Journal of Institute of Control, Robotics and Systems (2025) 31(5):463-480
https://doi.org/10.5302/J.ICROS.2025.25.0018 ISSN:1976-5622 eISSN:2233-4335

거대 언어 모델 기반의 의미적 텍스트 분리 및 교차 어텐션 기반의 양방향 멀티모달 특징 융합을 활용한 일반화된 3차원 참조 표현 분할

Large Language Model-based Semantic Text Decoupling and Cross Attention-based Bidirectional Multimodal Feature Fusion for Generalized Three-dimensional Referring Expression Segmentation

배 혜 림, 박 경 민, 김 인 철*
(HyeLim Bae[1], Kyungmin Park[1], and Incheol Kim[1,*])
[1]Department of Computer Science Kyonggi University

Abstract: 3D referring expression segmentation (3D-RES) locates target objects within a 3D scene point cloud with precise 3D masks based on natural language descriptions. We propose a new benchmark dataset, OAR3DRef, and a baseline model, TDMF, to overcome the limitations of current research in generalized 3D RES. OAR3DRef provides detailed annotation of referring expressions with rich semantic components, helping a 3D-RES model in understanding the correct meaning of the expressions. TDMF, which is a baseline deep neural network model proposed for effective use of OAR3DRef, utilizes both a linguistic parser and a large language model to decouple the given natural language expression into multiple semantic components. The model also adopts a novel semantic alignment loss function to align decoupled text features with object visual features. It performs cross attention-based bidirectional feature fusion between expression text features and object visual features before their semantic alignment. Furthermore, the model makes use of additional 2D object visual features extracted from multiview RGB-D scene images to enhance exact recognition of object attributes such as color, material, and appearance. We demonstrate the superiority of the proposed baseline model, TDMF, through various quantitative and qualitative experiments using the new OAR3DRef benchmark dataset.

Keywords: generalized 3D referring expression segmentation, scene point cloud, multi-view RGB images, large language model, semantic text decoupling, multimodal feature fusion, semantic alignment loss

I. 서론

실세계 및 가상 환경에서 자율적으로 동작해야 하는 서비스 로봇(service robot), 자율 주행 자동차(autonomous car), 게임 캐릭터(non-player character, NPC) 등과 같은 자율 에이전트들에게는 주변 환경에 대한 3차원 장면 이해(3D scene understanding) 능력이 필수적으로 요구된다. 3차원 장면 이해를 위한 다양한 지능형 시각 인식 작업 중에서 특히 3차원 시각적 그라운딩(3D visual grounding, 3D VG)은 Fig. 1의 예와 같이 주변 환경을 나타내는 3차원 장면 포인트 클라우드(scene point cloud)에서 자유 형식의 자연어(free-form natural language)로 묘사된 특정 목표 물체(target object)를 시각적으로 찾아내는 작업이다. 한편, 장면 포인트 클라우드 내의 목표 물체 영역을 3차원 경계 상자(bounding box)로 표시*하는 작업은 특별히 3차원 참조 표현 이해(3D referring expression comprehension, 3D REC) [4]라고 부르고, 반면에 목표 물체 영역을 보다 정밀한 3차원 마스크(mask)로 표시하는 작업은 3차원 참조 표현 분할(3D referring expression segmentation, 3D RES) [5]이라고 부른다. 본 논문에서는 3차원 참조 표현 분할(3D RES) 연구를 위한 새로운 데이터 집합과 베이스라인 모델을 제안한다.

3차원 참조 표현 이해(3D REC)와 참조 표현 분할(3D RES)과 같은 3차원 시각적 그라운딩(3D VG) 작업은 모두 목표 물체를 묘사하는 자연어 참조 표현의 의미를 언어적으로 정확히 이해하는 능력, 장면을 구성하고 있는 다양한 환경 물체들을 입력 포인트 클라우드에서 시각적으로 검출낼 수 있는 능력, 그리고 자연어 참조 표현과 장면 포인트 클라우드 내의 시각적 물체 간의 의미적 정렬을 통해 목표 물체를 정확히 가려내는 능력 등이 복합적으로 요구되는 매우 난이도가 높은 지능 작업이다. 특히 픽셀(pixel)들이 구조적으로 균일하게 잘 배

*Corresponding Author
Manuscript received February 2, 2025; revised March 24, 2025; accepted April 7, 2025
배혜림: 경기대학교 일반대학원 컴퓨터과학과 석사과정(thvk654@kyonggi.ac.kr, ORCID 0000-0003-4179-0339)
박경민: 경기대학교 일반대학원 컴퓨터과학과 석사과정(gkalsrudals@kyonggi.ac.kr, ORCID 0000-0006-9648-2391)
김인철: 경기대학교 AI컴퓨터공학부 교수(kic@kyonggi.ac.kr, ORCID 0000-0002-5754-133X)
※ This work was supported by Kyonggi University's Graduate Research Assistantship 2024.

치된 한 장의 2차원 장면 영상(scene image)에서 목표 물체를 찾아내는 기존의 2차원 시각적 그라운딩(2D VG) 작업에 비해, 3차원 공간상에 비-균일하게 포인트(point)들이 분포하는 3차원 장면 포인트 클라우드(scene point cloud)에서 목표 물체를 찾아내야 하는 3차원 시각적 그라운딩(3D VG) 작업은 더 많은 어려움을 내포하고 있다.

Fig. 1의 예에서 보듯이, 자연어 참조 표현들 중에는 목표 물체(target object)를 직접 가리키는 desk와 같은 목표 물체 단어 이외에, 목표 물체의 색상과 텍스처를 나타내는 brown, wooden 같은 목표 물체 속성 단어들, 목표 물체의 식별을 도와주기 위한 보조 물체(auxiliary object)를 나타내는 shelf와 같은 보조 물체 단어, 보조 물체의 색상을 나타내는 gray와 같은 보조 물체 속성 단어, 그리고 목표 물체와 보조 물체 간의 관계를 나타내는 positioned to the right of와 같은 관계 단어들을 함께 포함하는 경우가 많다. 그뿐만 아니라 Fig. 1의 예와 같이, 자연어 참조 표현에 부합하는 목표 물체 desk들이 장면 포인트 클라우드 안에 여러 개 존재하는 경우(multiple targets) 혹은 하나도 존재하지 않은 경우(zero target)도 종종 있다. 이와 같이 다양한 3차원 참조 표현 분할 작업들을 효과적으로 수행할 수 있는 심층 신경망(deep neural network) 모델을 설계하기 위해서는 고려해야 할 몇 가지 중요한 설계 이슈(design issue)들이 존재한다.

첫 번째 이슈는 복잡한 자연어 참조 표현의 의미를 심층적으로 이해하기 위해, 참조 표현 문장을 어느 수준까지 언어적 구성들로 분리(decoupling)할 것인가 하는 것이다. 이와 연관된 두 번째 이슈는 자연어 참조 표현을 언어적 구성 요소들로 분리한다면, 이렇게 분리된 참조 표현(decoupled referring expression)을 모델 학습에 어떻게 활용할 것인가 하는 것이다. Fig. 1의 모델 구조에서 보듯이, TGNN [1]과 같은 기존 모델들에서는 자연어 참조 표현에 대한 언어적 분리가 전혀 이루어지지 않은 채 구성 단어들을 임베딩하여 텍스트 특징을 추출하였다. 한편, 3D-STMN [2]과 같은 기존 모델들은 자연어 참조 표현에 대한 구성 단어들 간의 의존성 분석(dependency analysis)을 수행하여 의존성 그래프를 도출하기는 하였으나, 각 단어가 목표 물체 혹은 보조 물체를 나타내는지 목표 물체나 보조 물체의 속성 중 하나를 나타내는지 물체 간의 관계를 나타내는지 등을 명확히 분석해내지 못했다. 한편, MDIN [3]과 같은 기존 모델들은 언어적 구성 요소들로 자연어 참조 표현에 대한 분리가 이루어졌다. 하지만 이러한 언어적 구성 요소들에 대한 역할 분석은 대부분 목표 물체(target object) 위주로만 수행됨으로써, 목표 물체의 식별을 도와주는 중요한 역할을 수행하는 보조 물체들(auxiliary objects)의 속성들(attributes)과 물체들 간의 관계(relationship between objects)들을 나타내는 단어들에 대한 분석과 분리는 이루어지지 못했다.

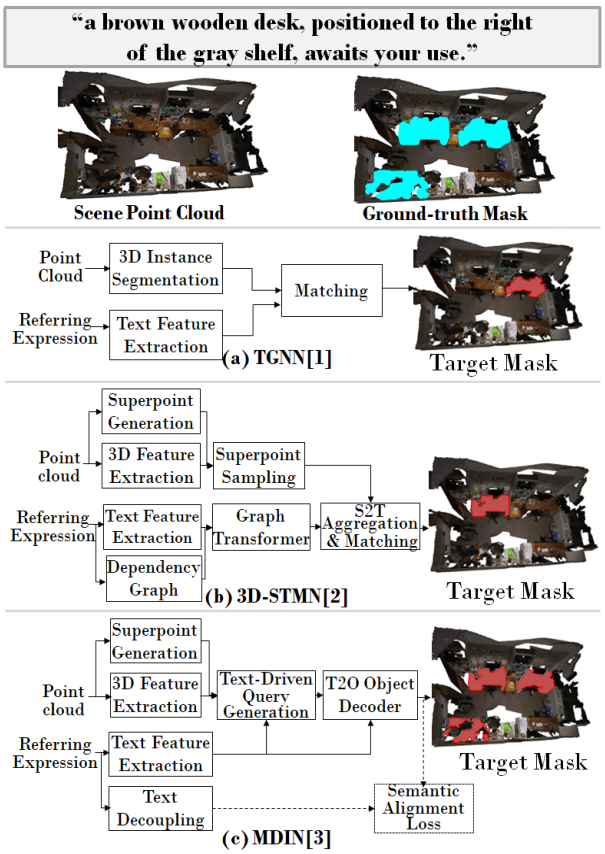


그림 1. 3차원 참조 표현 분할을 위한 기존 모델들.

Fig. 1. Existing models for 3d referring expression segmentation.

자연어 참조 표현에 대한 언어적 분리가 이루어지지 않은 TGNN [1]과 같은 기존 모델들은 두 번째 설계 이슈에 해당사항이 없지만, 자연어 참조 표현에 대한 언어적 분리가 부분적으로 이루어진 3D-STMN [2], MDIN [3]과 같은 기존 모델들은 나름의 방식으로 분리된 자연어 참조 표현을 모델 학습에 이용하였다. Fig. 1에서 보듯이, 3D-STMN [2] 모델의 경우는 물체들의 시각적 특징과 매칭을 수행할 텍스트 특징 추출에 분리된 참조 표현을 이용하였고, MDIN [3] 모델의 경우는 물체들의 특징과 의미적 정렬 손실(semantic alignment loss)을 계산할 때 분리된 참조 표현의 각 단어의 역할에 따라 차등적으로 손실에 반영하였다. 하지만 앞서 설명한 대로, MDIN [3]은 보조 물체들의 속성과 관계들까지 심층적으로 분리가 이루어지지 않음으로써, 의미적 정렬 손실 계산에도 역시 자연어 참조 표현에 대한 이해가 충분히 반영될 수 없다는 한계점이 있다.

3차원 참조 표현 분할 모델 설계 시에서 고려해야 할 세 번째 중요한 이슈는 포인트 클라우드에 포함된 물체들의 시각적 특징과 자연어 참조 표현의 텍스트 특징 간의 특징 융합(vision-text feature fusion)을 어떻게 수행할 것인가 하는 것이다. 3차원 참조 표현 분할 작업에서 목표 물체 결정을 위해서는 물체들의 시각적 특징과 자연어 참조 표현의 텍스트 특징 간 매칭 혹은 정렬이 필수적이다. 그런데 이들 간의 매칭과 정렬을 효과적으로 수행하기 위해 이에 앞서 서로 모달이 다른 이질적인 두 특징 간의 특징 융합을 미리 수행하는 기법들이 3D-STMN [2], MDIN [3]과 같은 기존 모델들에서도 적용되어 왔다. 하지만 기존 모델들이 수행한 특징 융합 기법들은 모두 물체의 시각적 특징에서 참조 표현 텍스트 특징으로(vision-to-text), 혹은 그 반대 방향(text-to-vision)으로만 특징을 융합하는 단방향 특징 융합(uni-directional feature fusion)을 시도하였다. 하지만 두 이질적인 특징 간의 효과적인 매칭과 정렬을 위해서는 서로 간의 양방향 특징 융합(bi-directional feature fusion)이 이루어지는 것이 더 바람직할 것이다.

3차원 참조 표현 분할 모델 설계를 위해 고려해야 할 네 번째 이슈로는 입력 데이터로 장면 포인트 클라우드 외에 해당 장면을 나타내는 멀티-뷰 RGB-D 영상들(multi-view RGB-D images)의 활용에 관한 것이다. Fig. 1의 예와 같이, 많은 자연어 참조 표현들은 목표 물체 및 보조 물체들의 색상(color), 텍스쳐(texture), 모양(shape), 크기(size) 등을 묘사하는 다양한 속성 단어들을 포함하고 있다. 하지만 모든 기존의 3차원 참조 표현 분할을 위한 벤치마크 데이터 집합들과 인식 모델들은 이러한 멀티-뷰 RGB-D 영상 데이터를 제공하지도 않고 이들을 모델 학습과 추론에 활용하지도 않았다.

마지막으로 고려해야 할 모델 설계 이슈는 목표 물체가 존재하지 않는 제로 목표(zero target)와 다중 목표 물체들(multiple targets)까지 포함하는 일반화된 3차원 참조 표현 분할(generalized 3D referring expression, 3D GRES) 작업에 대한 지원 여부와 구현 방법이다. MDIN [3]을 제외한 거의 모든 기존 모델들은 장면 포인트 클라우드 내에는 반드시 단 하나의 목표 물체만 존재한다고 가정(single target assumption)하고, 포인트 클라우드에서 검출해낸 물체들 중 자연어 참조 표현에 상대적으로 가장 잘 부합되는 물체를 무조건 목표 물체로 결정한다. 하지만 이러한 제한된 가정에 기초한 기존 모델들은 자연어 참조 표현이 묘사하는 목표 물체가 해당 장면에는 없거나 참조 표현에 부합하는 목표 물체가 여러 개 존재할 수 있는 현실 세계의 응용 분야들에는 적용하는 데 한계가 있다. 현실 세계의 폭넓은 적용 가능성을 고려한다면, 일반화된 3차원 참조 표현 분할(3D GRES) 기능을 제공할 수 있도록 모델을 설계하는 것이 바람직할 것이다.

이와 같이 모델 설계 이슈별로 기존 연구들의 한계성을 살펴보았는데, 본 논문에서는 기존 모델들의 한계성을 극복하고자 일반화된 3차원 참조 표현 분할을 위한 새로운 벤치마크 데이터 집합 **OAR3DRef (Object Attributes and Relationships for generalized 3D Referring expression segmentation)와 이 데이터 집합을 효과적으로 이용할 수 있는 새로운 베이스라인 심층 신경망 모델인 TDMF (Text Decoupling and Multimodal Fusion)를 제안한다. 새롭게 제안하는 벤치마크 데이터 집합과 베이스라인 모델은 앞서 설명한 자연어 참조 표현의 의미적 분리와 활용 이슈들과 연관해서는 (1) 보조 물체들의 속성과 관계들을 나타내는 단어들까지 참조 표현에서 심층적으로 분리해낼 수 있도록 풍부한 주석 데이터들을 포함하는 새로운 형식의 데이터 집합을 제공할 뿐만 아니라, 복잡한 자연어 참조 표현의 효과적인 분리를 위해 언어적 분석기(linguistic parser)와 더불어 거대 언어 모델(larger language model, LLM)을 함께 상호 보완적으로 활용한다. 또한 (2) 제안 모델에서는 이렇게 심층적으로 분리된 자연어 참조 표현을 모델 학습에 효과적으로 이용하기 위해, 포인트 클라우드에서 검출한 시각적 물체들과의 의미적 정렬을 수행하는 새로운 의미적 정렬 손실 함수를 정의한다. 또한 (3) 제안 모델은 물체들의 시각적 특징과 자연어 참조 표현의 텍스트 특징 간 특징 융합 이슈와 관련해서는, 특징 융합의 효과를 향상시키기 위해 단방향의 특징 융합만을 수행한 기존 모델들과는 달리 텍스트 특징에서 시각적 특징으로 또 그 반대 방향으로의 융합이 함께 수행되는 교차 어텐션(cross-attention) 기반의 양방향 특징 융합을 수행한다. 또 (4) 제안 모델은 장면 포인트 클라우드에서 추출하는 3차원 기하학적 특징만을 이용한 기존 모델들과는 달리, 동일 장면을 나타내는 멀티-뷰 RGB-D 영상들에서 추출하는 2차원 시각적 특징들도 물체들의 특징에 함께 포함시킴으로써, 다양한 색상, 재질, 크기를 가진 목표 및 보조 물체에 대한 식별과 분할 능력을 향상시킨다. (5) 마지막 설계 이슈와 관련해서, 제안 모델은 단일 목표 물체 가정에서 벗어나서 제로 목표(zero target) 및 다중 목표 물체들(multiple targets)도 찾아낼 수 있는 예측과 추론 기능을 제공한다. 본 논문에서는 새로운 벤치마크 데이터 집합인 OAR3DRef를 이용한 다양한 정량적, 정성적 실험들을 통해, 새롭게 제안하는 베이스라인 모델 TDMF의 우수성을 입증한다.

본 논문의 2장에서는 3차원 참조 표현 분할과 관련한 기존 연구들에 관해 살펴보고, 3장에서는 본 논문에서 제안하는 새로운 벤치마크 데이터 집합에 대해 설명하고, 이어서 4장에서는 이 데이터 집합을 효과적으로 활용할 수 있는 베이스라인 모델 TDMF의 설계에 대해 자세히 설명한다. 5장에서는 제안하는 베이스라인 모델의 TDMF 세부 구현과 다양한 성능 실험 결과들을, 마지막으로 6장에서는 결론 및 향후 연구에 관하여 정리한다.

II. 관련 연구

1. 3차원 참조 표현 분할 작업과 데이터 집합

3차원 참조 표현 이해(3D REC)와 3차원 참조 표현 분할(3D RES) 작업을 위한 대표적인 벤치마크 데이터 집합으로 ScanRefer [6]와 ReferIt3D [7]가 존재한다. ScanRefer [6] 데이터 집합은 ScanNet [8] 데이터 집합에서 추출한 장면 포인트 클라우드들을 기반으로, 11,046개의 단일 목표 물체에 대한 51,538개의 자연어 참조 표현 문장들과 더불어 각 장면 포인트 클라우드 내 목표 물체의 정답 경계 상자(ground-truth bounding box)를 함께 제공하고 있다. 한편, ReferIt3D [7] 데이터 집합 역시 ScanNet [8] 장면 포인트 클라우드들을 기반으로, 76개의 목표 물체들에 대한 자연어 참조 표현 문장들을 포함하고 있다. 이때 참조 표현들은 템플릿을 이용한 자동 생성 방식과 2인 참조 대화 게임을 통한 생성 방식 등 2가지 방식으로 확보하였다. 3차원 참조 표현 이해(3D REC) 작업을 수행하는 대표적인 모델인 EDA [4]는 ScanRefer [6]와 ReferIt3D [7] 데이터 집합을 이용하였다. 하지만, 3차원 참조 표현 이해 작업은 목표 물체의 영역을 3차원 경계 상자만으로 표시함으로써, 장면 포인트 클라우드 내 목표 물체의 정확한 영역을 알아내기 어렵다는 문제가 존재한다. 이를 극복하기 위해 목표 물체의 정밀한 3차원 마스크를 찾는 3차원 참조 표현 분할(3D RES) 작업이 등장하였다.

앞서 언급한 ScanRefer [6]와 ReferIt3 [7] 데이터 집합을 3차원 참조 표현 분할(3D RES) 작업에 이용하기 위해 초기 연구들에서는 목표 물체의 3차원 정답 경계 상자 대신 정답 마스크를 ScanNet [8] 데이터 집합에서 따로 추출하여 이용하였다.

** https://github.com/wooca12/OAR3DRef-dataset

대표적으로 [1,2,5,9,10]와 같은 기존 모델들은 정답 마스크를 갖는 ScanRefer [6] 데이터 집합을, [1,5]과 같은 기존 모델들은 정답 마스크를 갖는 ReferI3D [7] 데이터 집합을 사용하였다. 하지만, 이 두 데이터 집합들에 기초한 기존의 모델들은 모두 단일 목표(single target)를 찾는 작업들만 처리할 수 있어, 목표 물체가 없는 경우(zero target)나 다수인 경우(multiple targets)를 포함하는 다양한 현실세계 응용 분야들에 이용할 수 없다는 한계가 존재하였다. 이와 같은 문제를 해결하고자, 목표 물체의 개수에 제한을 받지 않는 다양한 참조 표현 분할 작업이 가능한 일반화된 3차원 참조 표현 이해(generalized 3D referring expression comprehension, 3D GREC) 작업 및 일반화된 참조 표현 분할(generalized 3D referring segmentation, 3D GRES) 작업이 등장하였다.

일반화된 3차원 참조 표현 이해(3D GREC) 작업을 위한 대표적인 데이터 집합으로는 Multi3DRefer [11]가 존재한다. Multi3DRefer [11] 데이터 집합은 ScanRefer [6] 데이터 집합을 확장한 61,926개의 자연어 참조 표현들을 포함하고 있다. 해당 데이터 집합에는 다양한 개수의 목표 물체를 갖는 참조 표현 작업들을 포함하고 있으며, 장면 포인트 클라우드 내 각 목표 물체들의 정답 경계 상자를 함께 제공하고 있다. 하지만 3차원 참조 표현 이해 작업 역시 3차원 경계 상자로는 목표 물체들의 정확한 영역을 표현하기 어려워, 최근에는 목표 물체들의 3차원 마스크를 찾는 일반화된 참조 표현 분할(3D GRES) 작업에 관한 연구가 활발하다.

일반화된 참조 표현 분할(3D GRES) 작업을 위한 기존의 데이터 집합으로는 Multi3DRES가 존재한다. Multi3DRES 데이터 집합은 앞서 설명한 Multi3DRefer [11] 데이터 집합을 기반으로 하되 목표 물체들의 정답 3차원 경계 상자 대신 정답 3차원 마스크를 제공한다. 이러한 Multi3DRES 데이터 집합을 이용해 일반화된 참조 표현 분할(3D GRES) 작업을 수행하는 모델로는 MDIN [3]이 새롭게 소개되었다. 하지만 Multi3DRefer [11]와 Multi3DRES 두 데이터 집합 모두 자연어 참조 표현 문장에 포함된 목표 및 보조 물체들의 속성 단어들과 물체들 간의 관계 단어들에 대한 심층적 정보를 제공하지 못함으로써, 이 데이터 집합들에 기초해 학습한 모델들은 자연어 참조 표현에 부합하는 물체들을 장면 포인트 클라우드에서 시각적으로 명확히 인식해내는데는 한계가 있다.

2. 3차원 참조 표현 분할 모델들

현재까지 발표된 대표적인 3차원 참조 표현 분할 모델로는 TGNN [1], 3D-STMN [2], MDIN [3], 3DRefTR [5], RefMask3D [9], X-RefSeg3D [10] 모델 등이 있다. 이와 같은 기존 모델들은 자연어 참조 표현에 대한 의미적 텍스트 분리(semantic text decoupling)의 관점에서 다음과 같이 나누어 볼 수 있다. TGNN [1], Transfer3d [12], InstaceRefer [13] 모델은 특별히 참조 표현 문장에 대한 의미적 분리를 진행하지 않고 다만 참조 표현 단어들을 임베딩하여 텍스트 특징을 추출한다. 반면에, 3D-STMN [2], MDIN [3], 3DRefTR [5], RefMask3D [9], X-RefSeg3D [10] 모델들은 참조 표현 문장의 의미적 텍스트 분리를 수행하였다. 하지만 3D-STMN [2] 모델의 경우에는 각 단어의 역할에 대한 분류가 명확하게 이루어지지 않으며, X-RefSeg3D [9], RefMask3D [8], FFL-3DOG [14] 모델들은 문장 구조 및 단어 간 관계에 대한 분석이 미흡하다. MDIN [3], 3DRefTR [5] 모델들은 참조 표현에서 목표 물체의 명칭과 속성 등을 나타내는 목표 물체 위주의 단어들로만 텍스트 분리가 이루어졌다. 또한 다른 모델들과는 달리, MDIN [3]과 3DRefTR [5] 모델들은 이렇게 분리된 참조 표현의 텍스트 특징을 모델 학습에 활용하기 위해, 물체들의 시각적 특징들과의 의미적 정렬 손실 계산에 이용하였다. 하지만 목표 물체 위주의 단어들로만 참조 표현 텍스트를 분리하고 의미적 정렬 손실 계산 또한 목표 물체 위주로만 범위를 한정함으로써, 다수의 보조 물체들을 포함한 구조가 복잡한 자연어 참조 표현을 모델이 심층적으로 이해하도록 학습하는 데는 한계가 있다.

기존 모델들은 장면 포인트 클라우드로부터 각 환경 물체의 시각적 특징(object visual feature)을 추출하는 방식에도 차이가 있다. TGNN [1], X-RefSeg3D [10] 모델들의 경우, 별도의 3차원 개체 분할(3D instance segmentation) [15] 모듈을 적용하여 포인트 클라우드에 포함된 물체들을 먼저 검출해낸 다음, 이들로부터 자연어 참조 표현과 매칭할 3차원 시각적 특징을 추출해내는 방식을 적용하였다. 이 방식은 채용하는 개체 분할 모듈의 정확도에 따라 각 물체의 3차원 시각적 특징이 크게 달라질 수 있다는 문제가 있다. 한편, RefMask3D [9] 모델은 TGNN과는 달리, 포인트별 3차원 시각적 특징을 추출하는 다계층 U-Net 구조에서 서로 이웃한 포인트들을 묶어 계층화된 포인트 그룹(point group)들을 형성한 다음, 이 포인트 그룹들의 3차원 시각적 특징을 기초로 더 큰 단위인 물체별 3차원 시각적 특징을 디코딩하는 방식을 취했다. 반면에 3D-STMN [2], MDIN [3], 3DRefTR [5] 와 같은 기존 모델들에서는 포인트 클라우드 내의 해당 영역들을 정확히 알지 못하는 물체들의 3차원 특징을 곧바로 디코딩했을 때의 오류와 비효율성을 줄이기 위해, 포인트들을 먼저 수퍼포인트(superpoint) 단위로 묶어서 수퍼포인트별 3차원 시각적 특징을 구한 다음, 이들로부터 더 큰 단위인 물체별 3차원 시각적 특징을 디코딩하는 방식을 취했다. 이 두 방식 간에 차이는 있지만, 포인트 클라우드를 구성하는 최소 단위인 각 포인트의 3차원 시각적 특징들로부터 곧바로 가장 큰 단위인 각 물체의 3차원 시각적 특징을 디코딩하지 않고 중간 단위인 포인트 그룹이나 수퍼포인트들을 거쳐서 각 물체의 3차원 시각적 특징을 추출함으로써, 보다 정확한 물체별 특징을 얻을 수 있었다. 하지만 언급한 거의 모든 기존 모델들은 자연어 참조 표현에 등장하는 다양한 목표 물체 혹은 보조 물체들의 색상, 텍스쳐 등을 보다 정확히 인식해내기 위해 RGB-D 장면 영상들과 2차원 시각적 특징을 활용해보려는 시도는 없었다.

기존 모델들은 포인트 클라우드에서 추출하는 물체들의 시각적 특징과 자연어 참조 표현의 텍스트 특징을 정렬하기 이전에 미리 서로 융합(multimodal feature fusion)하는 방법에도 차이가 있었다. TGNN [1], MDIN [3], RefMask3D [9] X-RefSeg3D [10] 모델들은 물체들의 시각적 특징에 참조 표현의 텍스트 특징을 단방향으로 융합(uni-directional text-to-vision feature fusion)하였고, 반면에 3D-STMN [2] 모델은 이와는 반대로 참조 표현의 텍스트 특징에 물체들의 시각적 특징을 단방향으로 융합(uni-directional vision-to-text feature fusion)하였다. 또 3DRefTR [5] 모델의 경우에는 각 포인트별 시각적 특

징과 참조 표현 단어별 텍스트 특징 간에 양방향 특징 융합을 수행하지만, 정작 물체별 특징을 디코딩하는 단계에서는 물체들의 시각적 특징에 참조 표현의 텍스트 특징을 단방향으로만 융합(uni-directional text-to-vision feature fusion)하였다. 이와 같은 기존 모델들의 단방향 특징 융합은 다른 한쪽의 이질적인 정보를 각 측에 충분히 반영하기 어렵기 때문에, 이후 진행될 두 특징 간의 정렬 과정에서 높은 효과를 얻기 어렵다는 문제가 있다.

III. 물체 속성과 관계 중심의 벤치마크 데이터 집합

1. 자연어 참조 표현의 의미적 분리

과거 연구들에서는 3차원 참조 표현 이해 및 분할 작업을 위한 다양한 데이터 집합을 이용하였다. 대표적으로 ScanRefer [6]는 3차원 참조 표현 이해(3D REC) 작업을 위한 데이터 집합으로서, Fig. 2의 예에서 보듯이 각 작업별 자연어 참조 표현 문장과 더불어 목표 물체의 정답 경계 상자(ground-truth bounding box)를 제공하고 있다. 하지만 이 데이터 집합은 장면 포인트 클라우드 내에 반드시 목표 물체가 하나만 존재한다는 가정을 기초로 하고 있기 때문에, 참조 표현에 해당하는 목표 물체가 존재하지 않거나(zero target) 여러 개 존재할 경우(multiple targets)들을 다루지 못한다. 그뿐만 아니라, 장면 포인트 클라우드 내의 목표 물체 영역을 3차원 마스크가 아닌 3차원 경계 상자로만 표현하고 있다는 문제점도 있다. 하지만, 무엇보다 가장 큰 문제점은 ScanRefer [6] 데이터 집합에서는 Fig. 2의 예와 같이 복잡한 자유 형식의 자연어 참조 표현을 구성하는 언어적 구성 요소들에 대해 별도로 분리된 주석 데이터는 일절 제공하지 않는다는 것이다. 따라서 이러한 학습 및 검증 데이터에 기초한 모델들은 다양한 속성을 가진 목표 물체와 주변 보조 물체들, 그리고 이들 간의 관계들을 묘사하는 복잡한 자연어 참조 표현을 심층적으로 이해하기는 어렵다.

한편, 목표 물체의 개수에 제한 없이 다양한 참조 표현 작업이 가능한 일반화된 3차원 참조 표현 분할(3D GRES) 작업을 위한 데이터 집합으로는 최근 Multi3DRES 데이터 집합이 소개되었다. Multi3DRES 데이터 집합은 단일 목표 물체뿐만 아니라, 목표 물체가 존재하지 않거나 여러 개가 존재하는 경우의 참조 표현을 다루고 있다.

참조 표현에 해당하는 목표 물체의 개수와 목표 물체를 식별해내기 어렵게 하는 방해물(distractor)이 있느냐 없느냐에 따라 참조 표현의 유형을 방해물이 있는 제로 목표(zero target with distractors), 방해물이 없는 제로 목표(zero target without distractors), 방해물이 있는 단일 목표(single target with distractors), 방해물이 없는 단일 목표(single target without distractors), 다중 목표(multiple targets) 5가지로 분류하여 제공하고 있다. 또한, Fig. 2의 예와 같이, 자연어 참조 표현 문장과 함께 목표 물체의 정답 마스크(ground-truth mask)을 포함하고 있다. 구체적인 데이터 표현을 살펴보면, 자연어 참조 표현이 공간(spatial), 색상(color), 텍스쳐(texture), 모양(shape)에 대한 단어를 포함하고 있는지를 참/거짓(true/false)로 제공하고 있다. 하지만 구체적으로 자연어 참조 표현 내 어떤 단어들이 목표 물체 혹은 보조 물체의 속성들을 나타내는지, 어떤 단어들이 목표-보조 혹은 보조-보조 물체들 간의 관계를 표현하는지 등에 관한 구체적인 정보들을 일절 제공하지 않는다. 따라서 이러한 데이터에 기초한 모델들은 자연어 참조 표현에 등장하는 목표 물체 및 보조 물체들의 속성 단어들에 의미적으로 정확히 부합하는 물체들을 장면 포인트 클라우드에서 찾아내는 데는 한계가 있다.

이와 같은 한계점들을 개선하고자, 본 논문에서는 새로운 데이터 집합인 OAR3DRef를 제안한다. Fig. 3은 본 논문에서 제안하는 OAR3DRef의 데이터 표현을 나타낸다. Table. 1은 기존의 ScanRefer와 Multi3DRES, 그리고 새로운 OAR3DRef 데이터 집합들 간의 차이점을 한눈에 파악할 수 있도록 작성한 비교표이다. Tabel 1과 같이, 목표 및 보조 물체를 구분하지 않으며 물체마다의 구체적인 속성 정보를 제공하지 않는 ScanRefer와 Multi3DRES 데이터 집합과는 달리, OAR3DRef 데이터 집합에서는 Fig. 3과 같이 먼저 참조 표현 내의 목표

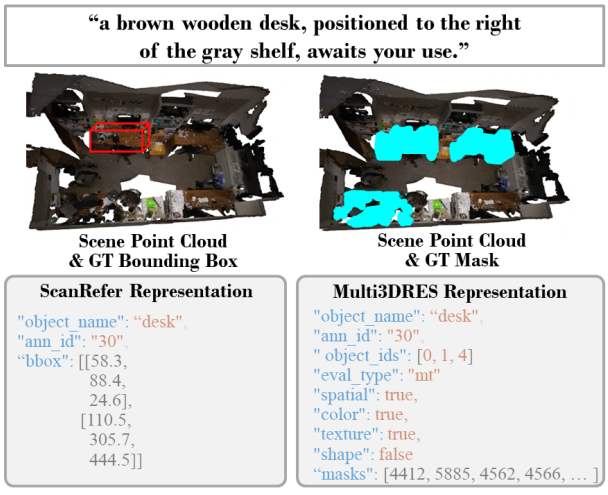


그림 2. ScanRefer과 Multi3DRES 데이터 집합.

Fig. 2. ScanRefer and Multi3DRES datasets.

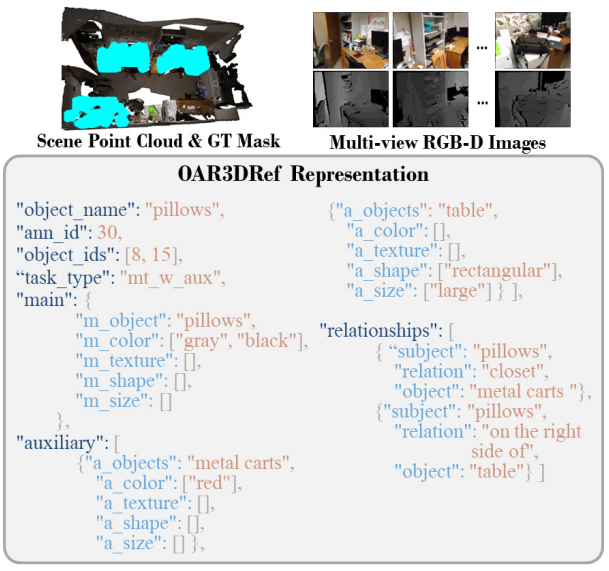


그림 3. OAR3DRef 데이터 집합.

Fig. 3. OAR3DRef dataset.

표 1. 세 가지 서로 다른 데이터 집합들 간의 비교.

Table 1. Comparison among three different datasets.

Dataset	ScanRefer[6]	Multi3DRES[3]	OAR3DRef(ours)
3D Target GT Mask	X	O	O
RGB-D Images	X	X	O
Target & Auxiliary Objects Words	X	X	O
Target & Auxiliary Attributes Words	X	provided only true / false	O
Relationships Words	X	provided only true / false	O
Subject-Object Words	X	X	O

물체와 보조 물체를 묘사하는 단어들을 서로 구분하여 제공하며, 또한 색상, 텍스쳐, 크기, 모양 등 각 물체의 외관 및 특징을 나타내는 속성(attribute) 단어들을 함께 제공한다. 또 참조 표현 내 물체들의 관계를 구분하지 않는 ScanRefer와 Multi3DRES 데이터 집합과는 달리, 참조 표현 내 포함되어 있는 목표 물체와 주변 보조 물체들간의 관계들(relationship)을 모두 "주어(subject) - 관계(relation) - 목적어(object)"의 트리플(triple) 형태로 제공한다. 그뿐만 아니라 OAR3DRef 데이터 집합은 기존의 장면 포인트 클라우드들 외에도 환경 내 목표 및 보조 물체들의 속성 정보 인식을 돕기 위해 ScanNet [8] 데이터 집합에서 제공하는 동일 장면에 대한 멀티-뷰 RGB-D 영상(mutli-view RGB-D images) 데이터들도 함께 제공한다. 따라서 이와 같은 OAR3DRef 데이터 집합을 모델 학습 및 검증에 충분히 활용한다면 자연어 참조 표현에 대한 심층적 이해와 이것에 대응되는 시각적인 목표 물체를 장면 포인트 클라우드에서 식별해내는데 더 용이할 것이다.

2. OAR3DRef 데이터 집합

이 절에서는 OAR3DRef 데이터 집합을 구축하기 위한 과정들을 설명한다. 새로 설계한 OAR3DRef 데이터 집합을 구축하기 위해서는 각 작업마다 목표 물체(target object)뿐만 아니라 보조 물체(auxiliary object)들을 나타내는 단어들, 그 물체들의 다양한 속성(attribute)들을 나타내는 단어들, 그리고 물체들 간의 관계(relationship)들을 나타내는 단어들까지 참조 표현 문장에서 분석해내어 주석 데이터로 담아야 한다. 하지만 ScanRefer, Multi3DRES와 같은 기존 데이터 집합들에 포함되어 있던 약 6만 개 이상의 참조 표현 문장들을 사람이 일일이 모두 수작업으로 이 수준까지 분석해내기란 시간과 노력이 많이 요구되는 매우 어려운 작업이며, 이렇게 수작업으로 분석한 속성 및 관계 분류도 사람마다 차이가 있을 수 있다는 문제점이 있다. 이러한 문제점을 극복하기 위해, 본 연구에서는 언어적 분석기(linguistic parser)와 거대 언어 모델(large language model, LLM)을 함께 효과적으로 활용하는 데이터 집합 구축 전략을 적용함으로써, 수작업의 비용 부담과 잠재적 오류를 대폭 줄였다.

만약 기존의 EDA 언어적 분석기[16,17]만을 이용해 자연어 참조 표현의 의미적 분리를 수행한다면, 자연어 참조 표현 문장에서 물체들을 나타내는 명사 단어들을 찾아낼 수는 있으나 이들이 목표 물체를 나타내는지 보조 물체를 나타내는지는 명확하게 구분해줄 수 없고, 각 물체의 속성 단어와 관계 단어들을 종종 오분류하거나 참조 표현에 포함된 일부의 속성 및 관계 단어들만을 분석해내는 한계성을 가지고 있다. 반면에, 자연어 참조 표현의 의미적 분리를 위해 ChatGPT와 같은 거대 언어 모델(LLM)만을 이용하는 경우에는, 자연어 참조 표현에서 목표 물체 및 보조 물체를 나타내는 단어들, 그들을 수식하는 속성 단어들, 그리고 물체들 간의 관계 단어들을 찾아내는 것도 가능하다. 하지만 자연어 참조 표현이 점차 더 복잡해질수록, 별도의 추가적인 초기 작업 지식을 제공하지 않으면 현재의 거대 언어 모델과 프롬프트 방식으로는 속성 및 관계 단어 분석에 간혹 오류가 발생하거나 자연어 참조 표현에는 존재하지도 않는 단어를 억지로 분류 항목으로 채워 넣는 소위 환각(hallucination) 현상들이 종종 발생한다. 이러한 문제점들을 극복하기 위해, 본 연구에서는 언어적 분석기와 거대 언어 모델, 그리고 사람에 의한 후처리 수작업(manual post-processing)을 함께 상호보완적으로 활용하는 효율적인 데이터 집합 구축 전략을 적용하였다.

Fig. 4와 같이, 먼저 전문화된 인간 지식에 기초한 언어적 분석기를 통해 자연어 참조 표현 문장에서 목표 물체 단어 및 보조 물체 단어과 이들 간의 관계 단어들을 일차적으로 찾아

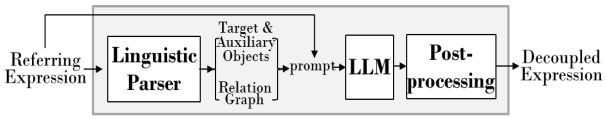


그림 4. 데이터 집합 구축 과정.

Fig. 4. Dataset construction process.

```
Given the input sentence:
  "{sentence}"
1. {main_object} is category name of the main object in the sentence.
2. List only words that appear in the sentence, that describe attribute words (like
   color, texture, shape, and size) of only {main_object}
3. {aux_objects} is category name of auxiliary objects only in the sentence.
4. Identify more category names of all auxiliary in the sentence, and for each
   auxiliary object, list any words that appear in the sentence, that describe their
   attribute words (like color, texture, shape and size) of the auxiliary object in the
   {sentence}.
5. Based on the relationships found in the {rel_graphs}, include only relationships
   that appear in the sentence between one object and another object, using the
   following format. For each relationship, identify the `subject`, 'sub_id`, `relation`,
   'object', 'obj_id.
6. Identify more relationships of between all {total_objects} in the sentence, and
   add them in {rel_graphs}.
7. If there are no words about color, texture, shape, size, and relationships in the
   list, return an empty list.
8. Format the extracted information into the following JSON structure and save it:
   ```json
  {{
    "main": {{
      "m_object": "[category name of main object]",
      "id": index of main object in total_objects,
      "m_attributes": ["[main object attribute words in the sentence]"]
    }},
    "auxiliary": [
      {{
        "a_object": "[category name of auxiliary object 1]",
        "id": index of auxiliary object 1 in total_objects,
        "a_attributes": ["[object 1 attribute words in the sentence]"]
      }},
    ],
    "relationship": [
      {{
          "subject": "[object 1 as the subject in the sentence]",
          "sub_id": "id of object 1's category",
          "relation": "[relation between object 1 and object 2]"
          "object": "[object 2 as the object in the sentence]",
          "obj_id": "id of object 2's category",
      }},
    ]
  }}
  '''
```

그림 5. 복잡한 참조 표현을 여러 의미적 구성 요소들로 분리하도록 거대 언어 모델에게 요청하는 프롬프트의 한 예.

Fig. 5. An example prompt for asking LLM to decouple the complex referring expression into multiple semantic components.
```

낸 다음, 이들을 초기 작업 지식으로 삼아 작업 요청 프롬프트(prompt)를 생성하여 거대 언어 모델에게 전면적이고 세부적인 참조 표현 분석 작업을 의뢰한다. 이후 거대 언어 모델로부터 응답으로 받은 참조 표현 분석 결과물에 대한 점검과 수정을 위한 인간의 후처리 수작업을 진행함으로써, 최종적으로 의미적으로 잘 분리된 참조 표현 주석 데이터들을 생성한다.

Fig. 5는 거대 언어 모델에게 보내는 프롬프트의 한 예를 나타낸다. 프롬프트의 구성을 자세히 살펴보면, 맨 먼저 자연어 참조 표현 문장을 제공한다. 이어서 언어적 분석기를 통해 획득한 목표 및 보조 물체가 무엇인지 알려주고, 이들 외에 혹시 참조 표현 문장 내에 다른 추가적인 보조 물체 단어들이 존재하는지를 판단하도록 한다. 이후 참조 표현 문장에서 목표 및 보조 물체 각각에 대한 속성 단어들을 찾도록 요구한다. 다음으로 언어적 분석기를 통해 참조 표현 문자에서 추출한 관계 단어들과 관계 표현 형식을 알려주고, 이들 외에 추가할 새로운 관계 단어들이 참조 표현에 포함되어 있다면 추출해 내도록 요청한다. 마지막으로 참조 표현 문장을 분석한 결과물들을 JSON 형식에 맞추어 반환하도록 요청한다. 거대 언어 모델은 위와 같은 프롬프트를 입력받아 최종적으로 JSON 형식의 분석 결과물을 출력한다. 이후 거대 언어 모델이 잘못 처리한 오류들을 찾아내기 위한 후처리 작업을 진행한다. 1차 후처리 작업에서는 알고리즘 기반의 프로그램을 이용하여, 참조 표현 문장에 등장하지 않는 단어들임에도 거대 언어 모델의 환각 현상에 의해 주석 데이터로 추출된 단어들을 찾아 제거한다. 이후 2차 후처리 작업에서는 훈련된 연구원들을 통해 거대 언어 모델이 복잡한 자연어 참조 표현들에서 미처 분석해내지 못한 물체들의 속성 단어나 이들 간의 관계 단어들을 찾아 주석 데이터에 포함하도록 한다.

이와 같은 방법으로 구축된 OAR3DRef 데이터 집합의 구성을 목표 물체의 개수 및 보조 물체의 유무에 따라 참조 표현 및 작업 유형(task type)을 다음과 같이 나눈다. 보조 물체가 있는 제로 목표(zero target with auxiliary, zt_w_aux), 보조 물체가 없는 제로 목표(zero target without auxiliary, zt_wo_aux), 보조 물체가 있는 단일 목표(single target with auxiliary, st_w_aux), 보조 물체가 없는 단일 목표(single target without auxiliary, st_wo_aux), 보조 물체가 있는 다중 목표(multiple targets with auxiliary, mt_w_aux), 보조 물체가 없는 다중 목표(multiple target without auxiliary, mt_wo_aux) 6가지 유형으로 나눈다. 6가지 유형별 데이터 분포는 Fig. 6과 같다. 전체 참조 표현 문장 개수는 훈련(train) 데이터는 43838개, 검증(validation) 데이터는 11120개로 이루어져 있다. 6 가지 유형별 훈련 및 검증 데이터의 분포를 살펴보면, zt_w_aux 유형은 각각 4809개, 887개, zt_wo_aux 유형은 각각 53개, 19개, st_w_aux 유형은 각각 28831개, 7367개, st_wo_aux 유형은 각각 407개, 97개, mt_w_aux 유형은 각각 8993개, 2516개, mt_wo_aux 유형은 각각 745개, 241개로 구성되어 있다. 훈련 및 검증 데이터를 모두 합쳐 st_w_aux 유형의 개수가 가장 많았으며, 반면에 zt_wo_aux 유형의 개수가 가장 적었다.

한편, 보조 물체가 존재하는 유형들(zt_w_aux, st_w_aux, mt_w_aux)이 보조 물체가 존재하지 않는 유형들(zt_wo_aux, st_wo_aux, mt_wo_aux) 보다 각각 41428, 10413개 더 많았으며, 이는 목표 물체를 찾기 위해서는 보조 물체들의 속성과 이들 간의 관계를 함께 이해해야 하는 작업들이 많은 것으로 판단

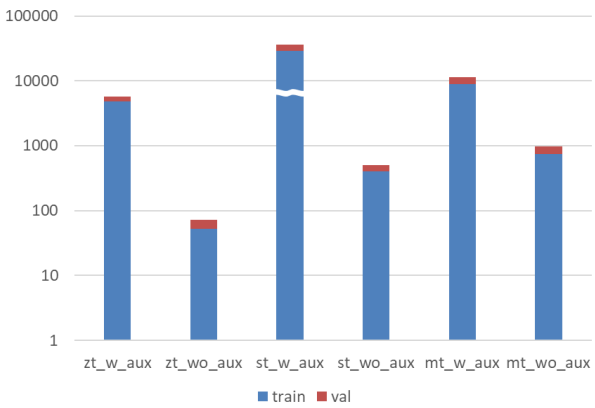


그림 6. OAR3DRef 벤치마크 데이터 집합의 데이터 분포.

Fig. 6. Data distribution of the OAR3DRef benchmark dataset.

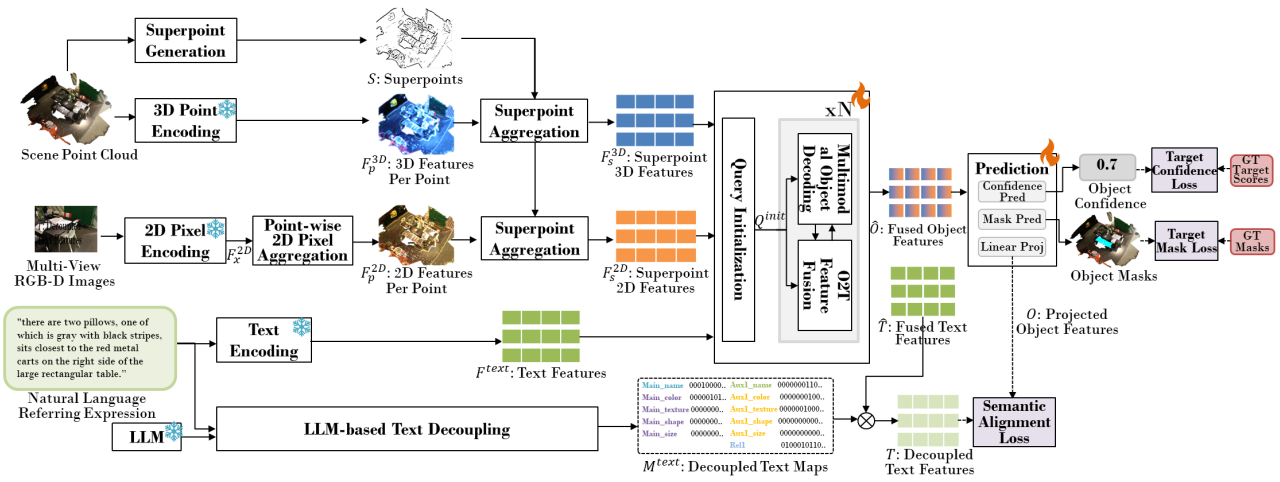


그림 7. 본 논문에서 제안하는 베이스라인 모델인 TDMF의 구조.

Fig. 7. Architecture of the proposed baseline model, TDMF.

된다. 이와 같이 자연어 참조 표현에 등장하는 물체들의 속성들과 물체들 간의 관계들을 심층적으로 표현한 새로운 데이터 집합 OAR3DRef를 효과적으로 모델 학습에 이용하기 위해서는 의미적으로 분리된 참조 표현이 반영된 텍스트 특징 인코딩 기법, 각 물체의 멀티모달 특징 디코딩 기법, 분리된 참조 표현에 적합한 새로운 손실 함수의 설계 등이 이루어져야 한다. 이와 관련한 자세한 내용은 베이스라인 모델 설계를 설명하는 4장에서 자세히 다룰 예정이다.

IV. 베이스라인 모델의 설계

1. 모델 개요

본 논문에서는 물체들의 속성(attribute)과 그들 간의 관계(relationship)들을 중심으로 표현된 새로운 일반화된 3차원 참조 표현 분할(3D-GRES) 데이터 집합인 OAR3DRef에 적용하기 위한 효과적인 베이스라인 모델(baseline model)로서, TDMF (Text Decoupling and Multimodal Fusion)을 제안한다. 제안하는 TDMF 모델은 기존 모델 중 대표적인 3차원 참조 표현 분할 모델인 MDIN [3]을 토대로, (1) 언어적 분석기와 거대 언어 모델(LLM)을 이용해 참조 표현 텍스트로부터 목표 물체 외에 보조 물체들의 속성들과 관계들까지 상세하게 분석할 수 있는 참조 표현 텍스트 분리(text decoupling) 기능 추가와 더불어, (2) 이렇게 분리된 텍스트 구성 요소들을 활용하여 포인트 클라우드에서 도출해낸 물체들의 시각적 특징들과 효과적으로 의미적 정렬을 수행하는 새로운 손실 함수(loss function) 정의, (3) 물체들의 속성 인식 능력 개선을 위해 멀티-뷰 RGB-G 영상들에서 추출해내는 2차원 시각적 특징(2D visual feature)들을 효과적으로 활용하는 방식, (4) 각 물체의 2차원 및 3차원 시각적 특징(visual feature)들과 자연어 참조 표현의 텍스트 특징(text feature)들을 융합하는 새로운 양방향 특징 융합 방식 등을 추가적으로 포함하도록 확장한 것이다.

Fig. 7은 본 논문에서 제안하는 베이스라인 모델 TDMF의 전체 구성을 나타낸 그림이다. TDMF는 크게 (1) 2차원 및 3차원 수퍼포인트 인코딩(2D & 3D Superpoint Encoding), (2) 텍스트 인코딩(Text Encoding), (3) 거대 언어 모델(LLM) 기반의 텍스트 분리(LLM-based Text Decoupling), (4) 멀티-모달 물체 디코딩(Multi-Modal Object Decoding)과 O2T 특징 융합(Object-to-Text Feature Fusion), 그리고 (5) 예측 및 손실(Prediction and Loss) 단계로 나뉜다. 그리고 2차원 및 3차원 수퍼포인트 인코딩 단계는 다시 수퍼포인트 생성(Superpoint Generation), 3차원 포인트 인코딩(3D Point Encoding), 2차원 픽셀 인코딩(2D Pixel Encoding)과 2차원 픽셀 집계(2D Pixel Aggregation). 그리고 수퍼포인트 집계(superpoint aggregation) 모듈로 구성된다.

먼저 수퍼포인트 생성 모듈에서는 유사한 기하학적 특징을 갖는 포인트들끼리 그룹핑한 수퍼포인트들(superpoints) S을 생성한다. 이후 포인트 클라우드로부터 사전 학습된 3차원 포인트 인코딩 모듈인 SparseNet [18]을 통해 포인트별 3차원 특징들 F_p^{3D}을 추출한다. 한편 멀티-뷰 2차원 RGB 영상들로부터 사전 학습된 2차원 픽셀 인코딩 모듈로 ViT [19]를 채용하여 픽셀별 2차원 특징들 F_x^{2D}을 추출한다. 이후 2차원 픽셀 집계 모듈을 통해 포인트별 2차원 특징들 F_p^{2D}을 생성한다. 이후 수퍼포인트 집계 모듈에서는 앞서 생성한 수퍼포인트 S를 포인트별 2차원 및 3차원 특징들에 적용 및 집계하여 수퍼포인트 2차원 특징들 F_s^{2D}과 수퍼포인트 3차원 특징들 F_s^{3D}을 생성하게 된다.

제안 모델의 멀티-모달 물체 디코딩 단계에서는 쿼리 초기화 모듈(query initialization)과 N개의 계층으로 구성된 멀티모달 물체 디코딩(multimodal object decoding) 모듈과 O2T 특징 융합(object to text feature fusion) 모듈로 구성되어 있다. 먼저 앞서 생성된 수퍼포인트 2차원 특징들 F_s^{2D}과 3차원 특징들 F_s^{3D}, 그리고 텍스트 특징들 F^{Text}을 입력받아, 초기 쿼리들 Q^{init}을 생성한다. 이후 수퍼포인트 2차원 특징들 F_s^{2D}과 3차원 특징들 F_s^{3D}, 초기 쿼리들 Q^{init}, 텍스트 특징들 F^{Text}을 입력받아 물체 디코딩 모듈에서는 융합된 물체 특징들(fused object features) $\mathcal{O}$을, 특징 융합 모듈에서는 융합된 텍스트 특징들(fused text features) $\mathcal{T}$을 생성한다. 이때 두 모듈 간의 양방향 융합을 통해 텍스트와 시각적 특징이 서로 상호작용하여 융합되게 된다.

한편, 제안 모델의 예측 단계는 목표 신뢰도 예측(confidence prediction), 목표 마스크 예측(mask prediction), 그리고 선형 투영(linear projection) 모듈로 구성되어 있으며, 융합된 물체 특징들인 $\mathcal{O}$을 입력받아 목표 물체일 가능성을 예측한 신뢰도(confidence)와 물체 영역을 예측한 마스크(mask), 투영된 물체 특징들(projected object features)인 O를 생성한다.

마지막 제안 모델의 손실 단계에서는 목표 신뢰도 손실(target confidence loss), 목표 마스크 손실(target mask loss), 그리고 의미적 정렬 손실(semantic alignment loss) 모듈로 구성된다. 목표 신뢰도 손실 모듈에서는 예측된 목표 신뢰도와 정답(ground truth, GT) 목표 신뢰도 간의 손실을, 목표 마스크 손실 모듈에서는 예측된 물체의 마스크와 정답 목표 물체의 마스크 간의 손실을 각각 계산한다. 한편, 의미적 정렬 손실 모듈에서는 분리된 참조 표현에서 추출한 텍스트 특징들 T과 장면 포인트 클라우드에서 추출한 물체 특징들 O 사이의 의미적 정렬을 통해 손실을 계산한다. 제안 모델은 이 3가지 서로 다른 손실들을 이용해 모델 학습을 수행하게 된다. 제안 모델의 주요 모듈과 해당 모듈에 적용된 새로운 기법에 대해서 후속 절들에서 자세히 설명한다.

2. 시각 및 텍스트 특징 인코딩

앞서 모델 개요에서 설명하였듯이, 제안 모델 TDMF는 2차원 및 3차원 수퍼포인트 인코딩(2D & 3D superpoint encoding) 단계를 통해서는 장면 포인트 클라우드와 멀티-뷰 RGB-D 영상들로부터 각 수퍼포인트별 2차원 및 3차원 시각적 특징들을 추출해내고, 반면에 텍스트 인코딩(text encoding) 단계를 통해서는 자연어, 텍스트 형태의 참조 표현으로부터는 단어별로 인코딩된 텍스트 특징들을 추출해낸다. 한편으로 거대 언어 모델(LLM) [20] 기반의 텍스트 분리(LLM-based text decoupling) 단계에서는 거대 언어 모델(LLM)을 이용하여 입력 참조 표현 텍스트로부터 물체들의 속성과 관계들을 상세히 분석함으로써 이들로부터 분리된 텍스트 맵(decoupled text map)들을 생성해낸다.

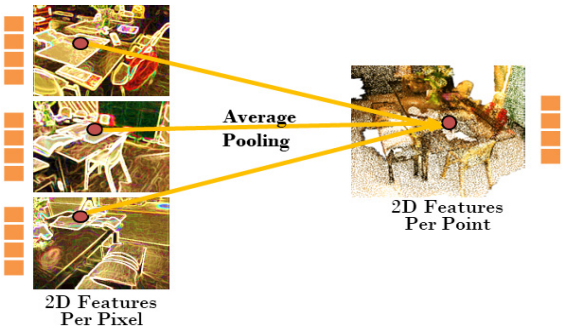


그림 8. 각 포인트별 2차원 시각적 특징들의 집계.
Fig. 8. Point-wise 2d feature aggregation.

자세히 설명하면, 먼저 2차원 및 3차원 수퍼포인트 인코딩 단계에서는 먼저 수퍼포인트 생성(Superpoint Generation) 모듈을 통해 포인트보다 큰 단위인 수퍼포인트(superpoint)들의 집합 S을 생성해낸다. 이와 동시에 다른 한편으로는 장면 포인트 클라우드로부터 3차원 포인트 인코딩(3D point encoding) 모듈을 통해 포인트별 3차원 시각적 특징들 F_p^{3D}을, 동일 장면에 대한 멀티-뷰 RGB-D 영상들로부터 2차원 픽셀 인코딩(2D pixel encoding) 모듈을 통해 픽셀별 2차원 시각적 특징들 F_x^{2D}을 추출해낸다. 이후 픽셀별 2차원 시각적 특징으로부터 포인트별 2차원 픽셀 집계(point-wise 2D pixel aggregation) 모듈을 통해 각 포인트에 매핑되는 2차원 픽셀 특징들을 집계하는 과정을 거친다.

Fig. 8과 같이, 멀티-뷰 RGB-D 영상들과 함께 입력된 영상별 카메라 포즈값을 이용하여 포인트 클라우드의 각 포인트들을 각 영상에 투영한다. 이후 투영된 포인트의 (x, y) 좌표를 이용하여 깊이 영상 내의 해당 좌표에 대한 깊이 값을 찾은 후, 투영된 포인트의 z좌표와의 차가 임계값보다 작은 경우 포인트에 대응되는 픽셀로 매핑시킨다. 영상마다 위 과정을 통해 각 포인트마다 대응되는 픽셀들을 찾아 이를 평균 풀링 연산(average pooling)을 통해 집계함으로써 포인트별 2차원 시각적 특징들 F_p^{2D}을 생성하게 된다. 이후 수퍼포인트 집계(superpoint aggregation) 모듈에서는 앞서 생성된 수퍼포인트들의 집합 S를 포인트별 2차원 및 3차원 시각적 특징들에 각각 적용하여 동일한 수퍼포인트에 해당하는 포인트들의 특징 벡터들을 찾아 이를 평균 풀링을 통해 하나의 특징 벡터로 집계함으로써 수퍼포인트별 2차원 시각적 특징들 F_s^{2D}과 3차원 시각적 특징들 F_s^{3D}을 생성해낸다.

한편, 제안 모델 TDMF의 텍스트 인코딩(text encoding) 단계에서는 텍스트 인코딩 모듈로 채용된 사전 학습된 RoBERTa [21]을 통해 입력된 자연어 참조 표현에서 단어 토큰별 텍스트 특징들 F^{Text}을 추출한다. 이후 거대 언어 모델(LLM) 기반 텍스트 분리 모듈(LLM-based text decoupling)에서는 거대 언어 모델(LLM)을 통해 입력된 참조 표현의 문장을 상세히 분석한다. 먼저 입력된 참조 표현은 언어적 분석기를 통해 거대 언어 모델(LLM)에 입력할 프롬프트(prompt)를 생성한다. 이후 거대 언어 모델(LLM)은 해당 프롬프트를 입력으로 받아, 주어진 자연어 참조 표현을 상세히 분석하여 목표 물체 단어, 목표 물체의 속성 단어, 보조 물체 단어, 보조 물체 속성 단어, 목표 물체와 보조 물체 간의 관계 단어 등 분리된 텍스트 구성 요소들(decoupled text components)을 결과물로 출력한다. 이처럼 분리된 참조 표현의 텍스트 구성 요소들은 분리된 텍스트 맵들(decoupled text maps) M^{Text}에 저장된다.

3. 멀티모달 물체 디코딩 및 02T 특징 융합

앞서 설명한 대로, 기존 연구들에서도 포인트 클라우드에서 검출하는 물체들의 시각적 특징과 자연어 참조 표현의 텍스트 특징 간의 효과적인 의미적 정렬(semantic alignment)을 위해 미리 두 특징 간의 융합(feature fusion)을 시도하였다. 하지만 수퍼포인트들의 2차원 및 3차원 시각적 특징들로부터 각 물체별 특징을 얻기 위한 물체 디코딩(object decoding) 과정 동안 이와 같은 이종 특징 간의 융합을 효과적으로 이루기 위해서는 추가적으로 고려해야 할 몇 가지 중요한 요소들이 있다.

첫 번째는 Fig. 9의 (a), (b)와 같이 텍스트 특징만 물체 특징들에 일방적으로 반영하는 기존의 단방향 특징 융합(uni-directional fusion)을 수행할 것인가, 아니면 Fig. 9(c), (d)와 같이 물체 특징들도 텍스트 특징에 반영하는 양방향 특징 융합(bi-directional fusion)을 수행할 것인지를 결정해야 한다. 기존의 단방향 특징 융합에 비해, 양방향 특징 융합은 물체 특징들에 의미적 단위의 언어적 정보를 반영할 수 있을 뿐만 아니라, 동시에 텍스트 특징들에도 각 물체 단위의 특징 정보를 반영할 수 있기 때문에, 두 이질적인 특징들을 서로 상호보완할 수 있다는 장점이 있다. 따라서 본 제안 모델 TDMF에서는 양방향 특징 융합 방식을 선택한다.

고려해야 할 두 번째 중요한 요소는 물체 디코딩 과정동안 이종 특징들 간의 양방향 특징 융합을 수행한다고 가정했을 때, Fig. 9의 (c)와 같이 참조 표현의 텍스트 특징들을 먼저 물체 특징들에 융합한 후 물체 특징들을 다시 텍스트 특징들에 융합하는 순차적 융합(sequential fusion) 방식을 선택할 것인가, 아니면 Fig. 9의 (d)와 같이 텍스트 특징들을 물체 특징들에 융합한 후 텍스트 특징을 반영하지 않은 순수한 물체 특징

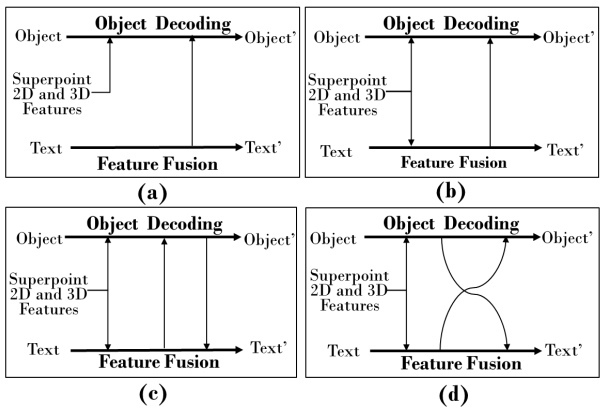


그림 9. 각 물체와 참조 표현 텍스트 간의 서로 다른 특징 융합 방법들.
Fig. 9. Different feature fusions between each object and referring expression text.

들을 텍스트 특징들에 융합하는 교차적 융합(cross fusion) 방식을 선택할 것인지를 결정해야 한다. 양방향 순차적 특징 융합의 경우, 텍스트 정보를 이미 포함하는 물체 특징들을 텍스트 특징들에 다시 반영하기 때문에, 순수한 물체 특징을 반영할 수 없다는 단점이 존재한다. 또한, 여러 개의 융합 계층을 반복할수록 물체 특징들에 텍스트 특징을 융합할 시 텍스트의 언어적 정보가 더 강조되어 반영될 수 있다는 문제가 존재한다. 반면, 양방향 교차적 특징 융합은 양방향 순차적 특징 융합의 문제점을 극복하고 물체 특징들과 텍스트 특징들 간의 융합이 골고루 이루어질 수 있다. 따라서 본 제안 모델 TDMF에서는 이러한 장점을 지닌 양방향 교차적 특징 융합 방식을 채택한다.

마지막으로 고려해야 할 중요한 요소는 수퍼포인트별 2차원 및 3차원 시각적 특징들을 미리 자연어 참조 표현의 텍스트 특징에도 융합할 것인지를 결정하는 문제이다. 물체 디코딩 과정은 본래 수퍼포인트별 2차원 및 3차원 시각적 특징들로부터 각 물체별 시각적 특징을 구해내는 과정이다. 따라서 Fig. 9의 (a)~(d)와 같이 물체 디코딩 과정에서 물체 특징들에 수퍼포인트별 2차원 및 3차원 시각적 특징들이 포함되는 것은 매우 일반적이다. 이에 반해, Fig. 9의 (b)~(d)와 같이 수퍼포인트별 2차원 및 3차원 시각적 특징들을 미리 자연어 참조 표현의 텍스트 특징에도 반영하는 방식은 아직 기존 연구들에서는 시도되지 않았다. 하지만 이와 같이 물체 특징과의 융합 이전에 미리 수퍼포인트별 2차원 3차원 시각적 특징을 텍스트 특징에 융합하는 방식은 수퍼포인트 수준의 풍부한 시각적 정보도 의미적 정렬에 활용함으로써 참조 표현이 가리키는 목표 물체를 더 정확하게 식별하는데 도움을 줄 수 있을 것으로 판단한다. 따라서 제안 모델 TDMF에서는 물체 특징과의 융합 이전에 미리 수퍼포인트별 2차원 3차원 시각적 특징을 텍스트 특징에 융합하는 방식을 선택한다. 이와 같은 설계 요소들을 종합적으로 고려함으로써, 최종적으로 제안 모델 TDMF는 멀티모달 물체 디코딩(multimodal object decoding, MOD) 모듈과 O2T 특징 융합(object to text feature fusion, O2T FF) 모듈을 이용해 Fig. 9의 (d)와 같은 양방향 교차적 특징 융합 방식을 구현한다.

제안 모델의 전체 구성도인 Fig. 7에서 보이듯이, 먼저 쿼리 초기화 모듈(query initialization)에서는 수퍼포인트별 2차원 및 3차원 시각적 특징들로부터 각 물체별 특징들이 될만한 초기 쿼리들(inital queries), 즉 초기 물체 특징들을 생성해낸다. 이후 멀티모달 물체 디코딩(MOD) 모듈은 앞서 포인트 클라우드 및 RGB-D 영상들에서 추출한 수퍼포인트(superpoint)별 2차원 시각적 특징들인 F_s^{2D}와 3차원 시각적 특징들인 F_s^{3D}, 그리고 참조 표현의 텍스트 특징들 F^{Text}로부터, 트랜스포머 디코더(transformer decoder) 신경망 구조를 이용하여 참조 표현에 해당하는 포인트 클라우드 내의 각 물체별 특징들(object features)을 추출해내는 역할을 수행한다. 반면, O2T 특징 융합(O2T FF) 모듈은 각 물체의 시각적 특징들과 의미적 정렬을 수행하므로 앞서 미리 자연어 참조 표현의 텍스트 특징들에 각 물체의 시각적 특징들을 융합하는 역할을 수행한다.

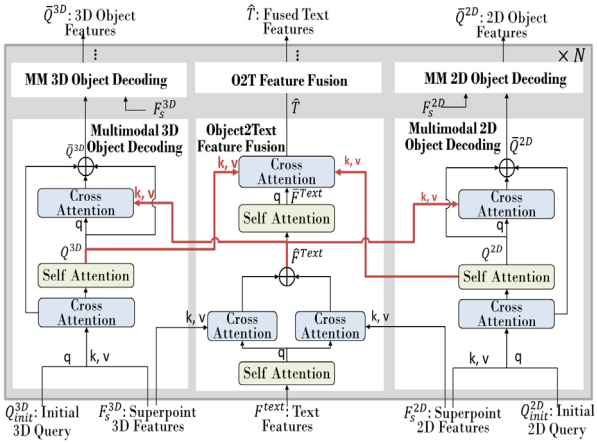


그림 10. 멀티모달 물체 디코딩 모듈과 O2T 특징 융합 모듈

Fig. 10. Multimodal object decoding and O2T feature fusion modules.

Fig. 10은 다계층(multi-layer)으로 구성된 멀티모달 물체 디코딩 및 O2T 특징 융합 과정을 나타내는 그림이다. 그림에서 보듯이, 각 계층(layer) 별 멀티모달 물체 디코딩은 먼저 2차원 시각적 특징을 중심으로 하는 2차원 물체 디코딩(2D object decoding)과 3차원 시각적 특징을 중심으로 하는 3차원 물체 디코딩(3D object decoding)이 병행적으로 수행된다. 2차원 물체 디코딩에서는 초기 쿼리들인 Q^{2D}와 수퍼포인트별 2차원 시각적 특징인 F_s^{2D}와의 교차 어텐션(cross attention)을 통해, 3차원 물체 디코딩에서는 초기 쿼리들인 Q^{3D}와 수퍼포인트별 2차원 시각적 특징인 F_s^{3D}와의 교차 어텐션(cross attention)을 통해, 각각 물체별 2차원 특징 Q^{2D} 및 3차원 특징 Q^{3D}을 얻게 된다.

한편, 이와 동시에 O2T 특징 융합 모듈에서는 교차 어텐션을 통해 자연어 참조 표현의 텍스트 특징들인 F^{text}에 수퍼포인트들의 2차원 및 3차원 시각적 특징들인 F_s^{2D}와 F_s^{3D}를 융합하여 새로운 텍스트 특징인 $\hat{F}^{text}$를 생성한다.

이후 과정에서는 참조 표현의 텍스트 특징 $\hat{F}^{text}$와 장면 포인트 클라우드에서 추출된 각 물체의 2차원 및 3차원 시각적 특징들인 Q^{2D}, Q^{3D}사이에 양방향 교차적 특징 융합(bidirectional cross feature fusion)이 수행된다. Fig. 10에서 빨간색 선(red line)으로 표시된 부분들이 교차 어텐션을 이용한 이종 특징들 간의 양방향 교차적 특징 융합을 나타낸다. 이와 같은 한 계층 과정을 거쳐 얻어지는 융합된 텍스트 특징 $\hat{T}$, 물체별 2차원 시각적 특징 $\hat{Q}^{2D}$, 물체별 3차원 시각적 특징 $\hat{Q}^{3D}$는 다음 계층(next layer)의 입력으로 공급된다. 이러한 한 계층의 처리과정을 미리 설정된 계층 개수 N만큼 반복해준 뒤에, 최후에 생성된 각 물체별 2차원 및 3차원 특징인 $\hat{Q}^{2D}$와 $\hat{Q}^{3D}$를 통합하여 각 물체의 최종 시각적 특징인 $\hat{Q}^{2D3D}$를 생성한다. 이렇게 얻어지는 각 물체의 통합 시각적 특징 $\hat{Q}^{2D3D} = \hat{O}$은 이후 예측 모듈을 통해, 목표 신뢰도 예측, 마스크 예측, 투영된 물체 특징들을 구하는데 이용된다. 또한 참조 표현의 텍

스트 특징 $\hat{T}$과 더불어 상호 간의 의미적 정렬 손실 계산에도 이용하게 된다.

4. 모델 예측과 손실 함수

제안 모델 TDMF의 예측 모듈에서는 각 후보 물체들의 특징 정보 $\hat{Q}^{2D3D} = \hat{O}$를 기초로, 목표 신뢰도 예측(target confidence prediction), 마스크 예측(mask prediction), 그리고 선형 투영(linear projection) 결과들을 생성한다.

먼저 마스크 예측은 각 물체별 특징 정보 $\hat{O}$를 토대로, 장면 포인트 클라우드 내 각 물체의 영역을 3차원 마스크로 예측하는 것이다. 이에 반해 목표 신뢰도 예측은 각 물체의 특징 정보로부터 해당 물체가 목표 물체(target object)에 속할 가능성을 0~1 사이 확률값으로 예측하는 것이다. 그리고 예측된 신뢰도가 일정한 임계치(threshold)보다 높은 경우에는 해당 물체를 목표 물체 중 하나로 간주한다. 모델 학습(model training) 단계가 완료되고, 모델 추론(model inference)을 수행할 때는 신뢰도가 임계치보다 높은 하나 이상의 물체들은 모두 목표 물체들로 간주하여, 이들의 3차원 마스크들을 모두 통합하여 최종 결과 마스크를 제시한다. 따라서 이와 같은 방식으로 제안 모델 TDMF는 단일 목표(single target)뿐만 아니라 다중 목표 물체들(multiple targets)도 함께 분할해낼 수 있다. 또한 동일한 방법으로 장면 포인트 클라우드에서 추출한 모든 물체들의 목표 신뢰도가 모두 임계치보다 낮은 경우에도 목표 물체가 존재하지 않는다(zero target)는 것을 의미하는 빈 마스크를 출력할 수 있다. 한편 예측 모듈이 생성하는 각 후보 물체의 특징 정보 $\hat{O}$에 대한 선형 투영 결과는 뒤에서 설명할 자연어 참조 표현과의 의미적 정렬에 이용된다.

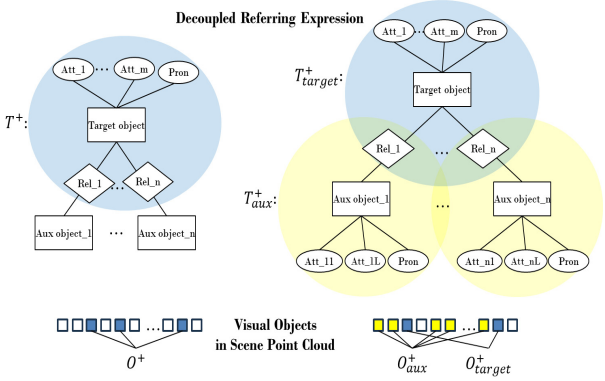
한편 제안 모델 TDMF의 모델 학습 단계에서는 예측 모듈의 결과를 토대로, 목표 신뢰도 손실(target confidence loss), 목표 마스크 손실(target mask loss), 의미적 정렬 손실(semantic alignment loss) 등 3가지 종류의 서로 다른 손실(loss)들을 계산하고, 이 손실들의 합을 줄이는 방향으로 모델 파라미터(parameter)들을 갱신한다. 첫 번째 목표 신뢰도 손실은 Equation 1과 같이, 각 후보 물체가 실제로 목표 물체를 포함하고 있는지를 1과 0으로 표시한 정답 레이블(target ground truth label) L^{target} 과 모델의 목표 신뢰도 예측치(predicted target confidence) P^{target}를 서로 비교하는 것이다. Equation 1에서 BCE는 binary cross entropy loss를 의미한다.

$$L_{conf} = BCE(P^{target}, L^{target}) \tag{1}$$

한편, 두 번째 목표 마스크 손실은 Equation 2와 같이, k번째 목표 물체의 정답 마스크(ground truth mask) M_k^{target}와 모델의 예측 마스크(predicted mask) $\hat{M}_k^{target}$를 서로 비교한다. Equation 2에서 DICE는 Dice loss를 의미한다.

$$L_{mask} = BCE(\hat{M}_k^{target}, M_k^{target}) + DICE(\hat{M}_k^{target}, M_k^{target}) \tag{2}$$

이와 같은 목표 신뢰도 손실과 목표 마스크 손실은 기존 연구들에서도 대부분 공통적으로 사용되어 온 손실들로서, 3차원 참조 표현 분할 모델 학습을 위해서는 필수적인 손실들이다.



(a) Existing models. (b) The proposed model, TDMF.

그림 11. 분리된 참조 표현과 시각적 물체 집합 간의 의미적 정렬 범위.

Fig. 11. Semantic alignment scopes on the decoupled referring expression and the visual object set.

반면에, 세 번째 의미적 정렬 손실은 의미적으로 잘 분리된 자연어 참조 표현들을 모델 학습에 효과적으로 활용하기 위해, 특별히 TDMF 모델에서 새롭게 개념을 확장한 것이다. 여기서 의미적 정렬이란 자연어 참조 표현의 텍스트 특징과 포인트 클라우드에서 검출한 후보 물체들의 시각적 특징 간의 의미적 연관성을 계산하는 것을 의미한다.

Fig. 11의 (a)는 EDA와 MDIN과 같은 기존 모델들에서 의미적 정렬을 시도했던 참조 표현 및 후보 물체들의 범위를 나타낸다. 그림에서 보이듯이 기존 모델들은 자연어 참조 표현에 대한 분리가 목표 물체 단어, 목표 물체 속성 단어, 목표 물체가 보조 물체들과 맺는 공간적 관계 단어 수준까지만 분리를 시도하고 이를 토대로 목표 물체 위주의 단어 집합 T^+를 생성한다. 또 한편으로는 포인트 클라우드에서 검출한 물체들 중 목표 물체 후보로 간주되는 물체들만을 가려내어 물체 집합 O^+를 구성한 다음, 단어 집합 T^+에 포함된 단어들의 텍스트 특징과 물체 집합 O^+에 포함된 물체들의 시각적 특징들끼리만 의미적 정렬을 수행하고 그 손실을 최소화하도록 모델을 학습하였다.

이에 반해, 본 논문의 제안 모델 TDMF에서는 Fig. 11의 (b)와 같이 목표 물체뿐만 아니라 목표 식별에 결정적인 역할을 수행하는 보조 물체들에 대해서도 해당 물체를 가리키는 단어와 속성 단어들, 관계 단어들까지 자연어 참조 표현을 분리한다. 이를 토대로 목표 물체 위주의 단어 집합 T^+_{target}과 더불어 보조 물체 위주의 단어 집합 T^+_{aux}를 함께 생성한다. 여기서 단어 집합 T^+_{target}와 T^+_{aux}에는 Equation 3과 같이 각각 해당 물체의 종류를 나타내는 단어(main, m), 속성 단어(attribute, a), 대명사 단어(pronoun, p), 관계 단어(relationship, r)들이 포함된다.

$$T^+_{target} = T^{target}_m \cup T^{target}_a \cup T^{target}_p \cup T^{target}_r$$

$$T^+_{aux} = T^{aux}_m \cup T^{aux}_a \cup T^{aux}_p \cup T^{aux}_r \tag{3}$$

한편, 목표 물체 후보 집합 O^+_{target}와 보조 물체 후보 집합

O^+_{aux}를 생성하는 방법은 다음과 같다. 먼저 OAR3DRef 훈련 데이터 집합에서 정답 목표 물체(GT target)의 마스크 정보 M^{GT}_{target}를 추출한다. 이 정답 목표 물체 마스크 영역에 해당하는 포인트 클라우드 내의 모든 후보 물체들을 가려내어 목표 물체 후보 집합 O^+_{target}을 생성한다. 정답 마스크를 이용해 보조 물체 후보 집합 O^+_{aux}를 생성하는 방법도 이와 거의 동일하다. 하지만 목표 물체와는 달리, 보조 물체들의 경우는 그 수도 가변적이고 형태도 너무 다양하여 OAR3DRef를 포함해 기존의 모든 데이터 집합에서 정답 마스크를 제공하지 못하고 있다. 따라서 본 연구에서는 보조 물체를 나타내는 단어를 OAR3DRef의 토대가 되는 ScanNet200 데이터 집합의 물체 레이블에 직접 대응시켜, 별도의 의사 정답 마스크(pseudo GT mask)를 구해낸다. 이렇게 구한 의사 정답 마스크를 보조 물체의 정답 마스크 M^{GT}_{aux}로 사용하여, 보조 물체 후보 집합 O^+_{aux}을 생성한다.

그런 다음, 목표 물체 위주의 단어 집합인 T^+_{target}은 목표 물체 후보 집합인 O^+_{target}와, 보조 물체 위주의 단어 집합인 T^+_{aux}는 보조 물체 후보 집합 O^+_{aux}와 각각 따로 의미적 정렬을 진행한 후, 각각의 손실을 결합한다. 따라서 기존 모델들에 비해, 제안 모델 TDMF의 의미적 정렬 손실 계산 방식이 복잡한 자연어 참조 표현의 의미를 심층적으로 이해할 수 있도록 모델을 학습하는데 더 효과적이다.

본 논문에서 제안하는 의미적 정렬 손실 $L_{semantic}$은 Equation 4와 같이 목표 물체에 대한 정렬 손실 $L^{target}_{semantic}$과 보조 물체들에 대한 정렬 손실 $L^{aux}_{semantic}$의 가중 합(weighted sum)으로 계산한다. λ_{target}와 λ_{aux}는 각 손실의 비중을 조절할 수 있는 가중치이다.

$$L_{semantic} = \lambda_{target} L^{target}_{semantic} + \lambda_{aux} L^{aux}_{semantic} \tag{4}$$

한편, 의미적 정렬 손실 $L^{target}_{semantic}$과 $L^{aux}_{semantic}$는 Equation 5와 같이 후보 물체(object)를 기준으로 참조 표현 구성 단어(word)들과 정렬하는 손실 $L_{o\to w}$와 그 반대 방향으로 정렬하는 손실 $L_{w\to o}$를 합해서 계산한다.

$$L^{target}_{semantic} = L^{target}_{o\to w} + L^{target}_{w\to o},\ L^{aux}_{semantic} = L^{aux}_{o\to w} + L^{aux}_{w\to o} \tag{5}$$

그리고 보조 물체의 의미적 정렬 손실 $L^{aux}_{o\to w}$와 $L^{aux}_{w\to o}$은 보조 물체 위주의 단어 집합 T^+_{aux}와 보조 물체 후보 집합 O^+_{aux}를 이용해, Equation 6, Equation 7과 같이 계산한다. 또한, 목표 물체의 손실 $L^{target}_{o\to w}$와 $L^{target}_{w\to o}$도 유사한 방식으로 계산한다.

$$L^{aux}_{o\to w} = \sum_{o_i \in O^+_{aux}} \frac{1}{|T^+_{aux}|} \sum_{t_k \in T^+_{aux}} -\log\left(\frac{\exp((o_i t_k/\tau))}{\sum_{t_j \in T} \exp((o_i t_j/\tau))}\right) \tag{6}$$

$$L^{aux}_{w\to o} = \sum_{t_i \in T^+_{aux}} \frac{1}{|O^+_{aux}|} \sum_{o_k \in O^+_{aux}} -\log\left(\frac{\exp((t_i o_k/\tau))}{\sum_{o_j \in O} \exp((t_i o_j/\tau))}\right) \tag{7}$$

끝으로 제안 모델 TDMF의 모델 학습을 위한 종합 손실 L은 Equation 8과 같이, 목표 신뢰도 손실 L_{conf}, 목표 마스크 손실 L_{mask}, 의미적 정렬 손실 $L_{semantic}$ 등 3가지 서로 다른 손실들의 가중 합(weighted sum)으로 구한다.

$$L = \lambda_{conf} L_{conf} + \lambda_{mask} L_{mask} + \lambda_{semantic} L_{semantic} \tag{8}$$

V. 구현 및 실험

1. 모델 구현과 학습

본 논문에서는 제안 모델 TDMF의 학습 및 성능 평가를 위해 OAR3DRef 데이터 집합을 사용하는데, 이 데이터 집합의 구성은 43838개의 참조 표현들을 포함한 훈련(train) 데이터들과 11120개의 참조 표현들을 포함한 검증(val) 데이터들로 이루어진다. 제안 모델에서는 참조 표현의 의미적 분리를 위해 거대 언어 모델(large language model, LLM)로 ChatGPT 3.5[20]를 이용하였다. 또한 제안 모델의 3차원 포인트 인코딩 모듈로는 사전 학습된 Sparse 3D Uet [18]을 채용하였으며, 2차원 픽셀 인코딩 모듈로는 사전 학습된 ViT [19]의 2차원 픽셀 인코딩 모듈을, 텍스트 인코딩 모듈로는 사전 학습된 RoBERTa [21] 모듈을 채용하였다. 끝으로 제안 모델은 Ubuntu 18.04.6 LTS 환경에서 Pytorch 딥러닝 라이브러리를 이용하여 구현하였으며, 2개의 GeForce GTX 3090 GPU로 구성된 하드웨어 환경에서 훈련과 성능 테스트를 진행하였다.

2. 실험 및 평가

본 논문에서는 제안 모델 TDMF의 성능을 평가하기 위해, 새로운 데이터 집합 OAR3DRef를 기초로 (1) 의미적 정렬 손실을 포함한 손실 함수들에 따른 성능 비교, (2) 멀티-모달 특징 융합 방식들에 따른 성능 비교, (3) 2차원 시각적 특징을 포함한 멀티-모달 물체 특징들에 따른 성능 비교, (4) 기존의 타 모델들과의 성능 비교의 4가지 정량적 실험들과 (5) 기존의 타 모델들과의 정성적 비교 실험까지 총 5가지 실험들을 수행한다. (4) 기존 모델들과의 성능 비교에서는 새로운 데이터 집합 OAR3DRef뿐만 아니라, 기존 데이터 집합인 ScanRef를 이용한 실험들도 함께 진행한다. 또한 본 논문에서는 제안 모델 TDMF과 기존 모델들 간의 3차원 참조 표현 분할 결과들을 서로 비교해보는 정성적 실험들도 진행한다. 정량적 실험들에서 사용하는 성능 평가 척도(metric)는 평균 교차 영역 비율(mean intersection over union, mIoU)과 Acc@0.5 IoU의 평균 정확도(mean accuracy, mAcc)이다. Acc@0.5 IoU는 목표 물체의 예측 마스크와 실제 마스크 간의 겹침 정도(IoU)가 0.5보다 큰 경우에 목표 물체를 올바르게 분할한 것으로 간주하는 분할 정확도 계산 방법이다.

첫 번째 실험은 제안 모델 TDMF에서 모델 학습을 위해 설계한 손실 함수들의 긍정적인 효과를 입증하기 위한 실험이다. 이 실험에서는 (a) 목표 물체의 마스크 손실 함수만을 적용한 경우(L^{target}_{mask}), (b) 목표 물체의 마스크와 신뢰도 손실 함수를 함께 적용한 경우($L^{target}_{mask} + L^{target}_{conf}$), (c) 목표 물체의 마스크와 신뢰도, 목표 물체 위주의 의미적 정렬 손실 함수들을 적용한 경우($L^{target}_{mask} + L^{target}_{conf} + L^{target}_{semantic}$), (d) 제안 모델 TDMF와 같이, 목표 물체의 마스크, 신뢰도, 의미적 정렬 손실 함수들

표 2. 서로 다른 손실 함수들 간의 비교.

Table 2. Comparison with different loss functions.

#	Losses	Task Type						Overall	
		zt_w_aux	zt_wo_aux	st_w_aux	st_wo_aux	mt_w_aux	mt_wo_aux	Acc@0.5	mIoU
(a)	L_{mask}^{target}	0.0	0.0	0.0	0.7	6.4	17.0	2.2	9.7
(b)	$L_{mask}^{target}+L_{conf}^{target}$	35.7	15.8	31.1	46.4	38.1	61.4	33.8	38.8
(c)	$L_{mask}^{target}+L_{conf}^{target}+L_{semantic}^{target}$	36.4	26.3	32.0	49.5	38.3	59.8	34.6	39.5
(d)	$L_{mask}^{target}+L_{conf}^{target}+L_{semantic}^{target}+L_{semantic}^{aux}$	40.2	26.3	32.3	45.4	38.7	59.3	35.1	39.9

외에, 보조 물체들의 의미적 정렬 손실 함수까지 적용한 경우 ($L_{mask}^{target}+L_{conf}^{target}+L_{semantic}^{target}+L_{semantic}^{aux}$) 간의 성능 비교 실험을 수행한다.

Table 2의 실험 결과를 살펴보면, 제안 모델 TDMF와 같이 보조 물체들의 의미적 정렬 손실까지 손실 함수에 반영한 (d) $L_{mask}^{target}+L_{conf}^{target}+L_{semantic}^{target}+L_{semantic}^{aux}$의 경우가 다른 손실 함수들의 조합보다 평균 정확도인 Acc@0.5와 평균 교차율인 mIoU 면에서 가장 높은 성능을 보였다. 예컨대, (d)의 경우가 Acc@0.5 면에서는 (a)와 (b), (c)의 경우보다 각각 약 100*(35.1-2.2)/2.2=1495.45%, 3.85%, 1.44%의 성능 향상을 보였고 mIoU 면에서는 각각 약 100*(39.9-9.7)/9.7=311.34%, 2.84%, 1.01%의 성능 향상을 보였다. 특히 목표 물체의 의미적 정렬 손실까지만 적용한 (c) $L_{mask}^{target}+L_{conf}^{target}+L_{semantic}^{target}$에 비해서는 Acc@0.5와 mIoU 면에서 각각 약 1.45%와 1.01%의 성능 향상을 보였을 뿐만 아니라, 보조 물체가 포함된 작업 유형들인 zt_w_aux, st_w_aux, mt_w_aux에서 각각 100*(40.2-36.4)/36.4=10.44%, 0.94%, 1.04%의 정확도 향상을 보였다. 이러한 실험 결과들을 통해, 우리는 분리된 참조 표현을 토대로 목표 물체뿐만 아니라 보조 물체들의 의미적 정렬 손실까지 손실 함수에 반영한 제안 모델 TDMF의 우수성과 손실 함수의 긍정적 효과를 확인할 수 있다.

이 밖에도 Table 2의 결과를 자세히 살펴보면, 목표 물체의 마스크 손실 함수만 적용한 (a) L_{mask}^{target}의 경우가 다른 모든 손실 함수들의 조합보다 모든 성능 척도 면에서 가장 낮은 성능을 보였다. 예컨대, 목표 물체의 신뢰도 손실만을 추가한 (b) $L_{mask}^{target}+L_{conf}^{target}$의 경우도 (a) L_{mask}^{target}에 비해 Acc@0.5과 mIoU 면에서 각각 약 1,436.3%, 300%의 아주 높은 성능 향상을 보였다. 이러한 결과를 통해, 목표 물체의 마스크 손실에 비해 목표 물체 신뢰도 손실이 모델 성능 향상에 더 큰 효과를 준다는 것을 확인할 수 있다. 한편, 목표 물체의 의미적 정렬 손실을 추가한 (c) $L_{mask}^{target}+L_{conf}^{target}+L_{semantic}^{target}$의 경우가 그렇지 않은 (b)에 비해 Acc@0.5와 mIoU에서 각각 약 2.37%와 1.8%의 성능 향상을 보였을 뿐만 아니라, 세부적인 모든 작업 유형들에서도 더 향상된 정확도를 나타내었다. 이러한 실험 결과를 통해, 목표 물체의 의미적 정렬 손실이 모델의 성능 개선에 미치는 긍정적 효과도 함께 확인할 수 있다.

두 번째 실험은 제안 모델 TDMF이 멀티모달 물체 디코딩 모듈(MM object decoding)과 물체-텍스트 특징 융합 모듈(O2T feature fusion)을 통해, 물체 특징과 텍스트 특징 간에 양방향 교차적 특징 융합(bidirectional cross feature fusion)을 수행하는 방식의 타당성을 입증하기 위한 실험이다. 이 실험에서는 포인트 클라우드에서 추출한 물체 특징과 자연어 참조 표현에서 추출한 텍스트 특징 간의 융합 방식에 따라, 앞선 4.3절의 Fig. 9와 같이 서로 다른 4가지의 융합 방식들을 채용했을 때 각각의 모델 성능을 서로 비교한다.

Fig. 9의 (a)와 (b)는 물체 특징에 텍스트 특징을 융합하는 단방향 융합 방식(uni-direction)을, Fig. 9의 (c)와 (d)는 물체 특징과 텍스트 특징 간의 양방향 융합 방식(bi-direction)을 나타낸다. 단방향 융합 방식 중에서 (a)는 수퍼포인트의 2차원 및 3차원 시각적 특징을 물체 특징 디코딩에만 공급하는 방식을, (b)는 수퍼포인트의 2차원 및 3차원 시각적 특징을 물체 특징뿐만 아니라 텍스트 특징에도 미리 반영하는 방식을 각각 나타낸다. 한편, 양방향 융합 방식 중에서 (c)는 물체 특징과 텍스트 특징 간의 융합을 순차적으로 수행하는 양방향 순차적 융합 방식을, (d)는 물체 특징과 텍스트 특징을 교차하여 융합하는 양방향 교차적 융합 방식을 나타낸다. 제안 모델 TDMF는 멀티모달 물체 디코딩과 물체-텍스트 특징 융합 모듈을 통해, (d)와 같은 양방향 교차적 특징 융합 방식을 채택하고 있다.

이 실험의 결과들을 나타내는 Table 3를 살펴보면, 제안 모델 TDMF와 동일한 양방향 교차적 특징 융합 방식인 (d)의 경우가 다른 모든 비교 대상들인 (a), (b), (c)의 융합 방식들에 비해 가장 높은 성능을 보였다. 예컨대, (d)의 융합 방식은 (a), (b), (c)의 융합 방식에 비해 평균 정확도 Acc@0.5면에서는 각각 약 100*(35.1-32.4)/32.4=8.33%, 6.36%, 3.24%의 성능 향상을 보였고, 평균 교차율 mIoU 면에서는 각각 약 100*(39.9-37.4)/37.4=6.68%, 4.72%, 2.57%의 성능 향상을 보였다.

또한, Table 3의 결과들을 살펴보면 물체 특징과 텍스트 특징 간의 양방향 특징 융합(bi-directional feature fusion)을 수행하는 (c), (d) 방식들이 물체 특징에 텍스트 특징을 오직 단방향으로 융합(uni-directional feature fusion)하는 (a), (b) 방식들에 비해, 전체적으로 더 높은 성능을 보인 것을 알 수 있다. 예컨대, 양방향 특징 융합 방식의 하나인 (c)의 경우가 단방향 특징 융합 방식의 하나인 (a)에 비해 성능 척도 Acc@0.5와 mIoU 면에서 각각 4.94%와 4.01%의 성능 향상을 보인 것을 확인할 수 있다. 이와 같은 실험 결과를 통해, 우리는 물체 특징과 텍스트 특징 간의 상호작용을 더 효과적으로 촉진할 수

표 3. 서로 다른 특징 융합 방법들 간의 비교.

Table 3. Comparison between different feature fusion methods.

Feature Fusion Methods		Task Type						Overall	
		zt_w_aux	zt_wo_aux	st_w_aux	st_wo_aux	mt_w_aux	mt_wo_aux	Acc@0.5	mIoU
Uni-direction	(a)	36.8	21.1	30.4	40.2	34.3	53.9	32.4	37.4
	(b)	34.6	21.1	31.1	42.3	35.8	52.7	33.0	38.1
Bi-direction	(c)	37.0	36.8	31.8	46.4	36.2	59.3	34.0	38.9
	(d)	40.2	26.3	32.3	45.4	38.7	59.3	35.1	39.9

있는 양방향 특징 융합 방식의 긍정적인 효과를 확인할 수 있다.

한편, Table 3에서 서로 다른 양방향 융합 방식들인 (c)와 (d)의 성능을 서로 비교해보면, 교차적 융합을 수행하는 (d)가 순차적인 융합을 수행하는 (c) 보다 Acc@0.5와 mIoU 면에서 각각 약 3.25%, 2.57% 더 성능이 향상된 것을 알 수 있다. 세부적인 작업 유형별로 성능을 비교해보면, zt_wo_aux와 st_wo_aux 같이 보조 물체가 존재하지 않는 작업들에 대해서는 (c)의 순차적인 특징 융합 방식이 다소 더 높은 성능을 보였으나, zt_w_aux와 st_w_aux, mt_w_aux와 같이 보조 물체들이 존재하는 작업들에 대해서는 (d)의 교차적 특징 융합 방식이 각각 약 8.65%, 1.57%, 6.91%의 뚜렷한 성능 향상을 보였다. 이와 같은 실험 결과들을 통해, 제안 모델 TDMF와 같은 양방향 교차적 특징 융합 방식이 보조 물체들을 활용해야 더 난이도가 높은 참조 표현 분할 작업들에 대해서는 양방향 순차적 융합 방식에 비해 더 효과적임을 확인할 수 있다.

한편, Table 3에서 수퍼포인트의 2차원 및 3차원 시각적 특징을 물체 특징에만 미리 반영하는 (a) 방식과 텍스트 특징에도 미리 반영하는 (b), (c), (d)의 방식들과 비교해보면, 수퍼포인트의 2차원 및 3차원 시각적 특징을 물체 특징뿐만 아니라 텍스트 특징에도 미리 반영하는 (b), (c), (d)의 방식들이 더 높은 성능을 보였다. 예컨대, 단방향 특징 융합 방식 중에서도 (b)의 융합 방식이 (a) 융합 방식에 비해 Acc@0.5와 mIoU 면에서 각각 약 1.85%, 1.87% 더 성능이 향상된 것을 알 수 있다. 이같은 실험 결과를 통해, 수퍼 포인트의 2차원 및 3차원 시각적 특징을 물체 특징 뿐만 아니라, 텍스트 특징에도 모두 반영하는 특징 융합 방식이 모델의 성능 향상에 도움이 된다는 것도 확인할 수 있다. 이와 같은 실험 결과들을 종합해볼 때, 제안 모델 TDMF에서 채택한 물체 특징과 텍스트 특징 간의 양방향 교차적 멀티모달 특징 융합 방식을 사용하면서, 수퍼포인트의 2차원 및 3차원 시각적 특징을 물체 특징과 텍스트 특징에 모두 반영하는 방식이 모델의 성능 향상에 도움을 준다는 것을 명확히 확인할 수 있다.

세 번째 실험은 장면 포인트 클라우드에서 추출하는 3차원 시각적 특징 외에, 추가로 멀티-뷰 RGB-D 영상들에서 추출하는 2차원 시각적 특징들도 함께 물체 특징으로 이용하는 제안 모델 TDMF 방식의 긍정적인 효과를 입증하기 위한 실험이다. 이 실험에서는 (a) 기존 모델들과 같이 3차원 시각적 특징만을 이용하는 경우, (b) 제안 모델 TDMF와 같이 3차원 시각적 특징과 더불어 2차원 시각적 특징도 함께 사용하는 경우의 성능을 서로 비교한다.

표 4. 서로 다른 물체 특징들 간의 비교.

Table 4. Comparison between different object features.

Object Features			Task Type						Overall	
#	3D	2D	zt_w_aux	zt_wo_aux	st_w_aux	st_wo_aux	mt_w_aux	mt_wo_aux	Acc@0.5	mIoU
(a)	O		31.1	5.3	31.7	42.3	37.1	58.5	33.5	39.1
(b)	O	O	**40.2**	**26.3**	**32.3**	**45.4**	**38.7**	**59.3**	**35.1**	**39.9**

Table 4의 이 실험의 결과들을 살펴보면, 3차원 시각적 특징과 2차원 시각적 특징을 함께 이용한 (b)의 경우가 3차원 시각적 특징만 사용한 (a)의 경우에 비해, 모든 성능 척도 면에서 더 높은 성능 향상을 보였다. 즉 (b)의 경우가 (a)에 비해 Acc@0.5와 mIoU 면에서 각각 약 4.78%, 2.05%의 성능 향상을 보였다. 여섯 가지 세부 작업 유형별로 성능을 자세히 비교해보면, 보조 물체가 등장하지 않는 zt_wo_aux, st_wo_aux, mt_wo_aux 등의 작업들에 대해서는 각각 약 396.22%, 7.33%, 1.367%의 성능 향상을 보였으며, 그 중에서 특히 zt_wo_aux 작업에서 성능이 가장 큰 폭으로 증가하였음을 확인할 수 있다. 또한, 보조 물체가 등장하는 zt_w_aux, st_w_aux, mt_w_aux 작업들에 대해서는 (b)의 경우가 (a)에 비해 각각 약 29.26%, 1.89%, 4.31% 향상된 성능을 보였다. 이와 같은 실험 결과들을 통해, 포인트 클라우드에서 추출하는 3차원 시각적 특징 외에 멀티-뷰 RGB-D 영상들에서 추출하는 2차원 시각적 특징들도 물체 특징으로 함께 이용하는 제안 모델 TDMF 방식의 긍정적인 효과를 확인할 수 있다.

네 번째 실험은 기존 모델들과의 비교를 통해, 제안 모델 TDMF의 우수성을 입증하기 위한 실험이다. 먼저 제로 및 다중 목표 물체들도 분할해야 하는 일반화된 3차원 참조 표현 분할(3D GRES) 작업들에서 제안 모델 TDMF의 우수성을 입증하기 위해, 데이터 집합 OAR3DRef를 이용해 기존 모델들인 3D-STMN [2], MDIN [3]들과 제안 모델 TDMF의 성능을 비교한다. 이어서 단일 목표 물체의 3차원 참조 표현 분할(3D RES) 작업들에서 제안 모델 TDMF의 우수성을 입증하기 위해, 데이터 집합 ScanRefer를 이용해 TGNN [1], X-RefSeg3D [10], 3D-STMN [2]과 같은 기존 모델들과 제안 모델 TDMF의 성능도 비교한다. Table 5는 OAR3DRef 데이터 집합을 이용한 일반화된 3차원 참조 표현 분할(3D GRES) 작업들에서의 실험 결과를, Table 6는 ScanRefer 데이터 집합을 이용한 단일 목표 3차원 참조 표현 분할(3D RES) 작업들에서의 실험 결과들을 각각 나타낸다.

먼저 Table 5의 OAR3DRef 데이터 집합에 대한 모델 성능들을 비교해보면, 제안 모델TDMF가 Acc@0.5와 mIoU 면에서 3D-STMN 모델보다는 각각 약 12.14%와 23.53%의 성능 향상을, MDIN 모델보다는 각각 4.776%, 3.37%의 성능 향상을 보인 것을 알 수 있다.

또한 세부 작업 유형별로 모델 성능을 비교해보면, 제안 모델 TDMF가 역시 보조 물체가 존재하는 작업들에서 다른 기존

표 5. OAR3DRef 데이터 집합을 이용한 다른 모델들과의 정량적 비교.

Table 5. Quantitative comparison with other models on OAR3DRef dataset.

Models	OAR3DRef Dataset							
	zt_w_aux	zt_wo_aux	st_w_aux	st_wo_aux	mt_w_aux	mt_wo_aux	overall	
							Acc@0.5	mIoU
3D-STMN[2]	00.0	00.0	39.8	57.7	17.5	19.9	31.3	32.3
MDIN[3]	32.1	15.8	31.3	50.5	37.5	59.8	33.5	38.6
TDMF(ours)	40.2	26.3	32.3	45.4	38.7	59.3	35.1	39.9

표 6. ScanRefer 데이터 집합을 이용한 다른 모델들과의 정량적 비교.

Table 6. Quantitative comparison with other models on ScanRefer dataset.

Models	Unique	Multiple	Overall	
			Acc@0.5	mIoU
TGNN[1]	57.8	26.6	32.7	28.8
X-RefSeg3D[10]	-	-	33.8	29.9
3D-STMN[2]	84.0	29.2	39.8	39.5
TDMF(ours)	86.9	44.1	52.5	47.6

모델들에 비해 더 높은 성능을 나타내었다. 예컨대, 제안 모델 TDMF는 보조 물체가 존재하는 zt_w_aux, st_w_aux, mt_w_aux 작업들에서 기존의 MDIN 모델보다 각각 약 25.23%, 3.19%, 3.2%의 성능 향상을 보였다. 따라서 이와 같은 실험 결과들을 도대로, 제로 및 다중 목표 물체들도 분할해야 하는 일반화된 3차원 참조 표현 분할(3D GRES) 작업들에서 제안 모델인 TDMF가 기존 모델들에 비해 더 높은 성능을 보인다는 것을 확인할 수 있다. 한편, Table 5의 OAR3DRef 데이터 집합에 대한 세부 작업 유형별 모델 성능들을 살펴보면, 기존의 3D-STMN 모델이 단일 목표 물체 작업인 st_w_aux와 st_wo_aux 작업에서는 가장 높은 성능을 보였으나, 제로 목표 물체 작업인 zt_w_aux와 zt_wo_aux에서는 모두 0.0%의 정확도를 나타내었다. 또한 다중 목표 물체 작업들인 mt_w_aux와 mt_wo_aux에서는 MDIN과 제안 모델 TDMF보다 현저히 낮은 성능을 보인 것을 알 수 있다. 그 이유는 기존의 3D-STMN 모델은 앞서 소개한 바와 같이 본래 단일 목표 3차원 참조 표현 분할(3D RES) 작업을 위해 설계된 모델로서, 일반화된 3차원 참조 표현 분할(3D GRES) 작업을 수행하는 데는 한계가 있음을 이 실험을 통해 보여준 것으로 판단된다.

한편, Table 6의 ScanRefer 데이터 집합을 이용한 3차원 참조 표현 분할(3D RES) 작업들에 대한 모델들의 성능을 비교해보면, 제안 모델인 TDMF가 이 데이터 집합에서도 다른 기존 모델들인 TGNN [1], X-RefSeg3D [10], 3D-STMN [2]들에 비해 가장 높은 성능을 보인 것을 알 수 있다. 예컨대, 제안 모델 TDMF는 Acc@.5와 mIoU 면에서 TGNN보다는 60.55%, 65.28%, X-RefSeg3D보다는 55.33%, 59.20%, 3D-STMN보다는 31.91%, 20.51% 더 높은 성능 향상을 보였다. 특히 복수의 다중 목표 물체들이 포함된 Multiple 작업들에서 제안 모델 TDMF는 TGNN보다는 65.79%, 3D-STMN보다는 51.03% 더 높은 성능 향상을 나타내었다. 이와 같은 실험 결과들은 ScanRefer 데이터 집합을 이용한 기존의 3차원 참조 표현 분할(3D RES) 작업들에도 제안 모델 TDMF가 기존 모델들에 비해 높은 분할 성능을 보인다는 것을 확인할 수 있다.

마지막으로 본 논문에서는 Table 7과 같이 OAR3DRef 데이터 집합의 몇 가지 사례들을 중심으로, 제안 모델 TDMF과 기존 모델들의 3차원 참조 표현 분할 결과들을 정성적으로 비교해본다. Table 7에 표시한 사례 (a)와 사례 (b)는 제안 모델 TDMF이 성공적인 분할 결과를 보여준 경우이고, 반면에 Tabel 7의 사례 (c)는 만족스럽지 못한 분할 결과를 보여준 경우이다.

Table 7 (a)의 경우는 보조 물체가 존재하는 환경에서 다중 목표 물체(mt_w_aux)를 찾아 분할하는 작업의 하나이다. 자연어 참조 표현을 살펴보면, 이 경우는 문(door)의 옆에 위치한 검은색 선반(shelf)을 찾아서 목표 물체로 분할해야 하는 작업이다. 여기서 목표 물체 단어는 선반이며, 보조 물체를 나타내는 단어는 문이다. 입력 장면 포인트 클라우드와 정답 목표 마스크(ground-truth target mask)를 살펴보면, 문을 기준으로 양옆에 위치한 검은색 선반 모두가 목표 물체로 분할되어야 한다. 3D-STMN의 분할 결과를 살펴보면, 문의 왼쪽에 위치한 검은색 선반 하나는 목표 물체로 찾아내었으나, 나머지 오른쪽에 위치한 검은색 선반은 목표 물체로 찾아내지 못하였다. 즉, 다중 목표 물체(multiple targets) 분할에 실패한 것이다. 한편, MDIN의 경우, 문의 왼쪽에 위치한 검은색 선반은 목표 물체로 분할하였으나, 문의 바로 오른쪽에 위치한 검은색 선반 대신 멀리 떨어진 하얀색 선반을 목표 물체로 잘못 분할한 것을 볼 수 있다. 하지만 이에 반해, 제안 모델 TDMF는 자연어 참조 표현에 등장하는 목표 물체 단어인 선반과 보조 물체 단어인 문, 그리고 이 둘 간의 위치 관계를 나타내는 관계 단어, 그리고 목표 물체 선반의 색상 단어들의 의미를 이해하고 효과적으로 활용함으로써, 최종적으로 참조 표현에 부합하는 목표 물체들로서 두 개의 검은색 선반들만을 정확히 분할해내었다. 이와 같은 결과를 통해, 다중 목표 물체 분할 작업에 대한 제안 모델 TDMF의 우수성을 다시 한번 확인할 수 있다.

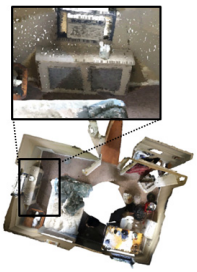
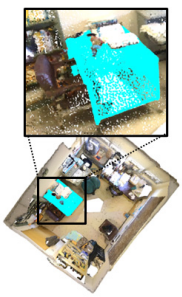
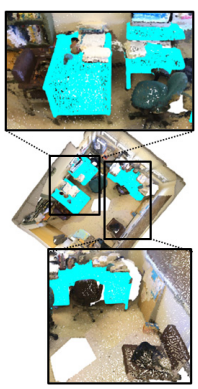
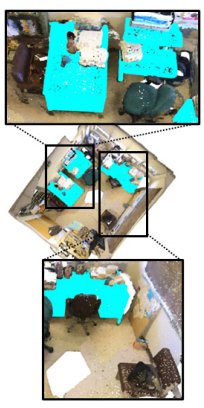
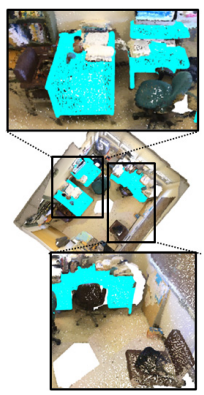
Table 7 (b)의 경우는 자연어 참조 표현에 부합하는 목표 물체가 장면 포인트 클라우드에는 하나도 존재하지 않는 제로 목표 물체(zt_w_aux) 작업의 하나이다. 자연어 참조 표현을 살펴보면, 이 경우는 싱크대(sink)의 왼쪽 모서리에 위치하고 있으면서 검은색 하단과 황갈색 상단을 가지고 있고 바닥부터 천장까지 이르는 긴 사물함(cabinet)을 찾아 목표 물체로 분할해야 하는 작업이다. 여기서 목표 물체 단어는 사물함, 보조 물체 단어는 싱크대가 된다. 이 경우에는 장면 포인트 클라우드 내에 사물함들이 여러 개 존재하지만 정작 참조 표현에 부합되는 목표 물체 사물함은 존재하지 않기 때문에, 정답 목표 마스크도 표시되지 않았다. 하지만 3D-STMN의 분할 결과를 살펴보면, 벽에 붙어있는 하얀색 사물함을 목표 물체로 잘못 분할한 것을 확인할 수 있다. 또한, MDIN 모델도 참조 표현에서 묘사하는 사물함의 위치와 속성들과는 무관하게 장면 포인트 클라우드 내의 모든 사물함들을 찾아 목표 물체들로 잘못 분할한 결과를 보여준다. 따라서 두 기존 모델들은 모두 이 제로 목표 물체 분할 작업에 실패한 것이다. 이에 반해, 제안 모델 TDMF의 경우는 정답 목표 마스크와 마찬가지로 자연어 참조 표현에 부합하는 목표 물체가 존재하지 않는다는 것을 인식하고 목표 물체 영역을 마스크로 표시하지 않았다. 이와 같은 결과를 통해, 제로 목표 물체 분할 작업에 대한 제안 모델 TDMF의 우수성을 다시 한번 확인할 수 있다.

마지막 Table 7 (c)의 경우는 보조 물체가 존재하는 환경에서 단일 목표 물체(st_w_aux)를 찾아 분할하는 작업의 하나이다. 먼저 자연어 참조 표현을 살펴보면, 갈색 의자(chair)를 뒤쪽 배경으로 두고 있으면서 이 의자와 매칭되는 갈색의 책상(desk)을 찾아 목표 물체로 분할해야 하는 작업이다. 여기서 목표 물체 단어는 책상, 보조 물체 단어는 의자가 된다. 이 장

면 포인트 클라우드 내에는 갈색 의자가 문 앞쪽과 벽쪽에 각각 하나씩 총 두 개가 존재하며, 갈색 책상 역시 문 앞쪽과 창문 바로 앞에 각각 하나씩 총 두 개가 존재한다. 참조 표현에 부합하는 목표 물체는 정답 목표 마스크가 가리키듯이 문 앞쪽의 갈색 의자 앞에 놓여있는 단 하나의 갈색 책상이어야 한다. 하지만 3D-STMN과 MDIN의 분할 결과를 살펴보면, 두 기존 모델 모두 장면 내의 모든 책상들을 목표 물체로 잘못 분할한 것을 볼 수 있다. 한편, 제안 모델 TDMF의 분할 결과를 살펴보면, 갈색 색상을 갖는 두 개의 책상들을 모두 목표 물체로 잘못 분할한 것을 알 수 있다. 이와 같이 기존 모델들뿐만 아니라 제안 모델까지 잘못된 분할 결과를 얻게 된 원인들로는 여러 가지를 추측해볼 수 있다. 그 중에서 해당 장면의 복잡성과 자연어 참조 표현의 모호성 문제도 있는 것으로 추측된다. 해당 장면에서는 보조 물체 역할을 하는 갈색 의자가 문 앞쪽에도 하나 놓여있고 벽 쪽에도 하나 놓여있다. 따라서 문 앞쪽의 의자를 기준으로 목표 물체인 갈색 책상을 찾는다면 바로 앞에 있는 책상을, 벽 쪽의 의자를 기준으로 목표 물체인 갈색 책상을 찾는다면 창문 앞의 책상이 목표 물체가 될 수

표 7. 다른 모델들과의 정성적 비교.

Table 7. Qualitative comparison with other models.

	Task	GT Result	3D-STMN	MDIN	TDMF(ours)
(a)	**task_type : mt_w_aux** "the black shelf is located by the door."				
(b)	**task_type : zt_w_aux** "these black base cabinets feature a tan top and floor-to-ceiling cabinets. they are located to the left of the sink and in the corner."				
(c)	**task_type : st_w_aux** "the desk, with its rich brown hue, is positioned in the backdrop, behind the matching brown chair."				

있다. 이와 같이 유사한 속성들을 가진 보조 물체 혹은 목표 물체들이 다수 혼재 되어 있는 복잡한 장면에서는 참조 표현에서 묘사된 목표 물체를 정확히 분할해내기는 어려울 수 있다. 또한 참조 표현의 묘사 중 “the matching brown chair” 부분은 여러 가지 의미로 해석이 가능하므로 이것을 토대로 갈색 의자와 목표 물체가 될 수 있는 갈색 책상 간의 관계를 정확하게 파악하기가 다소 어렵다. 이와 같은 복합적인 원인들로 인해 이 사례의 작업에 대해서는 기존 모델들 뿐만 아니라 제안 모델 TDMF도 모두 공통적으로 목표 물체를 정확하게 분할하기 어려웠던 것으로 추측한다. 이러한 문제점을 해결하고 현재 제안 모델 TDMF의 분할 성능을 더욱 향상시키기 위해서는 참조 표현을 더 심도 있게 이해할 수 있도록 언어적 이해 능력을 강화하는 연구와 더불어, 여러 물체들이 공존하는 복잡한 장면 포인트 클라우드에서도 참조 표현의 언어적 구성 요소들에 대응되는 시각적 요소들을 정확히 식별해낼 수 있도록 시각적 이해 능력을 더 심화하는 후속 연구들이 추가적으로 필요할 것으로 판단한다.

VI. 결론

본 논문에서는 모델 설계 이슈별로 3차원 참조 표현 분할을 위한 기존 연구들의 한계성을 살펴보고, 이러한 한계성을 극복하고자 새로운 벤치마크 데이터 집합 OAR3DRef와 이 데이터 집합을 효과적으로 이용할 수 있는 새로운 베이스라인 심층 신경망 모델인 TDMF를 제안하였다. 새롭게 제안하는 벤치마크 데이터 집합과 베이스라인 모델은 보조 물체들의 속성과 관계들을 나타내는 단어들까지 참조 표현에서 심층적으로 분리해낼 수 있도록 새로운 형식의 데이터 집합을 제공할 뿐만 아니라, 복잡한 자연어 참조 표현의 효과적인 분리를 위해 언어적 분석기와 더불어 거대 언어 모델을 함께 상호 보완적으로 활용한다. 또한 제안 모델에서는 이렇게 심층적으로 분리된 참조 표현을 모델 학습에 효과적으로 이용하기 위해, 물체들의 시각적 특징들과 특징 정렬을 수행하는 새로운 의미적 정렬 손실 함수를 정의하였다. 또한 제안 모델은 물체들의 시각적 특징에서 자연어 참조 표현의 텍스트 특징으로 또 그 반대 방향으로의 융합이 함께 수행되는 교차 어텐션 기반의 양방향 특징 융합을 수행한다. 또 제안 모델은 멀티-뷰 RGB-D 영상들로부터 추출하는 2차원 시각적 특징들도 물체들의 특징에 함께 포함시킴으로써, 다양한 색상, 텍스쳐, 모양, 크기를 가진 목표 및 보조 물체에 대한 식별과 분할 능력을 향상시켰다. 이 밖에도 제안 모델은 단일 목표 가정을 벗어나 제로 목표와 다중 목표 물체들도 함께 처리할 수 있는 예측과 추론 기능을 제공한다. 본 논문에서는 새로운 벤치마크 데이터 집합인 OAR3DRef를 이용한 다양한 정량적, 정성적 실험들을 통해, 새롭게 제안하는 베이스라인 모델의 우수성을 입증하였다.

한편, 앞서 정성적 평가 실험에서 언급한 것과 같이, 현재의 제안 모델은 장면 포인트 클라우드와 자연어 참조 표현의 복잡도가 높아질수록 다소 제한적인 분할 성능을 보여주고 있다. 따라서 향후에는 이러한 문제점을 극복하고 현재 제안 모델 TDMF의 분할 성능을 더욱 향상시키기 위해, 참조 표현에 대한 언어적 이해 능력을 더욱 강화하고 시각적 이해 능력도 더 심화할 수 있는 후속 연구들이 추가적으로 필요하다고 판단한다.

REFERENCES

[1] P. H. Huang, H. H. Lee, H.. T. Chen, and T. L. Liu, “Text-guided graph neural networks for referring 3d instance segmentation”, *Proc. of the AAAI Conference on Artificial Intelligence,* vol. 35, no. 2, pp. 1610-1618, 2021. doi: https://doi.org/10.1609/aaai.v35i2.16253

[2] C. Wu, Y. Ma, O. Chen, H. Wang, G. Luo, J. Ji, and X. Sun, “3D-STMN: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 5940-5948, 2024. doi: https://doi.org/10.1609/aaai.v38i6.28408

[3] C. Wu, Y. Liu, Y. Ma, H. Wang, G. Luo, J. Ji, H. Ding, X. Sun, and R. Ji, “3D-GRES: Generalized 3d referring expression segmentation,” arXiv preprint arXiv:2407.20664, 2024.
doi: https://doi.org/10.48550/arXiv.2407.20664

[4] Y. Wu, X. Cheng, R. Zhang, Z. Cheng, and J. Zhang, “EDA: Explicit text-decoupling and dense alignment for 3d visual grounding,” *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 19231-19242, 2023.
doi: 10.1109/CVPR52729.2023.01843.

[5] H. Lin, Y. Luo, X. Zheng, L. Li, F. Chao, T. Jin, D. Luo, Y. Wang, L. Cao, and R. Ji, “A unified framework for 3d point cloud visual grounding,” arXiv preprint arXiv:2308.11887, 2023.
doi: https://doi.org/10.48550/arXiv.2308.11887

[6] D. Z. Chen, A. X. Chang, and M. Nießner, “ScanRefer: 3d object localization in rgb-d scans using natural language,” *Proceedings of European Conference on Computer Vision (ECCV),* pp. 202-221, 2020.
doi: https://doi.org/10.1007/978-3-030-58565-5_13

[7] P. Achlioptas, A. Abdelreheem, F. X. M. Elhoseiny, and L. Guibas, “ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes,” *Proceedings of European Conference on Computer Vision (ECCV),* pp. 422-440, 2020.
doi: https://doi.org/10.1007/978-3-030-58452-8_25

[8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3D reconstructions of indoor scenes,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 5828-5839, 2017.
doi: 10.1109/CVPR.2017.261.

[9] S. He, and H. Ding, “RefMask3D: Language-guided transformer for 3d referring segmentation,” *arXiv preprint*

arXiv:2407.18244, 2024.
doi: https://doi.org/10.48550/arXiv.2407.18244

[10] Z. Qian, Y. Ma, J. Ji, and X. Sun, "X-RefSeg3D: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 38, no. 5, pp. 4551-4559, 2024.
doi: https://doi.org/10.1609/aaai.v38i5.28254

[11] Y. Zhang, Z. Gong, and A. X. Chang, "Multi3DRefer: Grounding text description to multiple 3d objects," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 15225-15236, 2023.
doi: https://doi.org/10.1109/ICCV51070.2023.01397

[12] D. He, Y. Zhao, J. Luo, T. Hui, S. Huang, A. Zhang, and S. Liu, "TransRefer3D: entity-and-relation aware transformer for fine-grained 3d Visual grounding," *Proceedings of the 29th ACM International Conference on Multimedia,* pp. 2344-2352, 2021.
doi: 10.48550/arXiv.2108.02388.

[13] Z. Yuan, X. Yan, Y. Liao, R. Zhang, S. Wang, Z. Li, and S. Cui, "InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 1791-1800, 2021.
doi: 10.1109/ICCV48922.2021.00181.

[14] M. Feng, Z. Li, Q. Li, L. Zhang, X. D. Zhang, G. Zhu, H. Zhang, Y. Wang, and A. Mian, "Free-form description guided 3d visual graph network for object grounding in point cloud," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 3722-3731, 2021.
doi: 10.1109/ICCV48922.2021.00370.

[15] S. H. Song, and I. C. Kim, "Transformer-based 3D instance segmentation with auxiliary denoising learning," *Journal of Institute of Control, Robotics and Systems (in Korean),* vol. 29, no. 12, pp. 954-965, 2023.
doi: 10.5302/J.ICROS.2023.23.0150.

[16] S. Schuster, R. Krishna, A. Chang, L. F. Fei, C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," *Proc. of the Fourth Workshop on Vision and Language,* pp. 70-80, 2015.
doi: https://doi.org/10.18653/v1/W15-2812

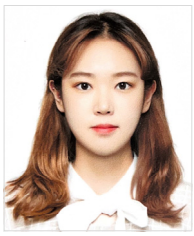
[17] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, and W. Y. Ma, "Unified visual-semantic embeddings: bridging vision and language with structured meaning representations," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),* pp. 6609-6618, 2019.
doi: https://doi.org/10.1109/CVPR.2019.00677

[18] B. Graham, M. Engelcke, and L. V. D. Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),* pp. 9224-9232, 2018.
doi: https://doi.org/10.1109/CVPR.2018.00961

[19] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. S. Torr, "LAVT: Language-aware vision transformer for referring image segmentation," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 18155-18165, 2022.
doi: https://doi.org/10.1109/TPAMI.2024.3468640

[20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," arXiv preprint arXiv:2203.02155, 2022.
doi: https://doi.org/10.48550/arXiv.2203.02155

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692,* 2019.
doi: https://doi.org/10.48550/arXiv.1907.11692



배 혜 림

2023년 경기대 컴퓨터공학부 졸업. 2025년 경기대 일반대학원 컴퓨과학과 석사. 관심분야는 인공지능, 기계학습, 컴퓨터비전, 로봇지능.

박 경 민

2024년 경기대 컴퓨터공학부 졸업. 2024년~현재 경기대 일반대학원 컴퓨과학과 석사과정. 관심분야는 인공지능, 로봇지능, 컴퓨터비전, 상식추론.

김 인 철

1985년 서울대 수학과 졸업. 1987년 동 대학원 이학석사. 1995년 동 대학원 이학박사. 1996년~현재 경기대학교 컴퓨터공학부 교수. 관심 분야는 인공지능, 로봇지능, 학습 및 지식추론.