," *Journal of IEEE Robotics and Automation Letters*, Vol.7, No.2, pp.1332-1339, 2021.

배 혜 림

https://orcid.org/0000-0003-4179-0339
e-mail : thvk654@kyonggi.ac.kr
2023년 경기대학교 컴퓨터공학부(학사)
2023년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 기계학습, 컴퓨터비전

김 인 철

https://orcid.org/0000-0002-5754-133X
e-mail : kic@kyonggi.ac.kr
1985년 서울대학교 수학과(학사)
1987년 서울대학교 전산과학과(석사)
1995년 서울대학교 전산과학과(박사)
1996년 ~ 현 재 경기대학교 컴퓨터공학부 교수
관심분야 : 인공지능, 기계학습, 로봇지능