The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.7, pp.335~347
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.7.335

Class-Agnostic 3D Mask Proposal and 2D-3D Visual Feature Ensemble for Efficient Open-Vocabulary 3D Instance Segmentation

Sungho Song† · Kyungmin Park†† · Incheol Kim†††

ABSTRACT

Open-vocabulary 3D point cloud instance segmentation (OV-3DIS) is a challenging visual task to segment a 3D scene point cloud into object instances of both base and novel classes. In this paper, we propose a novel model Open3DME for OV-3DIS to address important design issues and overcome limitations of the existing approaches. First, in order to improve the quality of class-agnostic 3D masks, our model makes use of T3DIS, an advanced Transformer-based 3D point cloud instance segmentation model, as mask proposal module. Second, in order to obtain semantically text-aligned visual features of each point cloud segment, our model extracts both 2D and 3D features from the point cloud and the corresponding multi-view RGB images by using pretrained CLIP and OpenSeg encoders respectively. Last, to effectively make use of both 2D and 3D visual features of each point cloud segment during label assignment, our model adopts a unique feature ensemble method. To validate our model, we conducted both quantitative and qualitative experiments on ScanNet-V2 benchmark dataset, demonstrating significant performance gains.

Keywords : Point Cloud, Open-Vocabulary 3D Instance Segmentation, Class-Agnostic 3D Mask Proposal, 2D-3D Visual Feature Ensemble

효율적인 개방형 어휘 3차원 개체 분할을 위한 클래스-독립적인 3차원 마스크 제안과 2차원-3차원 시각적 특징 앙상블

송 성 호† · 박 경 민†† · 김 인 철†††

요 약

개방형 어휘 3차원 포인트 클라우드 개체 분할은 3차원 장면 포인트 클라우드를 훈련단계에서 등장하였던 기본 클래스의 개체들뿐만 아니라 새로운 신규 클래스의 개체들로도 분할해야 하는 어려운 시각적 작업이다. 본 논문에서는 중요한 모델 설계 이슈별 기존 모델들의 한계점들을 극복하기 위해, 새로운 개방형 어휘 3차원 개체 분할 모델인 Open3DME를 제안한다. 첫째, 제안 모델은 클래스-독립적인 3차원 마스크의 품질을 향상시키기 위해, 새로운 트랜스포머 기반 3차원 포인트 클라우드 개체 분할 모델인 T3DIS[6]를 마스크 제안 모듈로 채용한다. 둘째, 제안 모델은 각 포인트 세그먼트별로 텍스트와 의미적으로 정렬된 시각적 특징을 얻기 위해, 사전 학습된 OpenScene 인코더와 CLIP 인코더를 적용하여 포인트 클라우드와 멀티-뷰 RGB 영상들로부터 각각 3차원 및 2차원 특징들을 추출한다. 마지막으로, 제안 모델은 개방형 어휘 레이블 할당 과정동안 각 포인트 클라우드 세그먼트별로 추출한 2차원 시각적 특징과 3차원 시각적 특징을 상호 보완적으로 함께 이용하기 위해, 특징 앙상블 기법을 적용한다. 본 논문에서는 ScanNet-V2 벤치마크 데이터 집합을 이용한 다양한 정량적, 정성적 실험들을 통해, 제안 모델의 성능 우수성을 입증한다.

키워드 : 포인트 클라우드, 개방형-어휘 3차원 개체 분할, 3차원 마스크 제안, 2차원-3차원 시각적 특징 앙상블

1. 서 론

3차원 장면 이해(3D scene understanding) 기술은 자율 주행, 로봇 조작, AR/VR과 같은 다양한 분야에서 활용되는 중요한 시각 인식 기술이다. 3차원 포인트 클라우드 개체 분할(3D point cloud instance segmentation)은 3차원 장면 이해 기술들 중의 하나로서, 입력 장면 포인트 클라우드에서 각 개체들의 3차원 이진 마스크(3D binary mask)와 함께 클래스 레이블(class label)를 동시에 예측해내는 기술을 의미한다. 3차원 포인트 클라우드 개체 분할에 관한 연구들은 현재까지 상당한 발전을 이루었으나, 대부분의 기존 연구들[1-6]은 이미 훈련 데이터 집합(training dataset)에 포함되어 있는 개체 클래스 레이블로만 신규 포인트 클라우드의 개체 분할이 가능한 폐쇄형 어휘 개체 분할 방식(closed-vocabulary instance seg-

※ 본 연구는 2024년 경기대학교 대학원 연구원장학생 장학금 지원에 의하여 수행되었음.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 비 회 원 : 경기대학교 컴퓨터과학과 석사과정
††† 종신회원 : 경기대학교 AI컴퓨터공학부 교수
Manuscript Received : June 18, 2024
Accepted : July 10, 2024
* Corresponding Author : Incheol Kim (kic@kyonggi.ac.kr)

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

mentation)을 취하고 있다. 일반적으로 폐쇄형 어휘 개체 분할 방식은 포인트 클라우드에 포함된 각 개체 별 시각적 특징(visual feature)들을 추출해낸 뒤, 이를 바탕으로 훈련 단계에서 이미 제공되었던 클래스 레이블들을 이용해 각 개체들의 클래스를 분류(classification)한다. 따라서 폐쇄형 어휘 개체 분할 방식은 학습 단계에서는 제공되지 않았던 새로운 클래스의 개체들을 탐지하고 분할하는 것이 불가능하고, 이로 인하여 다양한 종류의 개체들이 등장할 수 있는 열린 실세계 응용분야들에 그대로 활용되기 어렵다는 한계점이 있다. 이러한 폐쇄형 어휘 개체 분할 방식의 한계를 극복하기 위해서, 대규모 자연어 어휘 지식(large language vocabulary knowledge)의 도움을 받아 훈련 데이터에 등장하지 않은 새로운 클래스 레이블의 개체들에 대해서도 탐지와 분할이 가능한 개방형 어휘 개체 분할 연구들(open-vocabulary instance segmentation) [7-10]이 최근에 등장하기 시작하였다.

개방형 어휘 시각 지능 기술들은 먼저 2차원 영상(2D image)들을 대상으로 한 영상 분류, 물체 탐지, 의미적 분할, 개체 분할 작업들에서 부터 적용되기 시작하였다. 이들은 대부분 CLIP[11], ALIGN[12] 등과 같은 사전 학습된 개방형 어휘 2차원 시각-언어 모델들(pretrained open-vocabulary 2D vision-language model)을 활용하여 구현되었다. CLIP[11]과 ALIGN[12] 모델은 영상-텍스트 쌍(image-text pair)들로 구성된 인터넷-규모(Internet-scale)의 대용량 영상 캡션(image caption) 데이터 집합들을 이용하여 사전 학습함으로써, 임의의 텍스트 질의들과 한 장의 영상으로부터 의미적으로 상호 정렬된 텍스트 특징과 시각적 특징(semantically aligned text and visual features)을 각각 추출할 수 있다. OpenSeg[13], ZegFormer[14], OVSeg[15], FreeSeg[16]와 같은 대부분의 개방형 어휘 2차원 영상 분할 모델들은 사전 학습된 CLIP[11] 혹은 ALIGN[12]을 활용하여, 영상 세그먼트들(image segments)의 시각적 특징과 개방형 어휘 레이블 텍스트들(open-vocabulary label texts)의 텍스트 특징을 추출하고, 시각적 특징들과 텍스트 특징들 간의 매칭을 통해, 각 영상 세그먼트에 가장 의미적으로 부합되는 레이블을 할당한다.

하지만 3차원 포인트 클라우드 개체 분할 혹은 의미적 분할 작업의 경우에는 아직까지 포인트 클라우드-텍스트 쌍(pointcloud-text pair)들로 구성된 인터넷-규모의 대용량 포인트 클라우드 캡션(pointcloud caption) 데이터 집합이 존재하지 않기 때문에, 포인트 클라우드에 적용 가능한 CLIP[11], ALIGN[12]과 유사한 사전 학습된 개방형 어휘 3차원 시각-언어 모델(pretrained open-vocabulary 3D vision-language model)들이 존재하지 않는다는 어려움이 있다. 이와 같은 현재의 여건과 상황들을 고려할때, 효과적인 개방형 어휘 3차원 포인트 클라우드 개체 분할 모델을 설계하기 위해서는 다음과 같은 중요한 도전적 과제(challenge) 혹은 이슈(issue)들을 해결해야 한다.

첫 번째 이슈는 입력 포인트 클라우드에 포함되어 있을 각 개체의 클래스-독립적인 3차원 마스크를 어떻게 정확히 예측해낼 것인가 하는 3차원 마스크 제안(3D mask proposal) 문제이다. 개체별 3차원 마스크 예측의 정확도는 이후 이를 기초로 포인트 클라우드에서 추출되는 개체별 시각적 특징의 품질과 최종적인 개체 분할의 성능에도 모두 큰 영향을 미칠 수 있기 때문에 매우 중요한 문제이다. OVIR-3D[9]와 SAM3D[17]는 3차원 마스크 제안을 위해 포인트 클라우드를 직접 이용하지 않았다. 대신 대응되는 멀티-뷰 RGB 영상들에 2차원 RGB 영상 분할 모델인 SAM[18]을 적용하여 각 개체의 2차원 마스크들을 구한 뒤 이들을 3차원 공간 상에 역투영 후 병합함으로써, 포인트 클라우드 내 개체별 클래스-독립적인 3차원 마스크들을 생성하였다. 하지만 멀티-뷰 RGB 영상들에서 얻은 많은 2차원 마스크들에는 3차원 개체의 일부 가시 영역들(visual portions)만을 포함되거나 배경(background)과 같이 개체와는 직접 관련이 없는 영역들도 함께 포함될 수 있어, 역투영과 병합을 통해 정확한 3차원 마스크를 얻기 어렵다는 문제점이 있다. 한편 OpenMask3D[7]와 OpenIns3D[8]는 사전 학습된 폐쇄형-어휘 3차원 개체 분할 모델인 Mask3D[5]을 직접 포인트 클라우드에 적용하여 개체들의 클래스-독립적인 3차원 마스크들을 생성하였다. 그러나 이 모델들에서 3차원 마스크 제안을 위해 공통적으로 이용하는 Mask3D[5]는 마스크 예측에 필요한 개체별 3차원 시각적 특징을 트랜스포머 디코더를 통해 집계하는 과정에서 디코더 쿼리의 정제 방식, 학습 가속화 방식 등 아직은 성능 개선의 여지가 많은 상태이다. 본 논문의 제안 모델에서는 이러한 Mask3D[5]의 한계점을 보완하고 성능을 향상시킨 T3DIS[6]를 클래스-독립적인 3차원 마스크 제안 모듈로 채용함으로써, 보다 정확한 개체별 3차원 마스크를 예측해내고자 한다.

두 번째 설계 이슈는 3차원 개체 마스크들로 분할된 포인트 클라우드 세그먼트들에서 의미적으로 텍스트와 잘 정렬된 시각적 특징(semantically text-aligned visual feature)을 어떻게 추출해낼 것인가 하는 문제이다. 분할된 3차원 포인트 클라우드 세그먼트들에 정확한 개방형 어휘 레이블들을 할당해주기 위해서는 각 세그먼트별로 의미적으로 텍스트와 잘 정렬된 시각적 특징을 추출해내는 것이 매우 중요하다. 기존의 연구들에서는 (1) CLIP[11]과 같은 개방형 어휘 2차원 시각-언어 모델로부터 3차원 시각-언어 모델로 지식 증류(knowledge distillation) 학습을 수행함으로써, 포인트 클라우드 세그먼트들에서 텍스트와 의미적으로 정렬된 3차원 시각적 특징을 얻는 방식[10, 19, 20], (2) 포인트 클라우드 세그먼트들에 대응되는 멀티-뷰 RGB 영상 세그먼트들을 찾고 이 영상 세그먼트들에서 CLIP 기반의 2차원 시각적 특징들을 추출하고 병합함으로써, 각 포인트 클라우드 세그먼트별로 텍스트와 정렬된 2차원 시각적 특징을 얻는 방식[7], 혹은 (3) 포인트 클라우드 세그먼트별로 시각적 특징을 따로 구하지 않고 대응

되는 멀티-뷰 RGB 장면 영상들에 SAM[18]과 같은 개방형 어휘 2차원 영상 분할 모델들을 적용하여, 그 분할 결과인 개방형 어휘 개체 레이블들을 3차원 포인트 클라우드 세그먼트들에도 그대로 할당하는 방식[8, 17] 등이 시도되었다. 하지만 이와 같이 대부분의 기존 모델들에서는 각 포인트 클라우드 세그먼트의 3차원 시각적 특징 혹은 2차원 시각적 특징 중 어느 한 쪽만을 추출하여 분할에 이용함으로써, 각 개체의 기하학적 특성(geometric characteristics)과 외관적 특성(appearance characteristics)을 함께 개체 분할에 반영할 수 없다는 한계점이 있다. 이와 같은 기존 모델들의 한계점을 극복하기 위해, 본 논문의 제안 모델에서는 포인트 클라우드 세그먼트들의 효과적인 레이블 결정을 위해 각 포인트 클라우드 세그먼트별 3차원 시각적 특징뿐만 아니라 2차원 시각적 특징도 함께 이용한다.

세 번째 설계 이슈는 클래스-독립적인 3차원 마스크들로 분할된 포인트 클라우드 세그먼트들에 최적의 개방형 어휘 레이블들을 할당하기 위해, 포인트 클라우드 세그먼트들의 시각적 특징(visual feature)과 개방형 어휘 레이블들의 텍스트 특징(text feature) 간의 매칭을 어떻게 수행할 것인가 하는 문제이다. 대부분의 기존 개방형 어휘 3차원 개체 분할 모델들[7-10]은 앞서 설명한대로 포인트 클라우드 세그먼트들의 2차원 혹은 3차원 시각적 특징만을 이용해, 레이블 텍스트 특징들과의 코사인 유사도(cosine similarity)를 계산한 후, 가장 높은 유사도를 가지는 클래스 레이블을 각 포인트 클라우드 세그먼트에 할당하는 방식을 공통적으로 적용하였다. 하지만 이러한 방식은 각 포인트 클라우드 세그먼트를 나타내는 시각적 특징으로 2차원 특징이나 3차원 특징만을 단일하게 사용하였을 경우에만 가능하다. 본 논문의 제안 모델과 같이 포인트 클라우드 세그먼트들의 레이블 결정에 각 포인트 클라우드 세그먼트의 2차원 시각적 특징과 3차원 시각적 특징을 상호 보완적으로 함께 사용하기 위해서는 2가지 서로 다른 모드의 시각적 특징들을 개방형 어휘 레이블 텍스트 특징들과 매칭하기 위한 새로운 방식이 도입되어야 한다. 본 논문의 제안 모델에서는 각 포인트 클라우드 세그먼트별로 2차원 시각적 특징과 3차원 시각적 특징 중 레이블 텍스트 특징과의 유사도가 더 높은 시각적 특징을 골라 선택적으로 매칭에 이용하는 특징 앙상블(feature ensemble) 기법을 적용한다.

본 논문에서는 이와 같은 개방형 어휘 3차원 포인트 클라우드 개체 분할 모델의 설계 이슈별 기존 모델들의 한계점들을 고려하여, 새로운 개방형 어휘 3차원 개체 분할 모델인 Open3DME(Open-vocabulary 3D instance segmentation with class-agnostic 3D Mask proposal and 2D-3D visual feature Ensemble)를 제안한다. 제안 모델은 3차원 마스크들의 품질을 향상시키기 위해, 트랜스포머 디코더의 쿼리 정제 방식 개선, 그리고 모델 학습 가속화를 위한 노이즈 제거 보조 학습 등이 적용된 트랜스포머 기반 3차원 개체 분할 모델인 T3DIS[6]를 3차원 마스크 제안 모듈로 채용한다. 또한, 제안 모델은 각 포인트 클라우드 세그먼트별로 사전 학습된 OpenScene[20] 기반의 3차원 포인트 인코더를 적용해 텍스트와 의미적으로 정렬된 3차원 시각적 특징을 추출할 뿐만 아니라, 멀티-뷰 RGB 영상 세그먼트들에도 사전 학습된 CLIP[11] 기반의 2차원 픽셀 인코더를 적용하여 텍스트와 의미적으로 정렬된 2차원 시각적 특징을 추출한다. 또 제안 모델은 개방형 어휘 레이블 텍스트 특징들과의 매칭에 각 포인트 클라우드 세그먼트별로 추출한 2차원 시각적 특징과 3차원 시각적 특징을 상호 보완적으로 함께 이용하기 위한 특징 앙상블 기법도 적용한다. 본 논문에서는 ScanNet-V2 벤치마크 데이터 집합을 이용한 다양한 정량적, 정성적 실험들을 통해, 제안 모델의 우수성을 입증한다. 서론에 이어 본 논문의 2장에서는 관련 선행 연구들을 살펴보고, 3장에서는 제안 모델의 구조와 모듈별 설계 세부 사항들에 대해 자세히 설명한다. 이어서 4장에서는 제안 모델의 구현 및 다양한 성능 평가 실험 결과들을 소개하고, 마지막 5장에서는 결론과 향후 연구 계획을 정리한다.

2. 관련 연구

폐쇄형 어휘 개체 분할 연구 방식들[1-6]은 모두 학습 시에 등장하는 클래스(seen class)에 대해서만 분할 및 분류를 수행하는 것을 목표로 한다. 기존의 폐쇄형 어휘 3차원 개체 분할 방식의 연구들은 크게 탐지 기반(detection-based approach) 접근법[1, 2], 군집화 기반 접근법(clustering-based approach)[3, 4], 그리고 트랜스포머 기반 접근법[5, 6](transformer-based approach)으로 나누어 볼 수 있다. 이들 중에서 트랜스포머 기반의 접근 방식[5, 6]은 다계층 트랜스포머 디코더(multilayer Transformer decoder)를 거치는 동안 개체 쿼리(instance query)들의 콘텐트(content)를 반복적으로 정제함으로써 포인트 클라우드에 포함되어 있을 각 개체별 3차원 시각적 특징 정보를 집계해낸다. 이렇게 개체 쿼리에 최종 집계된 각 개체별 3차원 시각적 특징 정보를 토대로, 각 개체의 3차원 이진 마스크와 클래스 레이블을 예측해낸다.

대표적인 트랜스포머 기반의 폐쇄형 어휘 3차원 개체 분할 모델인 Mask3D[5]는 다계층 디코더를 통한 개체 쿼리 콘텐트 정제 과정에서 개체 쿼리 콘텐트에 큰 영향을 미치는 각 개체의 위치 정보는 함께 정제하지 않는다. 따라서, 최종적으로 개체 쿼리의 콘텐트 영역에 집계되는 각 개체의 3차원 시각적 특징과 이를 기초로 예측되는 3차원 마스크의 품질에 제한이 있다. 또한 Mask3D[5]는 특별히 트랜스포머 디코더의 학습 가속화를 위한 보조 학습도 적용하지 않았다. T3DIS[6]는 이러한 Mask3D[5]의 한계점을 보완하고 성능을 향상시킨 모델로서, 다계층 트랜스포머 디코더를 이용한 개체 쿼리 콘텐트 정제 과정에서 중요한 각 개체의 위치 정보까지 병행적으로 함께 정제함으로써 최종적으로 집계되는 각 개체의 3차원 시각적 특징과 이를 기초로 예측되는 3차원 마스크의 품질을 향

상시켰다. 또한, T3DIS는 다계층 트랜스포머 디코더를 학습하는 과정에 노이즈 제거 보조 작업 학습을 적용함으로써, 학습 수렴성도 크게 가속화시켰다.

이와 같은 폐쇄형 어휘 3차원 개체 분할 연구들과는 달리, 개방형 어휘 3차원 개체 분할 연구들[7-10]은 모델 학습 단계에서는 제공되지 않았던 미지의 신규 클래스(unseen class)들도 모델 추론 단계에서 개체 분할이 가능하도록 모델들을 확장하고자 하였다. OpenMask3D[7]의 경우 포인트 클라우드 내 개체들의 영역으로 간주되는 클래스-독립적인 3차원 마스크들을 예측하기 위해 Mask3D[5]를 3차원 마스크 제안 모듈로 이용하였다. 그리고 제안된 3차원 마스크들을 멀티-뷰 RGB 영상들에 투영함으로써 각 개체의 2차원 영상 세그먼트들을 구해내고, CLIP 시각적 특징 인코더를 이용해 각 영상 세그먼트 별 2차원 시각적 특징을 추출해내었다. 이렇게 추출한 개체 세그먼트들의 2차원 시각적 특징들과 CLIP 텍스트 인코더를 이용해 얻은 개방형 어휘 레이블들의 텍스트 특징들 간의 코사인 유사도를 계산하여, 각 개체 세그먼트에 가장 유사도가 높은 클래스 레이블을 할당하는 방식을 취하였다. 한편, OpenIns3D[8]도 전반적으로 OpenMask3D[7]와 유사한 접근 방식을 취하였지만, 다만 포인트 클라우드에 대응하는 멀티-뷰 RGB 입력 영상들이 함께 주어진다고 가정하지 않고 대신 3차원 장면 포인트 클라우드를 여러 시점(view point)에서 투영하여 생성된 2차원 합성 영상(synthetic image)들을 이용한다는 차이점이 있다. 하지만 OpenMask3D와 OpenIns3D 모두 클래스-독립적인 3차원 개체 마스크 예측에 Mask3D를 공통적으로 이용하였는데, 앞서 설명한대로 Mask3D 내부의 트랜스포머 기반 디코딩 과정의 비효율성으로 인해 마스크 예측 정확도가 제한적이라는 문제점을 가지고 있다.

한편, OVIR-3D[9] 모델은 사전 학습된 2차원 개방형 어휘 개체 분할 모델인 Detic[21]를 사용해 멀티-뷰 RGB 영상들에서 2차원 개체 마스크들을 예측해내는 동시에, 각 개체 마스크에 해당하는 영상 세그먼트들의 2차원 시각적 특징들도 추출해내었다. 그리고 다수의 2차원 개체 마스크들을 3차원 포인트 클라우드 상에 역투영하고 병합하여, 개체별 3차원 마스크들을 생성해내었다. 또한 영상 세그먼트들의 2차원 시각적 특징들을 개방형 어휘 레이블 텍스트 특징과 매칭함으로써 각 개체 세그먼트에 의미적으로 가장 부합되는 레이블을 할당하였다. 하지만 이러한 OVIR-3D 모델은 다수의 2차원 개체 마스크들을 3차원 포인트 클라우드에 역투영하여 3차원 개체 마스크들을 구하는 과정에서 2차원 마스크들이 3차원 개체들의 일부 가시 영역들만 포함하거나 배경 영역까지 포함하고 있는 경우들이 다수 존재하여, 최종적으로 만들어지는 클래스-독립적인 3차원 마스크들의 품질이 낮다는 문제를 가지고 있다.

한편, Lowis3D[10] 모델의 경우는 사전 학습된 영상 캡션 생성(image captioining) 모델들을 이용해 포인트 클라우드의 멀티-뷰 RGB 영상들에 대한 설명 텍스트(caption text)들을 생성함으로써, 의미적으로 텍스트와 잘 정렬된 3차원 시각적 특징을 추출할 수 있는 3차원 포인트 인코더를 학습하였다. 그리고 이러한 포인트별 3차원 시각적 특징들을 토대로 그룹화를 수행하여 3차원 개체 마스크들을 생성할 뿐만 아니라, 개방형 어휘 레이블 텍스트 특징들과의 유사도 계산을 통해 레이블 할당도 수행하였다. 따라서 Lowis3D[10]는 각 개체를 분할 단위로 보는 개체 분할(instance segmentation)보다는 각 포인트를 분할 단위로 보는 의미적 분할(semantic segmentation)에 더 가까운 접근 방식을 취하고 있으며, 각 포인트의 기하학적 특성을 반영한 3차원 시각적 특징만 이용할 뿐 각 개체의 외관적 특성을 반영한 2차원 시각적 특징은 분할에 이용하지 못한다는 한계점이 있다.

3. 개방형 어휘 3차원 개체 분할 모델의 설계

3.1 모델 개요

제안 모델은 Fig. 1과 같이 3차원 포인트 클라우드, 멀티-뷰 RGBD 영상, 개방형 어휘 클래스 레이블을 입력으로 받고, (1) 클래스-독립적인 3차원 마스크 제안(Class-Agnostic 3D Mask Proposal), (2) 텍스트와 정렬된 3차원 특징 인코딩(Text-Aligned 3D Feature Encoding), (3) 텍스트와 정렬된 2차원 특징 인코딩(Text-Aligned 2D Feature Encoding), (4) 클래스 레이블 할당(Class Label Assignment) 단계로 분할과 분류를 수행한다. 클래스-독립적인 3차원 마스크 제안 단계에서 마스크 제안 모듈은 다해상도의 포인트별 3차원 시각적 특징들을 토대로, 개체별 3차원 시각적 특징들을 집계하고 예측 헤드를 통해 클래스 레이블이 아직 부여되지않은 3차원 마스크(Class-Agnostic 3D Mask)를 예측해낸다. 텍스트와 정렬된 3차원 특징 인코딩 단계에서는 예측된 3차원 마스크를 토대로 3차원 세그멘트들을 만들어내고 OpenScene[20]의 포인트 특징 인코더에 입력하여 세그먼트 단위의 3차원 시각적 특징 F^{3D}를 만들어낸다. 텍스트와 정렬된 2차원 특징 인코딩 단계에서는 멀티-뷰 RGB 영상에 3차원 마스크를 투영하고, 투영된 마스크 영역의 2차원 세그먼트들을 CLIP 시각적 특징 인코더에 입력하여 세그먼트 단위의 2차원 시각적 특징 F^{2D}를 구해낸다. 클래스 레이블 할당 단계에서는 클래스 정보를 포함하고있는 텍스트 프롬프트로부터 얻어진 텍스트 특징 F^{Text}과의 코사인 유사도를 토대로 F^{3D}와 F^{2D} 두 특징을 앙상블해낸 후, 각 시각적 특징이 나타내는 3차원 마스크 세그먼트들에게 클래스 레이블을 할당한다.

후속 절 들에서는 제안 모델의 핵심 단계들인 클래스 독립적인 3차원 마스크 제안(Class-Agnositc 3D Mask Proposal), 텍스트와 정렬된 3차원 특징 인코딩(Text-Aligned 3D Feature Encoding), 텍스트와 정렬된 2차원 특징 인코딩(Text-Aligned 2D Feature Encoding), 클래스 레이블 할당(Class Label Assignment)에 대하여 좀 더 자세히 설명한다.

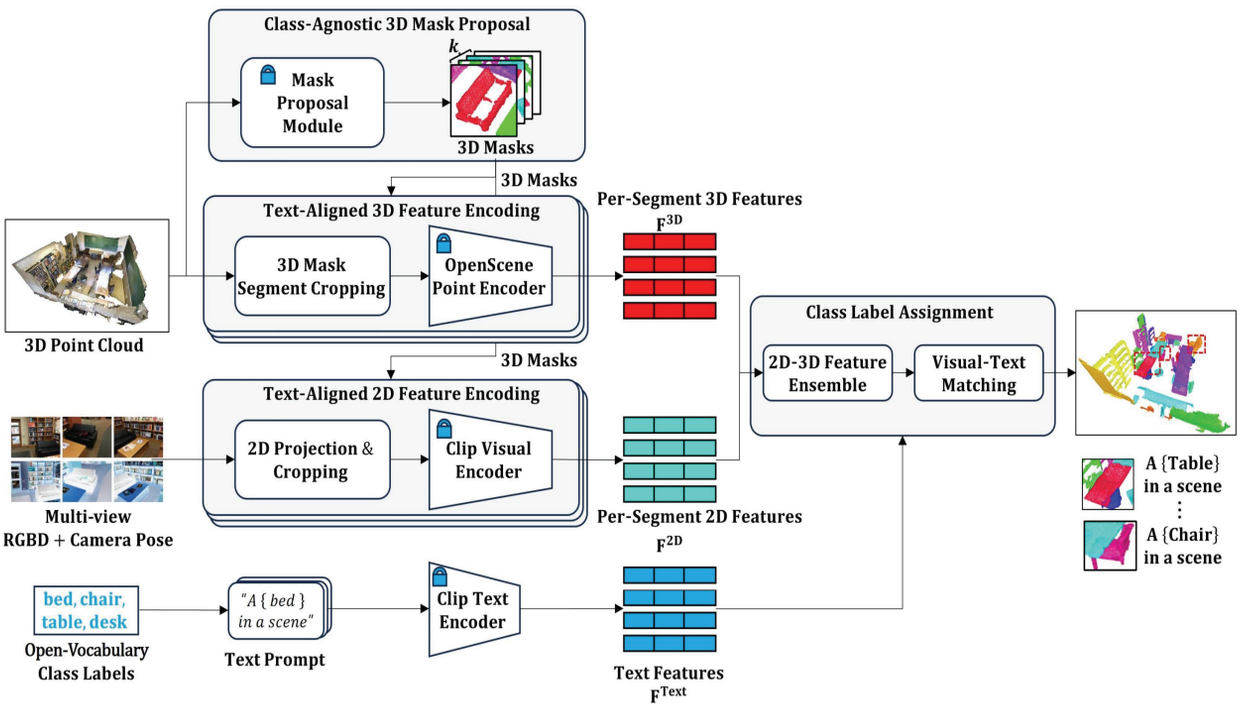

Fig. 1. Architecture of the Proposed Model

3.2 클래스-독립적인 3차원 마스크 제안

제안 모델은 3차원 마스크 제안 모듈로 ScanNet V2 데이터 집합에 사전학습된 트랜스포머 기반의 폐쇄형-어휘 3차원 개체 분할 모델 T3DIS[6]를 사용한다. T3DIS[6]는 기존의 연구들[7, 8]에서 대부분 3차원 마스크 예측에 사용했던 Mask3D[5]를 개선한 모델로, 디코더 계층 단위의 개체 쿼리 위치 인코딩 정보 정제, 노이즈 제거 보조 작업 학습이 적용되어 좀 더 정확한 마스크 예측이 가능한 모델이다. k개의 3차원 마스크를 예측하는 과정은 Fig. 2와 같다.

L계층의 개체 특징 디코더(Instance Feature Decoder)는 3차원 포인트 백본(3D Point Backbone)으로부터 전달받은 다해상도의 포인트별 3차원 시각적 특징 $F^0..F^L$을 토대로, 개체 쿼리와 마스크를 계층 단위로 정제한다. 쿼리 정제 과정에서는 개체 쿼리의 위치 정보 영역 P가 나타내는 개체별 예상 위치 정보와 동일 계층에서 정제된 마스크가 나타내는 영역 정보를 토대로 전체 포인트들의 시각적 특징 F^l로부터 개체별 시각적 특징을 개체 쿼리의 콘텐트 영역 C에 집계한다. 또한, 정제된 개체별 특징 C와 마스크를 토대로 쿼리의 위치 정보 영역 P도 정제한다. 마스크 정제 과정에서는 포인트별 시각적 특징 F^l과 개체별 시각적 특징 C를 내적하여 동일 계층의 쿼리 정제에 사용될 개체별 마스크를 정제해낸다. 모든 정제 과정이 완료된 이후에는, 최종 정제된 개체별 시각적 특징 $C^L \in k \times D$을 토대로 K개의 클래스 독립적인 3차원 마스크 $M=(m_1...m_k)$을 예측해낸다. 또한, 3차원 마스크 제안 모듈의 사전 학습 단계에서는 노이즈 제거 보조 작업 학습을 위해 생성된 보조 쿼리 Q_{Dn}을 추가로 디코더에 공급하고, 노이즈가 적용되기 전의 원본 데이터를 복원해내도록 추가로 학습한다. 이렇게 사전 학습된 3차원 마스크 제안 모듈이 예측한 k개의 3차원 마스크들은 텍스트와 정렬된 2차원, 3차원 특징 인코딩에 사용된다.

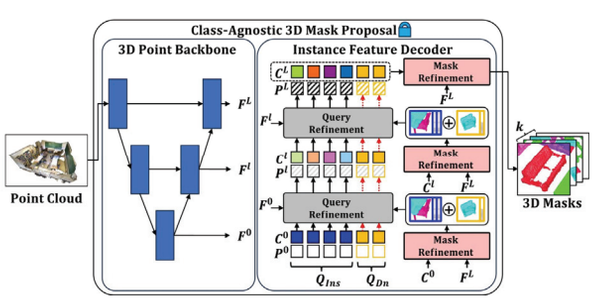

Fig. 2. Class-Agnostic 3D Mask Proposal

3.3 텍스트와 정렬된 3차원 특징 인코딩

이번 절에서는 텍스트와 정렬된 3차원 시각적 특징을 인코딩하는 과정에 대하여 설명한다. 앞서 예측된 3차원 마스크에 정확한 클래스 레이블을 할당해주기 위해서는 마스크 별로 텍스트와 의미적으로 잘 정렬되어 매칭할 수 있는 시각적 특징을 만들어내야한다. 3차원 시각적 특징을 인코딩하는 방식은 세 가지를 적용해볼 수 있었다.

(1) 첫 번째 방법은 이전 절에서 설명했던 3차원 마스크 제안 모듈의 개체 특징 디코더에서 최종 정제된 개체별 3차원

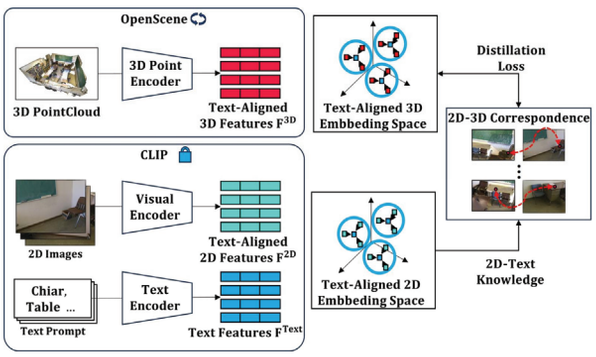

Fig. 3. 3D Visual Feature Enocoding with OpenScene

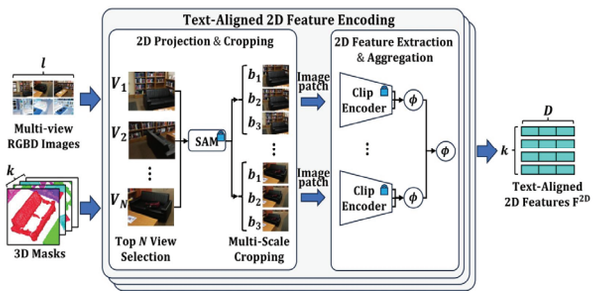

Fig. 4. 2D Visual Feature Encoding with CLIP

시각적 특징 C^L을 사용하는 방법(Built-in Features of 3D Mask Proposal Module)이다. (2) 두 번째 방법은 사전 학습된 3차원 포인트 백본[22, 23]을 통해 마스크에 포함된 포인트들의 3차원 시각적 특징을 추출하여 사용하는 방법이다. 첫 번째 방법과 두 번째 방법으로 얻어진 특징들은 개체별 기하학적 특성을 나타내고 있지만, 텍스트 표현과 의미적으로 정렬되어있지 않아 유의미한 시각-언어 간의 매칭이 불가능하다. 따라서, 시각-언어 유사도를 토대로 할당되는 개방형-어휘 레이블의 정확성을 기대하기 힘들다. 클래스 레이블 할당의 정확도를 향상시키기 위해선, 마스크를 나타내는 시각적 특징들은 기하학적 특성을 나타내는 것뿐만 아니라, 텍스트와 의미적으로 정렬되어 양쪽 특징 간의 매칭이 가능해야한다.

(3) 세 번째 추출 방법은 OpenScene[20]과 같이 개방형 어휘 2차원 시각-언어 모델 CLIP[11]으로부터 지식을 증류받은 3차원 시각-언어 모델[19, 20]기반의 텍스트와 정렬된 포인트 인코더(Text-Aligned 3D Point Encoder)를 사용하는 방법이다. 3차원 시각-언어 모델은 Fig. 3과 같이 학습 시간 동안 CLIP[11]의 2차원 시각-언어 지식(2D-Text Knowledge)을 전달받고, 이 지식을 토대로 2차원 픽셀과 3차원 포인트를 연관(2D-3D Correspondence)지어 텍스트와 의미적으로 정렬된 3차원 시각적 특징(Text-Aligned 3D Feature)을 추출할 수 있도록 훈련된다.

따라서 3차원 시각-언어 모델의 포인트 인코더를 통해서 추출된 특징들은 다른 두 가지 인코딩 방식들과는 다르게, 다양한 텍스트 표현들과 의미적 매칭이 가능하다. 제안 모델은 OpenScene[20]의 포인트 인코더를 통해 추출된 3차원 시각적 특징들을 사용한다. OpenScene[20]은 학습 시에 지정한 클래스 레이블에 대해서만 지식을 증류 받는 CLIP2Scene[19]과는 다르게, 증류 받을 개방형-어휘 레이블을 한정 짓지 않고 학습하여 더 다양한 텍스트 표현과 의미적으로 매칭이 가능하다는 장점을 가지고 있다.

OpenScene[20]의 포인트 인코더를 통해 3차원 시각적 특징이 인코딩되는 과정은 다음과 같이 이루어진다. 먼저 3차원 마스크 제안 모듈로부터 전달받은 k개의 3차원 마스크를 토대로, 포인트 클라우드 세그먼트들을 잘라낸다. 그런 다음 각 세그먼트들을 OpenScene[20]의 포인트 인코더에 입력하여 세그먼트별 3차원 시각적 특징을 추출해낸다. 이 과정은 세그먼트 수만큼 반복되어 최종적으로 k개의 세그먼트들을 나타내는 3차원 시각적 특징 $F^{3D} \in k \times D$가 얻어지게 된다.

3.4 텍스트와 정렬된 2차원 특징 인코딩

이번 절에서는 텍스트와 정렬된 2차원 시각적 특징 인코딩 해내는 과정에 대하여 구체적으로 설명한다. 2차원 시각적 특징을 인코딩해내는 과정은 Fig. 4와 같이 크게 3차원 마스크를 멀티-뷰 RGB 영상에 투영하고 잘라내는 과정(2D Projection & Cropping), 2차원 특징 추출 및 집계 과정(2D Feature Extraction & Aggregation) 두 단계로 구성된다.

첫 번째 단계에서는 입력받은 l개의 멀티-뷰 RGBD 영상과 k개의 3차원 마스크를 토대로 상위 N개의 시점의 RGB 영상을 선별해낸다. 구체적으로 설명하면, 식 (1)과 같이 3차원 마스크에 포함된 포인트 $P_{3D} = (x, y, z, 1)$를 각 RGB 영상 시점의 카메라 파라미터 I, $R \mid T$와의 연산을 통해 2차원 투영(2D projection)하여 $P_{2D} = (u, v, w)$를 구해낸다.

$$
\begin{aligned}
&P_{3D} = (x, y, z, 1) \\
&P_{2D} = I \bullet (R \mid t) \bullet P_{3D} = (u, v, w) \\
&\\
&depth_{proj} = w \\
&depth_{gt} = d \\
&\\
&w - d > threshold \rightarrow occluded \\
&w - d < threshold \rightarrow appeared
\end{aligned}
\tag{1}
$$

Equation (1)에서 (u, v)는 3차원 마스크에 포함된 포인트들의 RGB 영상에서의 픽셀(pixel) 좌표를, w는 깊이 값(depth value)을 나타낸다. 그런 다음, 구해진 P_{2D}의 픽셀 좌표 (u, v)에서의 깊이 값 w과 입력된 깊이 영상(depth image)에서의 실제 깊이 값 d와 비교하여, 마스크의 영상에서의 가시성을 판단한다. $w - d$가 임계값보다 큰 경우에는 포인트가 다른 사물에 의해 가려져 있다고(occluded) 판단하고, 작은 경우에는 영상에 등장한 것으로 판단하여 마스크 별로 영상에 등장한 포인트 수를 계산한다. 이렇게 계산된 등장 포인트 수가 가장 많은 상위 N개의 시점 $V_1 \cdots V_N$을 선별(top N

selection)하고 개방형 2차원 분할 모델인 SAM[18]에 입력하여 2차원 영상 세그먼트들을 구해낸다. 그런 다음, 각 시점의 영상을 세그먼트를 중심으로 m개의 다양한 크기의 이미지 패치(image patch)로 잘라내는 작업(multi-scale cropping)을 수행한다. 잘라낼 이미지 패치의 크기는 Equation (2)와 같이 계산된다.

$$
\begin{aligned}
&tight\ bbox\ b^1 = (x_1^1, y_1^1, x_2^1, y_2^1)\\
&(0 \le x_1^1 < x_2^1 < W, 0 \le y_1^1 < y_2^1 < H)\\
&x_1^m = \max(0, x_1^1 - (x_2^1 - x_1^1) \cdot k_{\exp} \cdot m)\\
&y_1^m = \max(0, y_1^1 - (y_2^1 - y_1^1) \cdot k_{\exp} \cdot m)\\
&x_2^m = \min(0, x_2^1 + (x_2^1 - x_1^1) \cdot k_{\exp} \cdot m, W-1)\\
&y_2^m = \min(0, y_2^1 + (y_2^1 - y_1^1) \cdot k_{\exp} \cdot m, H-1)
\end{aligned}
\tag{2}
$$

$b^1 = (x_1^1, y_1^1, x_2^1, y_2^1)$은 SAM[18]으로부터 예측된 영상 세그먼트를 빈틈없이 둘러쌓는 경계 상자의 좌표를, W와 H는 원본 영상의 너비와 높이를 나타내며, $k_{\exp}$는 잘라낼 크기를 조절할 확장 변수를 의미한다. m번 째로 잘라낼 b^m의 범위는 b^1의 너비$(x_2^1 - x_1^1)$와 높이$(y_2^1 - y_1^1)$에 각각 $k_{\exp} \cdot m$를 곱한 값을 각 꼭짓점에 더하거나 빼서 결정되며, b^1에서부터 점점 넓은 영역을 잘라내게 된다. 계산 과정에서 원본 RGB 영상의 경계 $(0 \sim W-1,\ 0 \sim H-1)$를 초과할 경우에는 경계 지점의 좌표로 대체 된다. 제안 모델은 확장 변수 $k_{\exp}$의 값은 0.2로, 잘라낼 영상 수 m은 3으로 설정하였다.

두 번째 단계인 2차원 특징 추출 및 집계 단계에서는 먼저, 같은 영상으로부터 다양한 크기로 잘라낸 m개의 이미지 패치들을 각각 CLIP 시각적 특징 인코더에 입력하여 특징들을 추출해내고 평균 집계(mean aggregation)한다. 그런 다음, 동일한 3차원 마스크로부터 구해진 영상 세그멘트들의 특징을 다시 한번 평균 집계하여 마스크 별로 텍스트와 정렬된 2차원 시각적 특징을 구해낸다. 앞서 설명한 모든 과정은 3차원 마스크의 개수인 k번만큼 반복되며 최종적으로 k개의 2차원 시각적 특징 $F^{2D} \in k \times D$가 구해지게 된다.

3.5 클래스 레이블 할당

제안 모델에서 각 포인트 클라우드 세그먼트의 2차원과 3차원 시각적 특징들의 앙상블(2D-3D Feature Ensemble)을 기초로, 개방형 어휘 레이블들의 텍스트 특징들과 시각-언어 매칭(Vision-Text Matching)을 통해 각 3차원 포인트 클라우드 세그먼트에 클래스 레이블을 할당(Class Label Assignment)하는 과정은 Fig. 5와 같다.

먼저, 2D-3D 특징 앙상블 모듈은 2차원 시각적 특징 $F^{2D} \in k \times D$, 3차원 시각적 특징 $F^{3D} \in k \times D$를 토대로, 각각 텍스트 특징 $F^{Text} \in q \times D$ 와의 코사인 유사도를 비교하여 클래스 레이블 할당에 사용될 특징들을 앙상블한다. 구체적으로

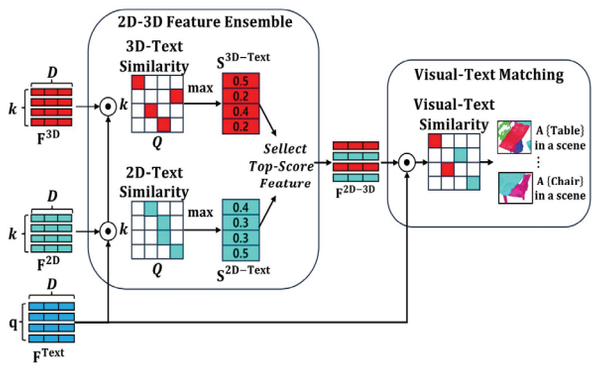

Fig. 5. Class Label Assignment

설명하면, 먼저 Equation (3)과 같이 $F^{2D} = (f_1^{2D}, f_2^{2D}, \ldots, f_k^{2D})$와 $F^{3D} = (f_1^{3D}, f_2^{3D}, \ldots, f_k^{3D})$을 각각 F^{Text}와 내적하여 코사인 유사도 점수 행렬을 생성하고, 각각의 점수 지도에서 가장 높은 하나의 클래스에 대한 유사도 점수 $S^{2D-Text}$, $S^{3D-Text}$를 구해낸다.

$$
\begin{aligned}
S_i^{2D-Text} &= \max\left\{\cos(f_i^{2D}, F^{Text}) \mid i = 1,..,k\right\}\\
S_i^{3D-Text} &= \max\left\{\cos(f_i^{3D}, F^{Text}) \mid i = 1,..,k\right\}
\end{aligned}
\tag{3}
$$

그런 다음, Equation (4)와 같이 동일 포인트 세그먼트를 나타내는 2차원과 3차원 특징의 최대 유사도 점수를 서로 비교하여 좀 더 높은 점수를 가지는 하나의 특징을 선택함으로써, 앙상블된 시각적 특징 F를 구해낸다.

$$
\begin{aligned}
&If\ S_i^{2D-Text} \ge S_i^{3D-Text}\\
&Then\ f_i^{2D-3D} = f_i^{2D}\\
&Else\ f_i^{2D-3D} = f_i^{3D}\\
&F^{2D-3D} = \left\{f_1^{2D-3D}, \ldots, f_K^{2D-3D}\right\}
\end{aligned}
\tag{4}
$$

이렇게 구해진 앙상블된 시각적 특징 F는 비전-언어 매칭 과정에서 다시 한번 F^{Text}와 내적을 통해 코사인 유사도를 계산하고, 유사도 점수가 가장 높은 텍스트 특징의 개방형-어휘 레이블을 특징들이 나타내는 3차원 마스크에 할당한다.

4. 구현 및 실험

제안 모델 Open3DME의 학습과 성능 테스트를 위해, 다양한 3차원 실내 환경 포인트 클라우드들과 이들에 대한 18개 클래스 개체 분할 레이블들로 구성된 대용량 벤치마크 데이터 집합인 ScanNet-V2를 이용하였다. 이 중 1201개의 장면 데이터들은 본 모델을 구성하는 각 모듈들의 사전 학습에 이용하였고, 312개의 장면 데이터들은 모델의 성능 테스트에 이용하였다. 제안 모델의 3차원 마스크 제안 모듈은 T3DIS[6]를 사

전 학습하여 사용하였다. 이때 사전 학습을 위해 학습률(learning rate)은 0.0001, 최적화 함수(optimizer)는 AdamW를 사용하였다. 한편, 제안 모델의 레이블 텍스트 인코더와 2차원 시각적 특징 인코더는 특징 차원이 768차원인 사전 학습된 CLIP ViT-L/14 모델[11]의 텍스트 특징 인코더와 시각적 특징 인코더를 채용하였다. 한편, 제안 모델의 3차원 시각적 특징 인코더로는 CLIP ViT-L/14 모델[11]의 지식 전이를 통해 사전 학습된 OpenScene[20]을 채용하였다. 제안 모델 Open3DME는 Ubuntu 18.04.6 LTS 환경에서 Pytorch 딥러닝 라이브러리를 이용하여 구현하였으며, GeForece GTX 3090 GPU 2개가 설치된 하드웨어 환경에서 학습 및 성능 테스트를 진행하였다.

첫 번째 실험은 제안 모델에서 3차원 마스크 제안을 위해 채택한 T3DIS[6]의 우수성을 입증하기 위한 실험이다. 이 실험에서는 (1) 마스크 제안 모듈로 Mask3D[5]를 사용한 경우, T3DIS[6]를 사용하되 (2) 위치 정제만 적용한 경우, (3) 노이즈 제거 보조 작업 학습만 적용한 경우, 그리고 (4) 제안 모델과 같이 위치 정보 정제와 노이즈 제거 보조 작업 학습이 모두 적용된 경우들을 비교하였으며, 이들의 3차원 마스크 예측 성능과 최종 개방형 어휘 개체 분할 성능을 각각 비교해본다.

Table 1의 실험 결과들을 살펴보면, 제안 모델과 같이 위치 정보 정제와 노이즈 제거 보조 작업 학습이 모두 적용된 T3DIS[6]를 사용한 모델이 다른 세 모델들에 비해 가장 높은 마스크 예측 성능과 개방형 어휘 개체 분할 성능을 보여주였다. 구체적으로 설명하면, 위치 정보 정제와 노이즈 제거 보조 작업 학습이 모두 적용된 T3DIS[6]를 사용한 분할 모델이 성능 척도 AP 면에서 마스크 예측 성능은 Mask3D[5]를 사용한 모델보다 16.05((0.535-0.461)/0.461)%, 위치 정보 정제만 적용된 T3DIS[6]를 사용한 모델보다 13.34%, 노이즈 제거 보조 작업 학습만 적용된 모델보다 12.63%의 성능 향상률을 보여주었다. 또한, 향상된 마스크 예측 성능에 따라 개방형 어휘 개체 분할 성능 면에서도 각각 4.27%, 2.87%, 2.09% 성능 향상이 있음을 확인할 수 있었다. 한편, T3DIS[6]에 위치 정보 정제만을 적용하여 사용한 모델이 Mask3D[5]를 사용한 모델에 비하여 마스크 예측 성능은 2.38%, 개방형 어휘 개체 분할 성능은 1.42%의 성능 향상률을, 노이즈 제거 보조 작업 학습만을 적용하고 사용한 모델이 Mask3D[5]를 사용한 모델과 비교하여 마스크 예측 성능은 3.03%, 개방형 개체 분할 성능은 2.13% 성능 향상률을 보였다. 이와 같은 실험 결과들을 통해서, 제안 모델의 마스크 제안 모듈인 T3DIS[6]에 적용된 위치 정보 정제와 노이즈 제거 보조 작업 학습이 마스크 예측뿐만 아니라 개방형 어휘 개체 분할의 성능 향상에도 도움을 줄 수 있다는 것을 확인할 수 있었다.

Table 1. Performance Comparison with Different 3D Mask Proposal Methods

Models	Position Refine	Denoising Aux Training	3D Mask AP	Segmentation AP
Mask3D	-	-	0.461	0.281
T3DIS	✓	-	0.472	0.285
	-	✓	0.475	0.287
	✓	✓	0.535	0.293

두 번째 실험은 제안 모델에서 각 포인트 세그먼트의 3차원 시각적 특징을 추출하기 위해 채택한 OpenScene의 긍정적 효과를 확인해보기 위한 실험이다. 이 실험에서는 Table 2와 같이 (a) 3D 마스크 제안 모듈로 Mask3D[5]와 T3DIS[6]를 사용하였을 때, 최종 정제된 개체 쿼리 콘텐트 영역의 개체별 특징 벡터를 세그먼트들을 나타내는 3차원 시각적 특징으로 사용하였을 경우와 (b) 3D 포인트 특징 백본인 Mink-Unet[22]과 PT-V2[23]에 포인트 세그먼트들을 입력하여 얻어진 특징을 사용한 경우, 그리고 (c) 3D 시각-언어 모델 Clip2Scene[19]과 OpenScene[20]의 텍스트와 정렬된 3차원 포인트 인코더에 포인트 세그먼트들을 입력하여 얻어진 3차원 시각적 특징을 사용한 경우들의 성능을 비교해본다. 또한, 이 실험에서는 3차원 시각적 특징만을 사용하였을 경우(Only 3D Features)의 개방형 어휘 개체 분할 성능뿐만 아니라, 3차원 시각적 특징이 2차원 시각적 특징과 앙상블(2D-3D Ensemble)되어 분할을 수행한 경우의 성능도 함께 비교해본다.

Table 2의 실험 결과들을 살펴보면, 제안 모델과 같이 3차원 시각적 특징 인코딩에 OpenScene[20]의 포인트 인코더를 사용한 모델이 3차원 시각적 특징만 사용한 경우와 2차원 특징과 앙상블하여 사용한 경우 모두 가장 높은 성능을 기록하였다. 먼저 3차원 시각적 특징만을 사용한 경우(Only 3D Feature)들을 결과를 살펴보면, (a) 마스크 제안 모듈[5, 6]을 3차원 시각적 특징 인코딩에 사용한 경우에는, 각각 AP 0.011%, 0.014%로 매우 저조한 성능을 보였다. 마찬가지로, (b) 3차원 포인트 특징 백본[22, 23]을 사용한 경우들도 각각 0.002%, 0.001%의 낮은 성능을 기록하였다. 이러한 결과를 통해, 시각(Vision)과 언어(Language)의 정렬과 매칭을 통해 3차원 마스크에 클래스 레이블을 할당하는 개방형 어휘 개체 분할 작업에서는, (a)와 (b) 같이 텍스트와 의미적으로 정렬되어 있지 않은 3차원 시각적 특징을 추출하는 방식으로는 성능 향상을 기대하기 어렵다는 것을 확인할 수 있었다.

Table 2. Performance Comparison with Different 3D Visual Feature Encoding Methods

Encoding Methods	Model	Only 3D Features	2D-3D Ensemble
(a) Built-in 3D Mask Proposal Module	(1) Mask3D	0.011	0.281
	(2) T3DIS	0.014	0.283
(b) 3D Point Backbone	(1) Mink UNet	0.002	0.283
	(2) PT V2	0.001	0.283
(c) Text-Aligned 3D Point Encoder	(1) Clip2Scene	0.237	0.241
	(2) OpenScene	0.276	0.293

반면, (c) 3D 시각-언어 모델[19, 20]의 3D 포인트 인코더를 3차원 시각적 특징 인코딩에 사용한 모델에서는, CLIP2Scene[19]의 포인트 인코더를 사용한 경우 AP 0.237, 제안 모델과 같이 OpenScene[20]의 포인트 인코더를 사용한 경우 AP 0.276로 가장 높은 성능을 기록하였다. 개방형-어휘 레이블을 제한하지 않고 지식을 증류받은 OpenScene[20]의 포인트 인코더를 사용한 경우가 제한된 개방형 어휘 레이블에 대해서만 지식을 증류받은 CLIP2Scene[19]의 포인트 인코더를 사용했을 때보다 약 16.45%의 성능 향상률을 보이면서, OpenScene[20]의 우수성을 확인할 수 있었다.

이어서 각 방식들을 통해 인코딩된 3차원 시각적 특징들을 2차원 시각적 특징들과 앙상블하여 분할 작업에 사용한 경우들의 결과를 살펴보면, 제안 모델과 같이 OpenScene[20]의 3차원 포인트 인코더를 사용한 경우가 AP 0.293%로 가장 높은 성능을 기록하였다. 또한, 3차원 시각적 특징만을 사용하였을 경우와 비교하여 6.15% 성능 향상률을 보이며, 앙상블되고 나서도 분할 성능 향상에 도움이 되는 것을 확인할 수 있었다. 반면에, (a)와 (b)의 경우에는 텍스트와 의미적으로 정렬되어 있지 않은 3차원 시각적 특징들로 인하여 앙상블되었을 때, 2차원 시각적 특징에만 의존하는 결과를 보여주었다. (c) Clip2Scene[19]의 포인트 인코더를 사용하였을 때는 3차원 시각적 특징만 사용한 경우보다 앙상블되어 사용한 경우가 1.68%의 성능 향상률을 보이며, OpenScene[20]의 포인트 인코더를 사용한 모델보다 비교적 낮은 성능 향상률을 보여주었다. 이러한 실험 결과들을 통해서 제안 모델이 채용한 3차원 비전-언어 모델 OpenScene[20]의 포인트 인코더를 사용하여 3차원 시각적 특징을 인코딩하는 방식이 단일 3차원 특징을 사용하였을 때뿐만 아니라, 2차원 특징과 앙상블되어 사용하였을 때에도 분할 성능 향상에 가장 도움이 된다는 것을 확인할 수 있었다.

세 번째 실험은 제안 모델에서 적용한 2차원 시각적 특징과 3차원 시각적 특징에 대한 특징 앙상블(feature ensemble) 방식의 성능 개선 효과를 입증하기 위한 실험이다. 이 실험에서는 2차원과 3차원 특징을 앙상블하는 방식으로 (1) 각 특징별 가장 높은 하나의 클래스 유사도 점수끼리 비교하여 앙상블하는 하드-보팅(Hard Voting) 방식과 (2) 모든 클래스에 대한 평균 유사도 점수를 비교하여 앙상블하는 소프트-보팅(Soft Voting)방식을 적용한 경우의 성능을 비교하였다. 그리고 앙상블 기법의 긍정적 효과를 확인해보기 위해, 2차원과 3차원 시각적 특징이 앙상블되지 않고 단일하게 사용된 경우들(Only 2D Feature, Only 3D Feature)과도 성능을 비교해본다. 또한, 앙상블 방식들에는 각 특징 별 유사도 점수에 동일한 가중치를 부여하는 방식(Rigid Score Weight)과, 단일 특징만을 사용하였을 때 높은 성능을 기록한 클래스들에 대하여 적응적으로 높은 가중치를 부여하는 방식(Adaptive Score Weight)을 적용하여 개방형 어휘 개체 분할 성능을 비교해본다.

Table 3의 실험 결과를 살펴보면, 적응적 하드 보팅 앙상블(Adaptive Hard Voting Ensemble, AHVE) 기법을 적용한 경우가 다른 경우들에 비해 가장 높은 분할 성능을 나타낸 것을 확인할 수 있었다. 구체적으로 살펴보면, 적응적 하드 보팅 앙상블 기법(Adaptive Hard Voting Ensemble, AHVE)을 적용한 분할 모델이 평가 지표 AP면에서 균등 소프트 보팅 앙상블(Rigid Soft Voting Ensemble, RSVE)보다 3.53%, 적응적 소프트 보팅 앙상블(Adaptive Soft Voting Ensemble, ASVE)보다 1.03%, 균등 하드 보팅 앙상블(Rigid Hard Voting Ensemble, RHVE) 기법을 적용한 경우보다 3.53%의 성능 향상률을 보였다.

Table 3. Performance Comparison with Different 2D-3D Ensemble Methods

Visual Features	Feature Ratio	AP
Only 2D Feature	100 : 0	0.283
Only 3D Feature	0 : 100	0.276
2D-3D Ensemble Methods		
RSVE	95 : 5	0.283
ASVE	64 : 36	0.290
RHVE	89 : 11	0.285
AHVE	67 : 33	0.293

또한, 각 앙상블 방식 별로 선별된 2차원과 3차원 특징들의 비율을 비교해보면, 2차원과 3차원 특징에 균등한 가중치를 부여한 RHVE, RSVE 방식들은 약 9:1의 비율로 3차원 특징이 거의 사용되지 않고 2차원 특징에만 의존적인 모습을 보였고, 2차원 특징만을 사용했을 경우(Only 2D Feature)와 별반 성능이 다르지 않은 결과를 확인할 수 있었다. 반면, ASVE, AHVE 방식들은 약 6:4의 비율로 비교적 2차원과 3차원 시각적 특징들을 골고루 사용하였고, 2차원 특징만을 사용한 경우(Only 2D Feature)보다 2.47%, 3.53%, 3차원 특징만을 사용한 경우(Only 3D Feature)보다 5.07%, 6.15%의 성능 향상률을 보이는 것을 확인할 수 있었다. 이러한 실험 결과들을 통해서 균일한 가중치를 부여하는 앙상블 방식들은 3차원 특징의 장점을 잘 활용하지 못하는 모습을 보이는 반면, 적응적 가중치를 부여하는 앙상블 방식들이 각 특징이 가지는 장점들을 비교적 잘 활용하였고, 그 중에서 적응적 하드 보팅 앙상블 방식이 가장 분할 성능 향상에 효과적인 것을 확인할 수 있었다.

네 번째 정량적 실험은 기존의 3차원 개체 분할 모델들과의 비교를 통해, 제안 모델 Open3DME의 우수성을 입증하기 위한 실험이다. 이 실험에서는 Table 4와 같이 대표적인 개방형-어휘 3차원 개체 분할 모델들인 OpenMask3D[7], OpenIns3D[8] 뿐만 아니라, 폐쇄형 어휘 3차원 개체 분할 모델들인 Mask3D[5]와 T3DIS[6]들의 성능도 제안 모델과 비교해보았다. 비교 모델인 OpenMask3D[7], OpenIns3D[8]는 마스크 제안 네트워크로 Mask3D[5]를 사용하여 제안 마스크의 품질이 저하되었고, 모두 2차원과 3차원 특징 앙상블을 적용하지 않았다. 또한, 이번 실험에서는 다른 모델들과의 성능 비교뿐만 아니라 제안 모델이 앙상블을 적용하지 않고 2차원 특징만을

Table 4. Performance Comparison with Other Models

Models	Avg	cab	bed	chair	sofa	table	door	wndw	bkshf	pic	cntr	desk	crtn	fridge	s.curt	toilet	sink	bath
Closed-Vocabulary 3D Instance Segmentation																		
Mask3D[5]	0.461	0.341	0.540	0.742	0.355	0.536	0.464	0.371	0.236	0.357	0.247	0.278	0.385	0.362	0.442	0.898	0.493	0.628
T3DIS[6]	0.535	0.463	0.542	0.816	0.455	0.581	0.518	0.352	0.315	0.536	0.286	0.336	0.474	0.572	0.562	0.965	0.507	0.745
Open-Vocabulary 3D Instance Segmentation																		
Open Inst3D[8]	0.252	0.220	0.403	0.582	0.473	0.312	0.305	0.305	0.421	0.001	0.257	0.128	0.370	0.142	0.003	0.385	0.129	0.058
Open Mask3D[7]	0.270	0.184	0.349	0.405	0.346	0.245	0.263	0.128	0.167	0.190	0.172	0.256	0.213	0.369	0.329	0.517	0.296	0.400
Ours (only 2d)	0.283	0.192	0.372	0.441	0.325	0.246	0.296	0.141	0.142	0.237	0.158	0.284	0.213	0.359	0.354	0.542	0.330	0.431
Ours (only 3d)	0.276	0.235	0.332	0.368	0.350	0.269	0.307	0.158	0.104	0.156	0.237	0.253	0.230	0.201	0.342	0.558	0.356	0.474
Ours (Ensemble)	0.293	0.225	0.364	0.438	0.327	0.246	0.317	0.149	0.151	0.235	0.158	0.287	0.214	0.344	0.373	0.546	0.385	0.499

사용한 경우(Only 2D)와 3차원 특징만을 사용한 경우(Only 3D Feature)들과도 성능을 비교해본다. Table 4의 실험 결과를 살펴보면, 본 논문의 제안 모델Open3DME가 기존의 개방형 어휘 개체 분할 모델들에 비해 전체 클래스에 대한 평균 분할 성능이 AP 면에서 가장 높은 성능을 보였으며, 17개의 분할 대상 클래스들 중 절반 이상 가장 높은 성능을 보여주었다. 일반적으로 추론 시에 훈련 단계에서 이미 학습한 분류 레이블들만을 개체 분할에 이용하는 폐쇄형 어휘 개체 분할 모델들은 추론 시에 훈련 단계에서는 주어지지 않은 새로운 분류 레이블들도 개체 분할에 이용해야 하는 개방형 어휘 개체 분할 모델들보다 대체적으로 성능이 높을 수 밖에 없다. 하지만 제안 모델은 Table 4에서 폐쇄형 어휘 개체 분할 모델들[5, 6]과 분할 성능을 비교하였을 때도, 이들과 큰 성능 차이를 보이지 않았음을 확인할 수 있었다. 구체적으로 살펴보면, 제안 모델은 케비넷(cabinet), 문(door), 액자(picture)를 포함한 8개의 클래스에 대해서 가장 높은 성능을 보였다. 반면, 침대(bed), 의자(chair)를 포함한 7개의 클래스에 대해서는 OpenIns3D[8]가 제안 모델보다 약간 높은 AP 성능을 보였다. 하지만 모든 개체 클래스에 대한 평균 성능은 제안 모델이 OpenMask3D[7], OpenIns3D[8]보다 각각 8.51%, 16.26%의 성능 향상률을 보였다. 이와 같은 실험 결과를 통해서, 제안 모델 Open3DME의 우수성을 확인할 수 있었다.

한편, 2차원 시각적 특징만을 사용한 경우(Only 2D), 3차원 시각적 특징만을 사용한 경우(Only 3D)의 제안 모델과도 분할 성능을 비교해보면, 2차원 시각적 특징만을 사용한 모델이 3차원 시각적 특징만 사용한 모델보다 낮은 정확도를 보였던 케비넷(cabinet), 소파(sofa), 테이블(table)을 포함한 10개의 클래스들에 대한 예측 성능이 앙상블 기법을 적용하였을 때 향상되었음을 확인할 수 있었다. 마찬가지로, 3차원 시각적 특징만을 사용한 모델이 부족했던 침대(bed), 의자(chair), 책꽂이(bookshelf)를 포함한 6개의 클래스에 대해서도 앙상블 기법이 적용된 특징을 사용한 제안 모델에서 성능 향상이 있었음을 확인할 수 있었다. 그러나 제안 모델은 문(door), 책상(desk), 샤워 커튼(s. curt) 등 5개의 클래스에 대해서는 단일 특징만을 사용한 모델들보다 높은 성능을 보였지만, 대부분의 클래스들은 단일 특징 사용 모델들의 중간 수준의 성능을 보여주었다. 이러한 결과를 통해서, 제안 모델이 2차원과 3차원 시각적 특징의 장점을 완전히 활용하지 못하는 한계점도 확인할 수 있었다.

마지막 실험은 몇 가지 개방형 어휘 개체 분할 사례들을 기초로, 제안 모델의 성능을 정성적으로 분석해보는 실험이다. 이 실험에서는 Fig. 6과 같은 서로 다른 3개의 장면 포인트 클라우드와 텍스트 프롬프트에 대해, 정답 개체 분할 결과뿐만 아니라, 비교 모델인 OpenMask3D[7]와 제안 모델의 개방형 어휘 개체 분할 결과들을 서로 비교해보았다. 이 실험에서는 텍스트 프롬프트와의 코사인 유사도가 높은 마스크일수록 붉은색에 가깝게, 유사도가 낮은 마스크일수록 푸른색에 가깝게 나타내었다.

Fig. 6의 첫 번째 사례는 마스크 제안 모듈로 T3DIS[6]를 채용한 제안 모델이 Mask3D[5]를 채용한 OpenMask3D[7]에 비하여 정답 결과에 가까운 분할 결과를 생성하였음을 확인할 수 있다. 제안 모델은 좀 더 정확한 마스크를 토대로, 입력 텍스트 쿼리 “a table in a scene”에 부합하는 장면 중앙의 탁자(table) 개체를 정확히 붉은색으로 분할한 것을 확인할 수 있다. 반면에, OpenMask3D[7]는 탁자를 두 개체로 나누어서 제안된 마스크들로 인해 개방형 어휘 개체 분할 결과도 두 개체로 나뉘어서 잘못 분할한 것을 확인할 수 있었다. 이와 같은 결과를 통해서, 제안 모델에서 채용한 T3DIS[6]의 마스크 품질 향상 효과를 확인할 수 있었다.

Fig. 6의 두 번째 사례는 3차원 마스크로 나뉘어진 포인트 세그먼트를 나타내는 시각적 특징으로 2차원과 3차원 특징을 앙상블하여 사용한 제안 모델이 2차원 특징만을 사용한 OpenMask3D[7]에 비하여 더 정확한 분할 결과를 생성하고, 텍스트 프롬프트와의 유사도를 좀 더 명확하게 구별해낸 사례

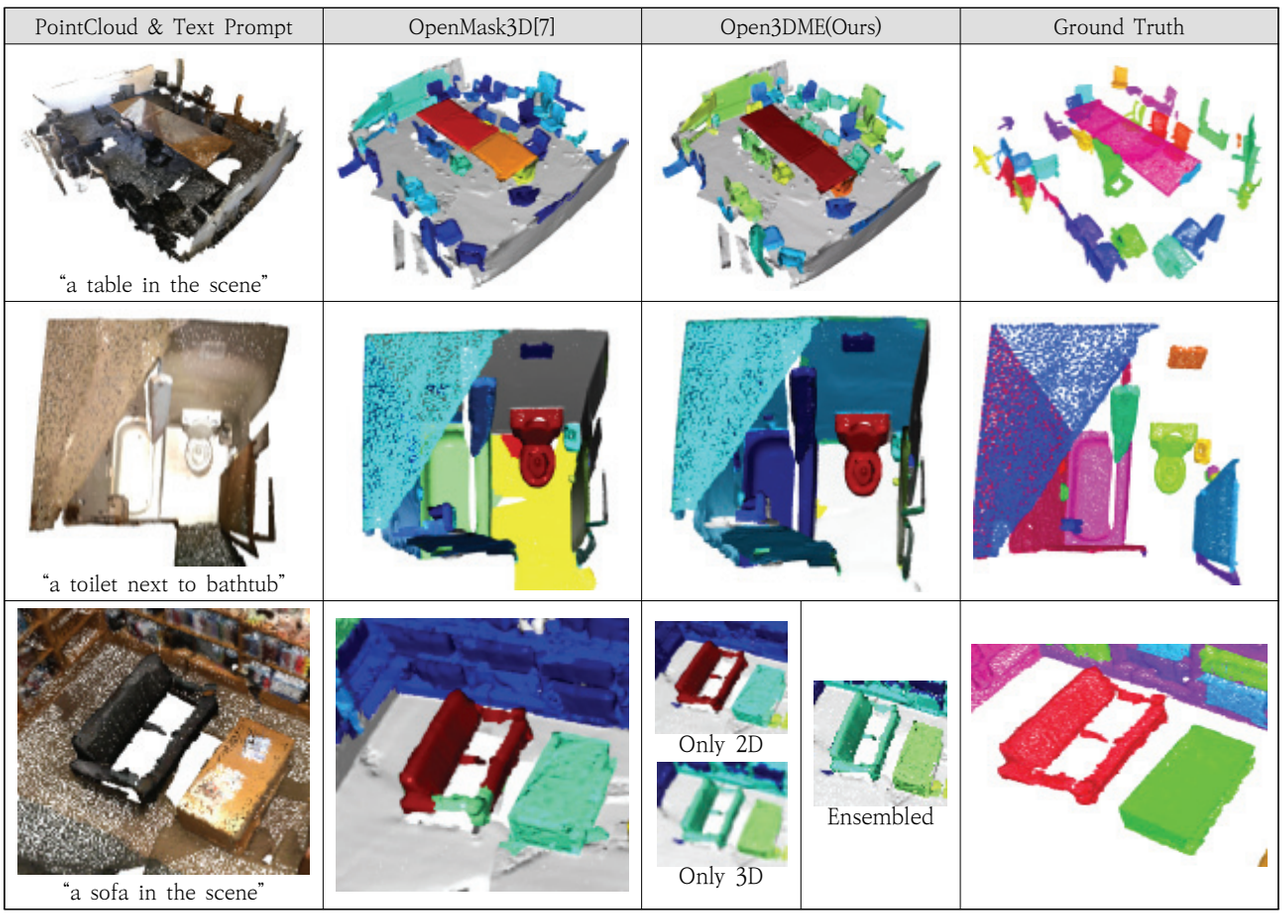

Fig. 6. Qualitative Evaluation of Open-Vocabulary 3D Instance Segmentation Results

이다. 구체적으로 설명하면, OpenMask3D[7]의 경우, 텍스트 프롬프트 "a toilet next to bathtub"에 부합하는 장면 우측 상단의 변기(toilet) 개체의 마스크를 유사도가 가장 높은 붉은색으로 분할하였지만, 욕조(bathtub)나 바닥(floor)에도 텍스트 프롬프트와 어느 정도 유사도를 가진다고 판단하여 푸른색이 아닌 녹색과 노란색으로 분할한 것을 확인할 수 있다. 반면에, 앙상블된 시각적 특징을 사용한 제안 모델은 텍스트 프롬프트에 부합되는 변기(toilet) 개체만을 붉은색으로 명확하게 분할한 것을 확인할 수 있었다. 이와 같은 결과들을 토대로, 제안 모델에서 적용한 2차원-3차원 특징 앙상블의 긍정적 효과를 확인할 수 있었다. 또한, 첫 번째 사례와 마찬가지로 제안 모델이 OpenMask3D[7]와 비교하여 변기 개체에 대한 3차원 마스크가 좀 더 정답 마스크에 가까운 것을 다시 한번 확인할 수 있었다.

Fig. 6의 세 번째 사례는 제안 모델이 OpenMask3D[7]에 비해, 오히려 품질이 낮은 분할 결과를 생성한 경우를 보여준다. 구체적으로 설명하면, 적응적 앙상블 기법을 적용한 제안 모델의 경우에는 입력 텍스트 프롬프트 "a sofa in the scene"에 부합하는 소파(sofa) 개체를 유사도가 가장 높은 마스크로 분할하지 못한 반면에, OpenMask3D[7]는 소파(sofa) 개체의 마스크를 텍스트 프롬프트와 유사도가 가장 높은 마스크로 분할한 것을 확인할 수 있다. 정확한 원인 파악을 위해 앙상블이 적용되기 전의 2차원 시각적 특징만을 사용한 경우(Only 2D), 3차원 시각적 특징만을 사용한 경우(Only 3D)의 분할 결과를 확인해본 결과, 3차원 시각적 특징만을 사용한 경우와는 달리, 2차원 시각적 특징만을 사용한 경우에서는 장면에 등장하는 소파 개체의 제안 마스크를 제대로 분할해내는 것을 확인할 수 있다. 이러한 결과를 통해서, 앙상블 과정에서 몇몇 클래스에 대해 3차원 특징의 유사도 점수에 좀 더 높은 가중치를 부여하는 적응적 앙상블 기법으로 인해 실제로는 2차원 시각적 특징보다 의미적으로 연관성이 낮은 3차원 시각적 특징이 사용되면서 오히려 분할 성능이 저하된 것을 확인할 수 있었다. 이와 같은 앙상블 방식이 각 특징의 장점을 잘 활용하지 못하는 한계점은 실험 4에서도 확인해볼 수 있었다. 이러한 문제 해결을 위해서 좀 더 유동적으로 각 특징 별 장점을 잘 살려낼 수 있는 방법에 대한 탐색과 연구가 필요해 보인다.

5. 결 론

본 논문에서는 새로운 개방형 어휘 3차원 개체 분할 모델인 Open3DME를 제안하였다. 제안 모델은 클래스-독립적인 3차원 마스크들의 정확성을 향상시키기 위해, 디코더의 쿼리

정제 방식 개선, 그리고 모델 학습 가속화를 위한 노이즈 제거 보조 학습이 적용된 트랜스포머 기반 3차원 개체 분할 모델인 T3DIS[6]를 마스크 제안 모듈로 채용하였다. 또한, 제안 모델은 각 포인트 클라우드 세그먼트별로 사전 학습된 OpenScene [20] 기반의 3차원 포인트 인코더를 적용해 텍스트와 의미적으로 정렬된 3차원 시각적 특징을 추출할뿐만 아니라, 멀티-뷰 RGB 영상 세그먼트들에도 사전 학습된 CLIP 2차원 픽셀 인코더를 적용하여 텍스트와 의미적으로 정렬된 2차원 시각적 특징을 추출하였다. 또 제안 모델 Open3DME는 개방형 어휘 레이블 텍스트 특징들과의 매칭에 각 포인트 클라우드 세그먼트별로 추출한 2차원과 3차원 시각적 특징을 상호 보완적으로 함께 이용하기 위한 특징 앙상블 기법도 적용하였다. 본 논문에서는 ScanNet-V2 벤치마크 데이터 집합을 이용한 다양한 정량적, 정성적 실험들을 통해, 제안 모델의 우수성을 입증하였다.

한편, 앞서 정량적 실험 4와 정성적 평가에서 확인했던 바와 같이, 적응적 앙상블 기법을 적용한 제안 모델이 실제로는 2차원 시각적 특징보다 텍스트 프롬프트에 대한 유사도가 훨씬 낮은 3차원 시각적 특징을 포인트 세그먼트를 나타내는 시각적 특징으로 사용하면서, 오히려 분할 성능이 저하되는 경우도 발생하는 것을 확인할 수 있었다. 따라서 이러한 현재 제안 모델 Open3DME의 한계점을 극복하기 위해, 2차원과 3차원 시각적 특징이 가지는 장점을 충분히 활용할 수 있는 앙상블 방식을 고안해보고, 추가로 주의 집중(attention) 메카니즘을 이용한 2차원과 3차원 시각적 특징 간의 교차 모달 특징 융합을 적용해보는 등 추가적인 성능 개선을 위한 향후 연구를 진행할 계획이다.

References

[1] B. Yang, J. Wang, R. Clark, Q. Hu, S. Wang, A. Markham, and N. Trigoni, "Learning object bounding boxes for 3D instance segmentation on point clouds," In *Proceedings of the Nueral Information Processing Systems (NeurIPS)*, 2019.

[2] S. Liu, S. Yu, S. Wu, H. Chen, and T. Liu, "Learning gaussian instance segmentation in point clouds," *arXiv preprint arXiv:2007.09860*, 2020.

[3] T. Vu, K. Kim, T. Luu, T. Nguyen, and C. D. Yoo, "Soft-Group for 3D instance segmentation on point clouds," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Z. Liang, Z. Li, S. Xu, M. Tan, and K. Jia, "Instance segmentation in 3D Scenes using semantic superpoint tree networks," In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[5] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3D: Mask transformer for 3D semantic instance segmentation," In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2023.

[6] S. Song and I. Kim, "T3DIS: Transformer-based 3D instance segmentation with auxiliary denoising learning," In *Proceedings of the Journal of Institue of Control, Robotics and Systems(J Inst Contr Robot Syst)*, Vol.29, No.12, pp, 954-965, 2023.

[7] A. Takmaz, E. Fedele, R. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "OpenMask3D: Open-vocabulary 3D instance segmentation," In *Proceedings of the Nueral Information Processing Systems (NeurIPS)*, 2023.

[8] Z. Huang, X. Wu, X. Chen, H. Zhao, L. Zhu, and J. Lasenby, "OpenIns3D: Snap and Lookup for 3D Open-vocabulary Instance Segmentation," *preprint arXiv:2309.00616*, 2023.

[9] S. Lu, H. Chang, E. Jing, A. Boularias, and K. Bekris, "OVIR-3D: Open-vocabulary 3D instance retrieval without training on 3D data," In *Proceedings of the Conference on Robot Learning(CoRL)*, 2023.

[10] R. Ding, J. Yang,, C. Xue, W. Zhang, S. Bai and X. Qi, "Lowis3D: Language-Driven Open-World Instance-Level 3D Scene Understanding," preprint arXiv:2308.00353, 2023.

[11] A. Radford et al., "Learning transferable visual models from natural language supervision," *preprint arXiv:2103.00020*, 2021.

[12] C. Jia et al., "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," In Proceedings of the International Conference on Machine Learning(ICML), 2021.

[13] G. Ghiasi, X. Gu, Y. Cui, and T. Lin, "Scaling Open-vocabulary image segmentation with image-level labels," In *Proceedings of the roceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[14] J. Ding, N. Xue, G. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted CLIP," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[16] J. Qin et al., "FreeSeg: Unified, universal and open-vocabulary image segmentation," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu, "SAM3D: Segment anything in 3D scenes," *arXiv preprint arXiv: 2306.03908*.

[18] A. Kirillov et al., "Segment Anything," In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[19] R. Chen et al., "CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "OpenScene: 3D Scene understanding with open vocabularies," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[21] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," In *Proceedings of the roceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[22] C. Choy, J. Gwak, and S. Savarese "4D Spatio-Temporal ConvNets: Minkowski convolutional neural networks," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] X. Wu, Y. Lao, L. Jiang, X. Liu, and H. Zhao, "Point Transformer V2: Grouped vector attention and partition-based pooling," In *Proceedings of the Nueral Information Processing Systems (NeurIPS)*, 2022.

송 성 호

https://orcid.org/0000-0003-3372-4737
e-mail : ssh10032@kyonggi.ac.kr
2022년 경기대 컴퓨터공학부(학사)
2022년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 기계학습, 3D 비전

박 경 민

https://orcid.org/0009-0006-9648-2391
e-mail : gkalsrudals@kyonggi.ac.kr
2024년 경기대 컴퓨터공학부(학사)
2024년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 컴퓨터비전, 3D 비전

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 AI컴퓨터공학부 교수
관심분야 : 인공지능, 기계학습, 컴퓨터비전