KIPS Trans. Softw. and Data Eng.
Vol.12, No.9 pp.407~418
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.9.407

Multi-Object Goal Visual Navigation Based on Multimodal Context Fusion

Jeong Hyun Choi† · In Cheol Kim††

ABSTRACT

The Multi-Object Goal Visual Navigation(MultiOn) is a visual navigation task in which an agent must visit to multiple object goals in an unknown indoor environment in a given order. Existing models for the MultiOn task suffer from the limitation that they cannot utilize an integrated view of multimodal context because use only a unimodal context map. To overcome this limitation, in this paper, we propose a novel deep neural network-based agent model for MultiOn task. The proposed model, MCFMO, uses a multimodal context map, containing visual appearance features, semantic features of environmental objects, and goal object features. Moreover, the proposed model effectively fuses these three heterogeneous features into a global multimodal context map by using a point-wise convolutional neural network module. Lastly, the proposed model adopts an auxiliary task learning module to predict the observation status, goal direction and the goal distance, which can guide to learn the navigational policy efficiently. Conducting various quantitative and qualitative experiments using the Habitat-Matterport3D simulation environment and scene dataset, we demonstrate the superiority of the proposed model.

Keywords : Multi-Object Goal Visual Navigation, Deep Reinforcement Learning, Multimodal Context Fusion, Global Mapping

멀티모달 맥락정보 융합에 기초한 다중 물체 목표 시각적 탐색 이동

최 정 현† · 김 인 철††

요 약

MultiOn(Multi-Object Goal Visual Navigation)은 에이전트가 미지의 실내 환경 내 임의의 위치에 놓인 다수의 목표 물체들을 미리 정해준 일정한 순서에 따라 찾아가야 하는 매우 어려운 시각적 탐색 이동 작업이다. MultiOn 작업을 위한 기존의 모델들은 행동 선택을 위해 시각적 외관 지도나 목표 지도와 같은 단일 맥락 지도만을 이용할 뿐, 다양한 멀티모달 맥락정보에 관한 종합적인 관점을 활용할 수 없다는 한계성을 가지고 있다. 이와 같은 한계성을 극복하기 위해, 본 논문에서는 MultiOn 작업을 위한 새로운 심층 신경망 기반의 에이전트 모델인 MCFMO (Multimodal Context Fusion for MultiOn tasks)를 제안한다. 제안 모델에서는 입력 영상의 시각적 외관 특징외에 환경 물체의 의미적 특징, 목표 물체 특징도 함께 포함한 멀티모달 맥락 지도를 행동 선택에 이용한다. 또한, 제안 모델은 점-단위 합성곱 신경망 모듈을 이용하여 3가지 서로 이질적인 맥락 특징들을 효과적으로 융합한다. 이 밖에도 제안 모델은 효율적인 이동 정책 학습을 유도하기 위해, 목표 물체의 관측 여부와 방향, 그리고 거리를 예측하는 보조 작업 학습 모듈을 추가로 채용한다. 본 논문에서는 Habitat-Matterport3D 시뮬레이션 환경과 장면 데이터 집합을 이용한 다양한 정량 및 정성 실험들을 통해, 제안 모델의 우수성을 확인하였다.

키워드 : 다중 물체 목표 시각적 탐색 이동 작업, 심층 강화학습, 멀티모달 맥락정보 융합, 전역 지도 작성

1. 서 론

최근 들어 자연어 이해 및 시각 지능 기술들이 발전함에 따라 가상 및 실세계에 놓여 종합적인 실시간 지능을 발휘해야 하는 체화 인공지능(Embodied AI, EAI)에 관한 관심이 더욱 높아졌다. 기존의 지능형 서비스 로봇이나 자율 주행차, 컴퓨터 게임 NPC(Non-Player Character), 증강 및 가상 현실 등도 모두 체화 인공지능 기술이 적용될 수 있는 대표적인 응용 영역들이다. 체화 인공지능을 위해 가장 활발하게 연구되어 오고 있는 작업 중 하나인 시각적 탐색 이동(visual navigation) 작업은 에이전트가 실시간 입력 영상에 의존하여 연속된 이동 동작(navigation action)들을 수행함으로써 환경 내에 놓인 특정 목표 지점(point)이나 목표 물체(object)를 찾아가

※ 본 연구는 정보통신기획평가원의 재원으로 정보통신방송 기술개발사업의 지원을 받아 수행한 연구 과제(No. 2020-0-00096 클라우드에 연결된 개별 로봇 및 로봇그룹의 작업 계획 기술 개발)입니다.
† 비 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 종신회원 : 경기대학교 AI컴퓨터공학부 교수
Manuscript Received : June 29, 2023
First Revision : August 7, 2023
Accepted : August 24, 2023
* Corresponding Author : In Cheol Kim(kic@kyonggi.ac.kr)

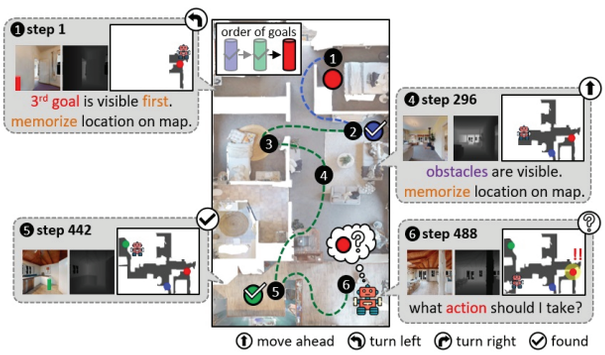

Fig. 1. An Example of Multi-Object Goal Visual Navigation

야 하는 작업이다. 이때 목표 지점은 주로 좌표 형태로, 목표 물체는 물체를 나타내는 영상이나 자연어 묘사 형태로 에이전트에게 주어진다. 특히 본 논문에서 다룰 다중 물체 목표 시각적 탐색 이동(Multi-Object Goal Visual Navigation, MultiOn) [1] 작업은 에이전트가 미지의 실내 환경 내 임의의 위치에 놓인 다수의 목표 물체들을 미리 정해준 일정한 순서에 따라 방문해야 하는 매우 어려운 시각적 탐색 이동 작업이다. Fig. 1은 하나의 다중 물체 목표 시각적 탐색 이동(MultiOn) 작업과 이를 수행하기 위한 에이전트의 입력 영상, 그리고 에이전트 내부에서 생성 중인 환경 지도 등을 나타내는 예시이다.

Fig. 1에서 볼 수 있듯이, 하나의 MultiOn 작업에서 에이전트가 찾아가야 할 목표들은 여러 개의 방과 가구들로 구성된 실내 환경 내에 임의의 위치에 놓여진 서로 다른 색상의 실린더(cylinder)들로 주어지며, 에이전트는 미리 정해진 일정한 순서에 따라 이들을 차례대로 방문해야 한다. 작업을 수행하는 동안 매 시간 단계(time step)마다 에이전트에게는 환경에 대한 RGB-D 영상이 실시간으로 입력되고, 이것에 대응하여 에이전트는 매번 {move ahead, turn left, turn right, found}의 4가지 이산화된 이동 행동들 중 하나를 선택하여 실행함으로써 목표 물체들까지 도달해야 한다. 이와 같은 일련의 과정을 통해, 제한 시간 내에 에이전트가 목표 물체들을 주어진 순서대로 모두 방문하였을 때, 해당 MultiOn 작업은 성공적으로 달성된 것으로 간주한다. Fig. 1의 예에서는 에이전트가 ①의 위치에서 출발하여 ②, ⑤, ①의 위치에 놓인 목표들을 순서대로 방문해야 하는 MultiOn 작업을 나타낸다. 이 예에서 에이전트는 이미 ①, ②, ③, ④, ⑤의 경로를 거쳐 ②, ⑤의 위치에 놓인 2개의 목표들은 방문하였고, 현재 ⑥의 위치에서 마지막으로 ①의 위치에 놓인 목표를 찾아 방문해야 하는 상황을 나타낸다. 만약 에이전트가 이 작업의 시작 위치인 ①에서 향후 방문해야 하는 마지막 목표를 미리 탐지하여 환경 지도에 저장하고 있다면, 해당 목표를 찾기 위한 불필요한 추가 탐험(exploration)없이 곧바로 해당 목표를 향해 이동해갈 수 있을 것이다.

이와 같이 일반적으로 한 에이전트가 매 순간 실시간 입력 영상을 토대로 전체 환경 구성과 목표 물체의 위치, 그리고 자신의 현재 위치와 같은 다양한 맥락정보들을 추출해내어 자신의 행동 선택에 효과적으로 활용할 수 있다면 MultiOn 작업을 보다 효율적으로 달성할 수 있을 것이다. 따라서 MultiOn 작업을 위한 에이전트 모델 설계에서 가장 중요한 이슈 중 하나는 실시간 입력 RGB-D 영상으로부터 어떤 맥락정보들을 추출하여 행동 선택에 이용할 것인가를 결정하는 문제이다. 이 문제에 관한 기존 모델들의 접근 방식은 (1) 단일 맥락 지도 이용 방식[1-3]과 (2) 멀티모달 맥락 지도 이용 방식[4, 5]으로 크게 나누어 볼 수 있다. 단일 맥락 지도를 이용하는 기존의 모델들은 대부분 RGB-D 입력 영상에서 추출하는 시각적 외관 특징 지도(visual appearance feature map)나 혹은 입력 영상에서 탐지한 목표 물체를 표시한 목표 지도(goal map)만을 행동 선택에 주로 이용하였다. 그러나, 이러한 단일 맥락 지도로는 에이전트가 복잡한 작업 환경 구성과 작업 상황을 다양한 관점에서 이해하기 어렵고, 이로 인해 목표 물체들에 도달하기 위한 효율적인 이동 정책(navigational policy)을 학습하기 어렵다는 한계점이 있다. 이러한 한계점을 극복하기 위해 입력 영상에서 추출하는 서로 다른 맥락정보를 함께 포함한 멀티모달 맥락 지도를 이용하는 방식들[4, 5]이 제안되었다. 하지만 기존 모델들[4, 5]에서는 점유 지도와 목표 지도만을 함께 이용하는 등 맥락정보들의 다양성과 관계성에 제한이 존재하며, 서로 다른 맥락정보들 간의 상호 보완성을 보장할 수 없는 등의 문제점을 가지고 있다.

한편, MultiOn 작업을 위한 에이전트 모델 설계에서 또 다른 중요한 이슈는 실시간 입력 영상에서 추출하는 서로 다른 맥락정보들을 어떤 방식으로 효과적으로 융합할 것인가하는 문제이다. 기존의 SGoLAM 모델[4]의 경우 MultiOn 작업을 위해 서로 다른 2가지 종류의 목표 지도(goal map)와 점유 지도(occupancy map)를 이용하였지만, 특별히 이들을 하나로 융합하지 않고 각기 독립적으로 목표까지 경로 계획 수립(path planning)과 목표 탐지를 위한 탐험(exploration)에 활용하였다. 한편, EXPO[5] 모델의 경우는 목표 지도와 점유 지도를 서로 단순 연결(concatenate)하여 하나의 환경 지도를 구성하였다. 이 경우 본래 두 지도가 보유하고 있는 서로 다른 맥락정보가 그대로 독립적으로 관리되고 활용됨으로써 맥락정보들 간의 심층적 융합을 통한 상호보완적 효과는 기대하기 어렵고, 이들을 저장하기 위한 메모리 사용에도 비효율성이 존재한다.

이와 같은 기존 모델들의 한계를 고려하여, 본 논문에서는 다중 물체 목표 시각적 탐색 이동 작업을 위한 새로운 에이전트 모델을 제안한다. 제안 모델 MCFMO(Multimodal Context Fusion for MultiOn tasks)은 입력 영상에서 추출하는 시각적 외관 특징외에 환경 물체의 의미적 특징, 목표 물체 특징을 함께 포함하는 멀티모달 맥락 지도를 행동 선택에 이용한다. 또한, 제안 모델은 다양한 맥락정보 간의 관계를 충분히 활용하기 위해 점-단위 합성곱 신경망(point-wise convolution neural network)을 이용하여 3가지 서로 이질적인 특징들을 하나의 전역 지도로 효과적으로 융합한다. 이 밖에도 효율적인 이동 정책 학습을 유도하기 위해 목표의 관측 여부와 방향

및 거리를 예측하는 보조 작업 학습 모듈을 추가로 채용한다. 본 논문에서는 제안 모델의 성능을 분석하기 위해, 3차원 시뮬레이터인 Habitat와 벤치마크 실내 환경 데이터 집합인 Matterport3D를 이용하여 다양한 실험들을 수행하고, 그 결과를 소개한다. 본 논문의 2장에서는 기존의 관련 선행 연구들을 살펴보고, 3장에서는 제안 모델의 설계에 대해 자세히 설명한다. 4장에서는 제안 모델의 구현과 실험에 관해 소개하고, 5장에서는 결론과 향후 연구를 정리한다.

2. 관련 연구

2.1 전역적 맥락 지도

MultiOn 작업에서 에이전트가 전역적 맥락 지도(global context map) 혹은 간단히 전역적 지도(global map) 작성을 통해 이동 작업을 수행해야 하는 환경과 현재 상황을 인식하는 것은 매우 중요하다. 로봇 또는 에이전트 분야에서 이동 작업을 위해 전통적으로 많이 이용되어 온 전역적 지도로는 점유 지도(occupancy map), 시각적 외관 특징 지도(visual appearance feature map), 그리고 의미적 지도(semantic map)가 있다. 이는 다시 단일 맥락 지도 이용 방식[1-3]과 멀티모달 맥락 지도 이용 방식[4, 5, 10-13]으로 구분된다.

먼저, 단일 맥락 지도 이용 방식은 탐색 이동 과정에서 얻은 한 가지 정보만을 표현한 전역적 지도를 사용하는 방식이다. 깊이 영상을 기반으로 그리드(grid) 내 각 셀(cell) 혹은 그래프(graph) 내 각 노드의 점유 여부를 나타내는 점유 지도는 시각적 탐색 이동 작업에서 휴리스틱 탐색 알고리즘(heuristic search algorithm) 중 하나인 A* 또는 빠른 전진 방식(fast marching method) 등 경로 계획 알고리즘과 결합되어 주로 사용되었다[8, 9]. 점유 지도는 에이전트 위치 및 이동 가능 영역 정보를 제공하지만, 복잡한 환경에서 세부적인 환경 구성 물체 등의 정보를 표현하기는 어렵다.

이를 보완하기 위해 입력 영상에서 신경망으로 추출하는 시각적 외관 특징 지도가 제안되었다[1, 2]. ProjNeuralMap[1], AuxTaskMap[2]와 같은 초기 MultiOn 작업 모델들에서는 입력 RGB-D 영상으로부터 합성곱 신경망(CNN)을 통해 추출한 결과를 투영한 시각적 외관 특징 지도(visual appearance feature map)를 작성하였다. 이와 같은 시각적 외관 특징 지도는 환경 장면의 질감, 색상 등과 같은 다양한 시각적 특징 정보를 표현할 수 있지만, 구체적으로 작업 환경에 놓여있는 구성 물체들과 그들 간의 공간적 관계와 같은 고수준의 의미적 정보는 제공해줄 수 없다.

한편, 심층 신경망 기반의 컴퓨터 비전(computer vision) 기술이 발전함에 따라, 물체 탐지(object detection) 및 영상 분할(image segmentation) 신경망 모델을 활용하여 입력 영상에서 환경 구성 물체 또는 목표 물체의 레이블(label) 정보를 추출하여 의미적 지도를 생성하고 이들을 로봇 및 에이전트 이동 작업에 이용하는 연구가 활발하게 진행되고 있다[1, 3]. ObjRecogMap[1]과 Modular-MON[3] 같은 MultiOn 모델들은 각각 1계층의 완전 연결 계층 혹은 Faster RCNN 기반의 목표 예측/탐지 모듈을 통해 입력 RGB 영상 내의 목표를 파악하고, 이를 토대로 목표 중심의 의미적 지도를 생성함으로써 목표와의 위치 관계를 집중적으로 파악하려고 하였다. 하지만, 대부분의 의미적 지도는 환경에 등장하는 소수의 주요 물체들은 표현할 수는 있지만, 아직은 물체 인식 정확도와 신뢰도가 높지 않고 환경의 나머지 상세 요소들이나 물체들의 재료, 질감, 밝기 등을 표현하기는 어렵다는 한계를 갖는다.

한편, 이러한 단일 맥락 지도들의 한계성을 극복하고자, 두 종류 이상의 서로 다른 맥락정보들을 함께 표현하는 멀티모달 맥락 지도를 이용하려는 모델들도 등장하였다[4, 5, 10-13]. SGoLAM[4]은 RGB 영상으로부터 생성한 목표 지도와 깊이 영상으로부터 생성한 점유 지도를 함께 이용하였다. 또, EXPO[5]는 입력 영상에서 실제로 인식 모듈을 적용해 목표 지도와 점유 지도를 생성하는 대신에 정답 목표 지도(ground truth goal map)와 정답 점유 지도(ground truth occupancy map)를 이동 작업에 이용하였다. 하지만 이러한 기존 모델들[4, 5]에서는 이용하는 맥락정보들의 다양성과 관계성에 제한이 존재하며, 서로 다른 맥락정보들 간의 상호 보완성을 보장할 수 없는 등의 문제점을 가지고 있다.

2.2 멀티모달 맥락정보 융합

서로 다른 다수의 이질적인 맥락정보들로부터 효과적인 멀티모달 맥락 지도를 작성하기 위해서는 맥락정보 각각의 특성과 함께 이들 간의 연관성이 잘 반영될 수 있는 방식으로 융합하는 것이 매우 중요하다. 초기의 시각적 탐색 이동(visual navigation) 작업에서는 환경에 대한 부분적인 관측 영상으로부터 추출한 지역적 멀티모달 맥락정보를 기준으로 이질적인 정보들을 융합하기 위해 단순 연결(concatenate)[14], 가중곱(weighted multiply)[15], 3계층 합성곱(convolution neural network)과 같은 융합 방식[16]들이 주로 적용되었다.

이후 환경 크기가 확장되어 전역적 지도를 사용하는 일부 물체 목표 탐색 이동(Object Goal Visual Navigation) 작업을 위해서도 단순 연결[10-12] 및 선형 결합[13] 방식으로 복수의 전역적 멀티모달 맥락 지도들을 융합하려는 시도들이 나타났다. 단순 연결 방식의 WS-MGMap[12] 모델은 U-Net을 통해 추출한 시각적 외관 특징 지도와 환경 물체 레이블 기반 의미적 지도를 단순 연결한 멀티모달 맥락 지도를 사용하였다. 이러한 경우, 본래 두 지도가 보유하고 있는 서로 다른 맥락정보가 그대로 독립적으로 관리되고 활용됨으로써 맥락정보들 간의 심층적 융합을 통한 상호보완적 효과는 기대하기 어렵고, 이들을 저장하기 위한 메모리 사용에도 비효율성이 존재한다.

이러한 한계를 극복하기 위해 각 지도를 임의의 가중치를 토대로 더하는 선형 결합(linear combination) 방식이 제안되었다. PONI[13]는 에이전트의 미 탐색 영역의 위치를 예측하는 잠재적 영역(area potential) 지도와 목표 위치를 예측한 잠재적 물체(object potential) 지도 2종류의 시각적 외관 특

징 지도를 $\alpha(=0.5)$와 $1-\alpha(=0.5)$의 비율로 더하여 사용하였다. 그러나 환경 구성 물체의 경계, 형태, 상대적인 위치 등 다양한 맥락을 고려한 가중치 설정이 어려우며, 이러한 복잡한 패턴에 대한 비선형 관계를 표현하는 데에 한계가 있다.

한편, 다수의 목표 물체들이 포함된 MultiOn 작업에서도 멀티모달 맥락 지도를 이용하는 모델들[4, 5]이 제안되었으나, 융합 기술과 관련한 깊은 연구는 이루어지지 않았다. SGoLAM 모델[4]의 경우 MultiOn 작업을 위해 서로 다른 2가지 종류의 목표 지도(goal map)와 점유 지도(occupancy map)를 이용하였지만, 특별히 이들을 하나로 융합하지 않고 각기 독립적으로 목표까지 경로 계획 수립(path planning)과 목표 탐지를 위한 탐험(exploration)에 활용하였다. 한편, EXPO[5] 모델의 경우는 목표 지도와 점유 지도를 서로 단순 연결(concatenate)하여 하나의 환경 지도를 구성하였다. 앞서 [10-12] 연구들과 동일하게 다양한 맥락정보 간의 깊은 관계를 충분히 활용할 수 없다는 문제와 메모리 비효율성의 한계를 보인다.

3. 다중 물체 목표 시각적 탐색 이동 모델

3.1 모델 개요

Fig. 2는 MultiOn 작업을 위한 제안 모델의 전체 구성을 나타낸다. 제안 모델은 매시간 단계마다 에이전트 시점의(agent-centric view) RGB-D 입력 영상 O_t과 직전에 실행한 행동 a_{t-1}을 기초로, 현재 방문해야 할 목표 g_t에 도달하기 위한 최적의 행동 a_t를 결정하고 실행한다. 제안 모델은 크게 공간적 맥락정보 임베딩(Spatial Context Embedding, SCE), 시간적 맥락정보 임베딩(Temporal Context Embedding, TCE), 행동 예측(Action Prediction, AP), 그리고 보조 작업 학습을 위한 목표 예측(Goal Prediction, GP) 모듈들과 같이 총 4개의 모듈들로 구성된다.

Fig. 2에서 보듯이, 공간적 맥락정보 임베딩(SCE) 모듈은 다시 지역적 맥락정보 임베딩(Local Context Embedding, LCE)와 전역적 맥락정보 임베딩(Global Context Embedding, GCE)의 서브 모듈로 구분된다. 지역적 맥락정보 임베딩(LCE)은 실시간 RGB-D 입력 영상 O_t에서 3가지 서로 다른 지역 맥락정보들(f^l_{vis}, f^l_{sem} f^l_{goal})을 추출한 뒤 이들을 하나의 지역적 맥락정보 특징 f^l_t으로 융합하는 역할을 수행하며, 전역적 맥락정보 임베딩(GCE)은 이동 작업동안 이러한 실시간 맥락정보 특징 f^l_t들을 차례로 모아 하나의 전역적 지도로 확장해가면서 이를 임베딩하여 전역적 맥락정보 특징 f^G_t를 생성하는 역할을 수행한다. 제안 모델의 공간적 맥락정보 임베딩(SCE) 모듈은 이 두 서브 모듈들이 생성하는 지역적, 전역적 특징들인 f^l_t와 f^G_t외에 추가로 현재 목표 g_t와 이전 행동 a_{t-1}을 나타내는 목표 특징 f^g_t, 행동 특징 f^a_t을 각각 구해낸 뒤, 이들을 모두 단순 연결(concatenate)하여 t시간 단계에서의 공간적 맥락정보 특징 f_t를 생성한다.

시간적 맥락정보 임베딩(TCE) 모듈에서는 공간적 맥락정보 임베딩(SCE) 모듈이 생성하는 현재 시간 단계의 맥락정보 특징인 f_t를 토대로, 에이전트의 연속적인 이동 행동들로부터 얻어지는 시간 순서에 따른 맥락정보를 반영하기 위해 순환신경망(Recurrent Neural Network, RNN)의 하나인 LSTM(Long Short-Term Memory)을 통해 이전 시간 단계들의 과거 맥락정보가 충분히 반영된 시공간적 맥락정보 특징 f^*_t를 구해낸다.

제안 모델의 행동 예측(AP) 모듈에서는 이렇게 구해진 시공간적 맥락정보 특징 f^*_t를 기초로 에이전트의 행동 정책 π_t을 학습하고 현재 시간 단계에서 수행할 행동 a_t을 결정한다. 행동 정책 π_t의 학습은 행동 a_t의 실행에 따른 보상 r_t과 PPO(Proximal Policy Optimization) 강화학습 알고리즘을 토대로 행위자(actor)와 비평가(critic) 간의 상호작용을 통해 이루어진다. 행위자와 비평가는 각각 1계층의 완전 연결 계층(fully connected layer)으로 구성된다.

마지막으로 목표 예측(GP) 모듈은 제안 모델의 효율적인 정책 학습을 돕고 작업 성공률을 향상시키기 위해, 모델 훈련 단계에만 이용되는 보조 작업 모듈(auxiliary task module)이다. MultiOn 작업의 특성상 환경 안에서 현재 방문해야 할

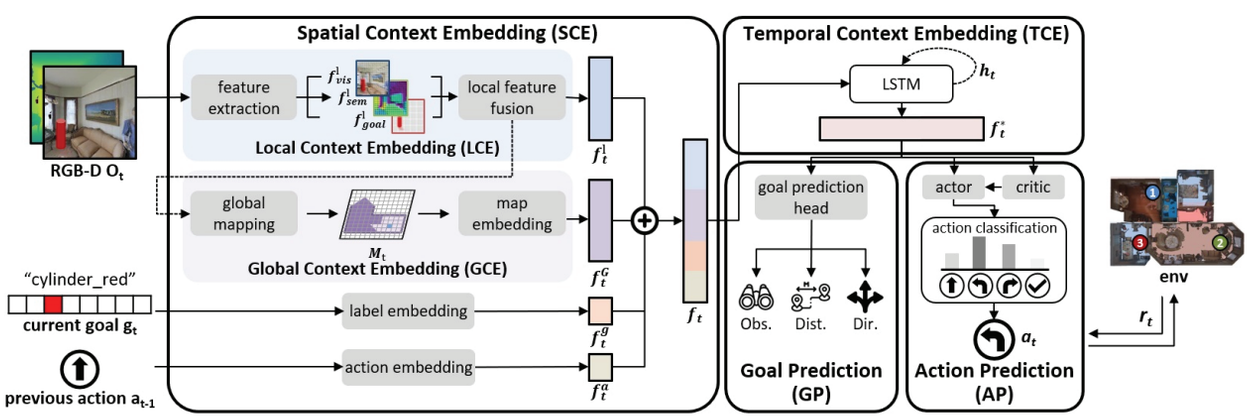

Fig. 2. Architecture of the Proposed MCFMO Model

목표 g_t를 신속히 찾아내는 일이 무엇보다 중요하다. 특히 목표 g_t가 에이전트의 현재 시야나 혹은 전역적 지도에 들어있지 않은 경우에는 목표 g_t를 찾아내기 위한 탐험 과정이 필수적으로 요구되고, 이때 아직 발견하지 못한 목표 g_t가 놓여있을 위치나 방향을 효과적으로 예측할 수 있다면 불필요한 탐험 과정을 줄이고 작업을 신속히 완료할 수 있을 것이다. 실제로 많은 MultiOn 작업들에서 이러한 미탐지 목표를 찾기 위한 탐험 과정으로 작업 시간의 상당 부분을 소모되는 경향이 많다. 이 점을 고려하여 제안 모델의 목표 예측 모듈에서는 맥락정보 특징 f_t^*로부터 목표 g_t의 관측 여부, 목표의 방향, 그리고 목표까지 거리를 각각 예측하는 3가지 목표 예측 분류기들을 포함한다. 각 분류기의 예측값 $\widehat{obs}_t$, $\hat{\phi}_t$, $\hat{d}_t$과 정답(ground truth) $^{obs}_t$, ϕ_t^*, d_t^* 간의 크로스 엔트로피 손실(cross-entropy loss)들인 L_{obs}, L_{dir}, L_{dist}에 가중치(weight)들을 부여해 결합함으로써, 이 보조 작업 학습을 위한 손실 $L_{AUX} = \lambda_{obs}L_{obs} + \lambda_{dir}L_{dir} + \lambda_{dist}L_{dist}$를 구한다.

3.2 멀티모달 지역적 맥락정보 융합

제안 모델의 지역적 맥락정보 임베딩(LCE) 단계에서는 실시간 RGB-D 입력 영상 O_t에서 작업 환경에 관한 시각적 외관 특징(visual appearance feature), 환경 구성 물체들을 나타내는 의미적 특징(semantic feature), 탐지된 목표 물체를 나타내는 목표 특징(goal feature) 등 서로 다른 맥락정보 특징들을 추출해내고, 이들을 점-단위 합성곱 신경망(point-wise convolution neural network)으로 융합하여 멀티모달 지역적 맥락정보 특징 f_t^l을 구해낸다. Fig. 3은 지역적 맥락정보 임베딩 과정의 세부 단계들을 나타내며, 이 과정은 특징 추출(Feature Extraction) 단계와 지역적 특징 융합(Local Feature Fusion) 단계로 구성된다. 특징 추출 단계는 RGB-D 입력 영상 O_t에서 시각적 외관 특징 f_{vis}^l, 의미적 특징 f_{sem}^l, 목표 특징 f_{goal}^l들을 추출한다. 시각적 외관 특징 f_{vis}^l 추출에는 3계층(layer)의 합성곱 신경망(CNN)을 이용하고, 환경 구성 물체들을 나타내는 의미적 특징 f_{sem}^l 추출을 위해서는 대표적인 영상 분할 모델인 U-Net과 3계층의 합성곱 신경망을 이용한다.

이때, 영상 분할을 위한 U-Net은 Matterport3D 실내 환경 데이터 집합을 이용하여 사전 학습(pre-training)시킨 후 이용하였다.

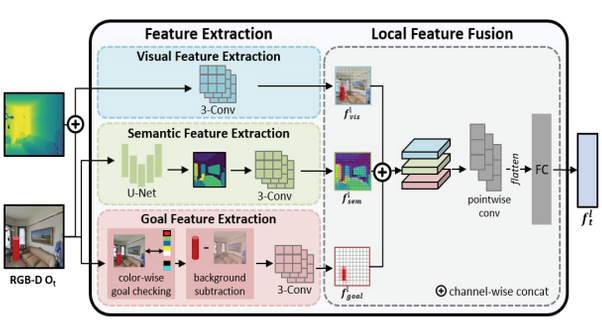

Fig. 3. Local Context Embedding

한편, 목표 특징 추출을 위해서는 8개의 목표 실린더 집합의 고유 색상값을 토대로 입력 RGB 영상 내 각 목표 실린더의 위치를 알아내는 색상별 목표 확인(color-wise goal checking) 단계와 입력 RGB 영상에서 탐지된 목표 실린더들을 제외한 나머지 영역은 모두 0으로 삭제하는 배경 제거(background subtraction) 단계를 거친 후, 결과 영상에서 3계층의 합성곱 신경망을 통해 목표 특징 f_{goal}^l을 구해낸다. 특히, 색상별 목표 확인 단계에서는 8개의 목표 실린더 집합마다 고유 색상값을 사전에 정의하고, RGB 영상과 픽셀 단위로 비교하여 관측 영상 내 목표 실린더의 위치를 파악한다. 이때, 오탐지를 줄이기 위해 실린더라고 판단한 픽셀들 중 이웃한 픽셀들을 연결하였을 때 크기가 임계치 δ 미만인 영역은 제거한다.

제안 모델의 지역적 특징 융합 단계에서는 RGB-D 입력 영상에서 추출된 서로 다른 맥락정보 특징들인 f_{vis}^l, f_{sem}^l, f_{goal}^l들을 Equation (1)과 같은 연산을 통해 심층적으로 융합한다.

$$f_t^l = W * \delta(F \circledcirc (f_{vis}^l \sqcup f_{sem}^l \sqcup f_{goal}^l)) + b \tag{1}$$

Equation (1)에서 $\sqcup$ 은 채널 단위 연결(channel-wise concatenate)을, F은 점-단위 합성곱 신경망 계층을, δ는 ReLU 함수를, $\circledcirc$ 은 행렬곱을, 그리고 W와 b는 완전 연결 계층의 가중치(weight)와 편향(bias)을 각각 의미한다. 특히, 1×1 크기의 커널을 통해 채널 방향의 합성곱 연산을 진행하는 점-단위 합성곱을 통해 다양한 관점에서의 이질적인 특징들 간의 관계를 반영하며, 이때 가중치를 학습하여 유연한 특징 융합이 가능하다. 해당 과정을 통해 3가지 정보가 융합된 지역적 맥락정보 특징 f_t^l를 구해낸다.

3.3 전역적 공간 맥락정보 임베딩

제안 모델의 전역적 맥락정보 임베딩(GCE) 단계에서는 실시간 입력 영상에서 추출한 지역적 멀티모달 맥락정보 특징 f_t^l들을 모아서 작업 환경 전체 공간을 나타내는 전역적 맥락 지도(global context map)를 생성하고, 이것을 토대로 전역적 맥락정보 특징 f_t^G을 구해낸다. Fig. 4는 전역적 맥락정보 임베딩 과정의 세부 단계들을 나타낸다. 이 과정은 전역적 지도 작성(Global Mapping)과 지도 임베딩(Map Embedding) 단계로 구성된다. 먼저, 전역적 지도 작성 단계는 멀티모달 맥락정보 특징 f_t^l로부터 전역적 맥락 지도 M_t를 구해내기 위해 다시 지면 투영(Ground Projection)과 등록(Registration)의 세부 단계로 나뉜다.

첫 번째로, 지면 투영 단계에서는 3차원 형태의 멀티모달 맥락정보 특징 f_t^l의 픽셀 (i,j)들로부터 Equation (2)와 같은 역-투영(inverse projection)을 통해 시뮬레이션 상의 실제 에이전트 위치에 대응하는 좌표 (X, Y, Z)을 구한다.

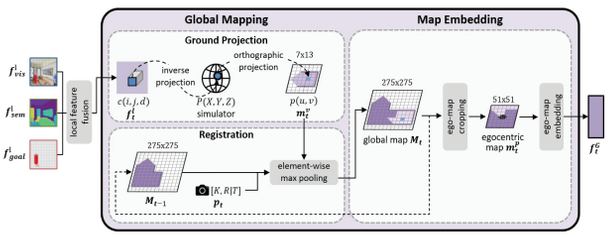

Fig. 4. Global Context Embedding

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = d_{i,j} R^{-1} K^{-1} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} - T \tag{2}$$

Equation (2)에서 $d_{i,j}$는 픽셀 (i,j)의 깊이를 의미하고, K, R, T는 각각 카메라의 내부 행렬과 회전 행렬, 그리고 변환 행렬을 의미한다.

두 번째로는, 실제 위치 좌표 (X, Y, Z)를 Equation (3)과 같이 정사영(orthographic projection)시켜 지도의 셀 (u, v)에 표시한다. 위의 두 과정을 통해 에이전트가 하단 간선의 중앙에 위치하는 에이전트 중심의 지역적 지도(agent-centric local map) m_t^v를 작성한다.

$$\begin{bmatrix} u \\ v \\ 0 \\ 1 \end{bmatrix} = K[R|T] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3}$$

마지막으로, 등록 과정에서는 $t-1$시간 단계에서의 전역 지도 M_{t-1}과 에이전트 시야 중심 지도 m_t^v를 통합하여 t시간 단계에서의 전역적 맥락 지도 M_t를 작성한다. 이때, 전역적 맥락 지도 내에서 에이전트의 위치 및 회전값 p_t를 기준으로 m_t^v를 회전 및 위치시킨 다음, 요소별 최대 풀링(element-wise max pooling) 과정을 통해 지도를 통합한다.

지도 임베딩 과정에서는 전역적 지도 M_t를 에이전트 위치 p_t를 중심으로 일정 크기로 자른 정사각형의 에이전트 중앙의 지역적 지도(agent-centered local map) m_t^p를 구해낸다. 이를 통해 효율적으로 에이전트 이동 범주 내 환경을 중점적으로 참조할 수 있다. m_t^p는 m_t^v와 다르게 에이전트가 지도의 중앙에 위치함을 가정한다. 이후 3계층의 합성곱 신경망과 1계층의 완전 연결 계층으로 구성된 에이전트 중심 지도 임베딩(ego-map embedding)을 통해 전역적 맥락정보 특징 f_t^G로 표현된다.

3.4 모델 학습

제안 모델 MCFMO에서 입력 영상의 의미적 분할을 위해 사용한 U-Net은 합성곱 신경망 블록과 스킵 연결(skip connection)로 구성된 4계층의 인코더(encoder) 및 4계층의 디코더(decoder)의 구조를 갖는다. 환경 내 존재하는 {배경, 벽, 기둥, 바닥, 천장, 의자, 테이블, 침대, 선반, 소파, 계단, 싱크대, 변기, 욕조, 캐비닛, 문, 스툴(stool)}의 주변 물체 17가지에 대해 예측하도록 학습하였다. U-Net 모델의 학습 데이터를 수집하기 위해 61개의 Matterport3D 실내 환경 내 물체를 목표로 하는 작업을 무작위로 정의하였으며, 에이전트가 탐색 이동 작업을 수행하는 과정에서 각 환경마다 관측 RGB 영상과 정답 분할 영상의 쌍을 2 만개씩 수집하였다. 학습을 위한 손실 함수는 예측 분할 영상과 정답 분할 영상 간의 픽셀 단위 크로스 엔트로피 손실(cross-entropy loss)을 사용하였으며, 총 2M 시간 단계만큼 학습을 진행하였다. 이후 전체 모델과 함께 사용 시에는 U-Net의 모든 가중치를 고정하였다.

제안 모델 MCFMO 전체의 종단 간(end-to-end) 학습은 아래와 같이 행동 정책 학습을 위한 PPO 손실 L_{PPO}과 목표 예측 보조 작업 손실 L_{AUX}을 줄이는 방향으로 진행된다. PPO 강화학습은 샘플링(sampling) 단계와 최적화(optimization)의 두 단계의 반복으로 진행된다. 샘플링 단계는 k시간 단계에서 고정된 최신 정책 π_θ로부터, T 길이를 가지는 에이전트의 경로 궤적 τ들을 수집하여 집합 U_k를 구성한다. 이때, θ는 정책 신경망의 가중치 집합을 의미하며, T는 전체 에피소드 제한 길이보다 작다고 가정한다. 이후, 최적화 단계에서는 정책 갱신을 위해 샘플링 단계에서 수집한 경로 궤적 τ들의 상태-행위 쌍으로부터 얻은 보상을 기반으로 손실을 계산하여 정책을 학습하며, 이때의 손실은 Equation (4)와 같이 계산한다.

$$L_{PPO} = \frac{1}{\mid U_k T \mid} \sum_{\tau \in U_k} \sum_{t=0}^{T-1} \left[\min(R_t(\theta)\widehat{A}_t, C(R_t(\theta), \epsilon)\widehat{A}_t) \right] \tag{4}$$

이때, $C(R_t(\theta), \epsilon)$은 $clip(R_t(\theta), 1-\epsilon, 1+\epsilon)$을 의미하며, $\widehat{A}_t$는 t시간 단계에서의 어드밴티지 함수(advantage function)인 $A^{\pi\theta}(s_t, a_t) = Q^{\pi\theta}(s_t, a_t) - V^{\pi\theta}(s_t)$의 측정치이다. 이때, $Q^{\pi\theta}(s_t, a_t)$와 $V^{\pi\theta}(s_t)$는 각각 Equation (5)과 Equation (6)과 같이 계산한다.

$$Q^{\pi\theta}(s_t, a_t) = E_{a_{t'} \sim \pi_\theta} \left[\sum_{t'=t}^{T} \gamma^{t'} r_{t'} \mid S_t = s_t, A_t = a_t \right] \tag{5}$$

$$V^{\pi\theta}(s_t) = E_{a_{t'} \sim \pi_\theta} \left[\sum_{t'=t}^{T} \gamma^{t'} r_{t'} \mid S_t = s_t \right] \tag{6}$$

$R_t(\theta)$는 갱신된 정책과 이전 정책 간의 확률비를 의미하며, $\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta\,old}(a_t \mid s_t)}$와 같이 계산한다. 또한, Equation (6)과 Equation (7)에서의 γ는 감가 계수를, s_t와 a_t는 각각 경로 궤적 내 t시간 단계의 상태와 행동을 나타낸다.

한편, 목표 물체에 관한 3가지 보조 손실은 Fig. 2의 목표 예측 모듈과 같이 시공간적 맥락정보 특징 f_t^*을 입력으로 받아 관측 여부, 목표의 방향, 그리고 목표까지 거리를 예측한 값으로부터 구해내며, 이와 관련한 Equation (7)부터 Equation (12)는 모두 [2]의 논문을 참고하였다. 정답은 50x50 크기로 자른 에이전트 중앙의 지역적 정답 목표 지도를 사용하며, 각

각 Equation (7), Equation (8)와 같이 계산한다.

$$\phi^*_{t,c} = (o_t, e) = -\text{atan2}(o_{t,x} - e_x, o_{t,y} - e_y) \tag{7}$$

$$d^*_t = \|o_t - e\|_2 \tag{8}$$

$e = [e_x, e_y]$는 에이전트 위치에 해당하는 (x, y) 좌표값으로, 모든 시간 단계에서 지도의 중심인 (25, 25)로 고정된다. $o_t = [o_{t,x}, o_{t,y}]$는 t시간 단계에서의 목표 물체의 중심 (x, y) 좌표값을 의미한다. 방향은 $[0, 2\pi]$를 총 $K(=12)$ 개의 클래스로, 거리는 총 $L(=36)$ 개의 클래스로 이산화한다. 각 손실 함수는 Equation (9)부터 Equation (11)까지와 같다.

$$L_{obs} = \frac{1}{\mid U_k T \mid} \sum_{\tau \in U_k} \sum_{t=0}^{T-1} -(1^{obs}_t \log p(\widehat{obs}_t) + (1 - 1^{obs}_t) \log(1 - p(\widehat{obs}_t))) \tag{9}$$

$$L_{dist} = \frac{1}{\mid U_k T \mid} \sum_{\tau \in U_k} \sum_{t=0}^{T-1} \left[-1^{obs}_t \sum_{c=1}^{L} d^*_{t,c} \log p(\widehat{d_{t,c}}) \right] \tag{10}$$

$$L_{dir} = \frac{1}{\mid U_k T \mid} \sum_{\tau \in U_k} \sum_{t=0}^{T-1} \left[-1^{obs}_t \sum_{c=1}^{K} \phi^*_{t,c} \log p(\widehat{\phi_{t,c}}) \right] \tag{11}$$

이렇게 계산된 예측 손실값들은 Equation (12)과 같이 PPO 손실에 더하여 전체 학습의 손실로 사용되며, 평가 시에는 사용하지 않는다.

$$L_{total} = L_{PPO} + \lambda_{obs} L_{obs} + \lambda_{dist} L_{dist} + \lambda_{dir} L_{dir} \tag{12}$$

이때, λ_{obs}, λ_{dist}, λ_{dir}는 가장 우수한 성능을 보이는 0.25의 가중치로 모두 동일하게 고정하였다.

4. 구현 및 실험

4.1 모델 구현과 학습

본 논문의 제안 모델 MCFMO는 Python 딥러닝 라이브러리인 Pytorch를 이용해 구현하였으며, 성능 평가 실험들은 모두 Intel i9-12900K CPU와 GeForce RTX 3090 GPU 2개가 장착된 컴퓨터 환경에서 수행되었다. 그리고 제안 모델의 학습과 성능 평가에는 3차원 실사(photo-realistic)의 대규모 실내 시뮬레이션 환경을 제공하는 Habitat-Matterport3D 장면 데이터 집합을 이용하였다. 총 91개의 가상 실내 환경 중, 61개의 환경은 모델 학습에, 나머지 30개의 환경은 검증 및 평가를 위해 사용하였으며, 이러한 실내 환경을 이용하여 정의된 에피소드들은 학습용 50,000개, 검증용 12,500개, 평가용 12,500개가 존재한다. 각 에피소드는 실내 환경 고유 번호(id), 에이전트의 초기 위치, 그리고 3가지 목표 물체의 색상 및 위치 등을 사전에 정의한다. 제안 모델의 학습 과정은 16개의 병렬 에이전트를 통해 총 70M 시간 단계(step)만큼 진행하였으며, 평가 시에는 18개의 병렬 에이전트를 통해 평가용 에피소드 1,000개로 측정한 성능을 기재하였다. 작업 학습을 위해 설계된 보상함수 r은 아래의 Equation (13)과 같이 정의된다.

$$r_t = r_{sucessss} + r_{doser} + r_{time-penalty} \tag{13}$$

$r_{success}$는 현재 목표 물체에 대해 성공적으로 방문한 경우 받는 작업 성공 보상이며, r_{closer}는 현재 목표 물체와 에이전트 간의 거리 감소에 따른 보상으로, 이전 시점에서 목표까지 거리 d_{t-1}와 현재 시점에서의 거리 d_t간 차이로 정의한다. $r_{time-penalty}$는 최단 시간과 가깝도록 작업을 성공시키는 행동 정책을 학습하기 위해 매 시간 단계마다 에이전트가 받게 되는 음의 보상이다. 이때 $r_{success}$와 $r_{time-penalty}$의 크기는 비교 실험 후 가장 우수한 성능을 보이는 3.0과 -0.01로 각각 고정하였다.

4.2 성능 평가 실험

본 논문의 정량 평가 실험들에서 사용한 성능 평가 지표들(metrics)은 (1) 작업 성공률인 Success, (2) 방문에 성공한 목표 물체들의 비율인 Progress, 그리고, Success와 Progress를 기초로 이동 궤적의 효율성을 평가할 수 있는 (3) SPL(Success weighted by Path Length)와 (4) PPL(Progress weighted by Path Length) 등 총 4가지이다. 이 중에서 SPL은 Equation (14), PPL은 Equation (15)과 같이 계산한다.

$$SPL = \frac{1}{n} \sum_{i=1}^{n} S_i \frac{l_i}{\max(p_i, l_i)} \tag{14}$$

$$PPL = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} S_{i,j} \frac{l_{i,j}}{\max(p_{i,j}, l_{i,j})} \tag{15}$$

Equation (14)의 S_i는 i번째 에피소드의 작업 성공 여부(0 or 1)을 나타내며, l_i는 실제 i번째 에피소드에서 마지막 목표까지의 최단 경로 길이, p_i는 i번째 마지막 목표 물체까지 지나온 에이전트의 경로 길이, n은 동일하게 에피소드 개수를 의미한다. Equation (15)의 m은 목표 물체 개수, i,j는 i번째 에피소드에서의 j번째 물체를 의미하며 변수의 의미는 Equation (14)와 동일하다.

첫 번째 실험은 제안 모델 MCFMO에서 이용하는 멀티모달 공간적 맥락 특징의 긍정적 효과를 입증하기 위한 실험이다. 이를 위해 이 실험에서는 (a) RGB-D 입력 영상으로부터 추출한 시각적 외관 특징만 이용한 경우(f_{vis}), (b) 시각적 외관 특징과 환경 물체의 의미적 특징을 함께 이용한 경우($f_{vis+sem}$), (c) 시각적 외관 특징과 목표의 의미적 특징을 함께 이용한 경우($f_{vis+goal}$), 마지막으로 (d) 제안 모델과 같이 3 가지 서로 다른 공간적 맥락 특징들을 모두 이용한 경우($f_{vis+sem+goal}$)의 성능을 서로 비교하였다.

Table 1. Comparison with Different Spatial Context Features

Context Features	Metrics(%)			
	Success	Progress	SPL	PPL
(a) f_{vis}	47	62	30	40
(b) $f_{vis+sem}$	49	62	33	41
(c) $f_{vis+goal}$	51	64	32	41
(d) $f_{vis+sem+goal}$	60	73	36	43

Table 2. Comparison with Different Fusion Methods

Fusion Methods	Metrics(%)			
	Success	Progress	SPL	PPL
(a) concatenate	52	66	34	41
(b) linear combination	49	63	35	45
(c) concatenate + point-wise convolution	60	73	36	43

Table 1의 실험 결과를 살펴보면, 제안 모델과 같이 3가지 멀티모달 맥락 특징들을 모두 이용한 (d) $f_{vis+sem+goal}$의 경우가 모든 평가 지표에서 다른 맥락 특징들의 경우보다 더 높은 성능을 보여주었음을 알 수 있다. 특히 시각적 외관 특징만 이용한 (a) f_{vis} 경우에 비해서는 Success, Progress, SPL, PPL의 성능 평가 지표에서 각각 27.7%, 17.7%, 20%, 7.5%의 성능 향상률을 보인 것을 확인할 수 있다. 또한 단일 맥락 특징을 이용한 (a) f_{vis} 경우에 비해, 환경 물체의 의미적 특징을 추가한 (b) $f_{vis+sem}$와 목표 특징을 추가한 (c) $f_{vis+goal}$의 경우도 성능 향상 효과를 확인할 수 있다. 반면에, (b) $f_{vis+sem}$와 (c) $f_{vis+goal}$의 경우를 비교해보면, 환경 물체의 의미적 특징에 비해 목표 특징이 조금 더 성능 향상에 더 큰 영향을 미쳤다는 실험 결과를 확인할 수 있다. 이러한 실험 결과들을 종합해 볼 때, 본 논문의 제안 모델 MCFMO과 같이 서로 다른 멀티모달 맥락 특징들을 상호 보완적으로 이용하는 것이 에이전트의 공간적 맥락 이해와 작업 성능 향상에 긍정적 효과를 준다는 사실을 확인할 수 있었다.

두 번째 실험은 제안 모델에서 채용한 점-단위 합성곱 신경망 기반 멀티모달 특징 융합 방식의 긍정적 효과를 입증하기 위한 실험이다. 이 실험에서는 시각적 외관 특징 f_{vis}, 환경 물체의 의미적 특징 f_{sem}, 목표 특징 f_{goal} 등 3 가지 맥락 특징들을 융합하는 서로 다른 방식들인 (a) 단순 연결(concatenate), (b) 선형 결합(linear combination), (c) 제안 모델과 같이 1×1 크기의 커널을 갖는 점-단위 합성곱 (concatenate + point-wise convolution)들을 서로 비교하였다. 이때, (b) 선형 결합의 경우에는 3가지 특징 모두 동일하게 0.3의 고정된 가중치를 곱한 후, 이들을 모두 더하는 방식으로 특징들을 융합하였다.

Table 2의 실험 결과를 살펴보면, 본 논문에서 제안한 (c) 점-단위 합성곱 신경망을 이용한 융합 방식이 대부분의 평가 지표들에서 가장 좋은 성능을 나타내었다. 특히 (c) 점-단위 합성곱 기반 융합 방식은 (a) 단순 연결 방식에 비해,각각 평가 지표면에서 15.4%, 9.6%, 5.9%, 4.9%의 성능 향상을 보였다. 또한, (c) 점-단위 합성곱 기반 융합 방식을 (b) 선형 결합 방식과 비교했을 때도 PPL을 제외한, Success, Progress, SPL 등 나머지 3가지 평가 지표에서 각각 16.1%, 10.6%, 2.9%의 성능 향상을 보였다. 이와 같은 실험 결과들을 종합해보면, 본 논문에서 제안한 점-단위 합성곱 신경망 기반 맥락정보 융합 방식이 다른 융합 방식들에 비해 서로 다른 맥락 특징들 간의 상호 연관성을 효과적으로 탐지해내고 이를 토대로 더 높은 융합 시너지 효과를 얻을 수 있음을 확인할 수 있었다. 다만, 점-단위 합성곱을 통한 차원 축소 과정에서 크기가 작은 물체들에 대한 일부 정보 손실이 발생하는 경우 이동 경로 계획에 영향을 주어 (b) 선형 결합 방식이 PPL 평가 지표에서 조금 더 우수한 결과를 보이기도 하였다.

세 번째 실험은 제안 모델에서 채용한 목표 예측 보조 작업 학습(auxiliary task learning)의 성능 향상에 미치는 긍정적 효과를 입증하기 위한 실험이다. 이 실험에서는 모델학습을 위해, (a) 보조 작업 손실없이 PPO 손실만 사용한 경우(L_{PPO}), 보조 작업 학습을 위해서 (b) 목표 관측 손실만을 추가 사용한 경우($L_{PPO}+L_{obs}$), (c) 목표 관측 손실과 목표 거리 손실을 추가 사용한 경우($L_{PPO}+L_{obs}+L_{dist}$), (d) 목표 관측 손실과 목표 방향 손실을 추가 사용한 경우($L_{PPO}+L_{obs}+L_{dir}$), 마지막으로 (e) 제안 모델과 같이 3 가지 서로 다른 목표 예측 손실들을 모두 추가 사용한 경우($L_{PPO}+L_{obs}+L_{dir}+L_{dist}$)의 성능을 서로 비교해본다. 또 이 실험에서는 모든 보조 손실들의 가중치는 0.25로 동일하게 설정하였다.

Table 3의 실험 결과를 살펴보면, (e) 제안 모델과 같이 3가지 보조 작업 손실을 모두 사용한 경우($L_{PPO}+L_{obs}+L_{dir}+L_{dist}$)가 모든 평가 지표면에서 가장 우수한 성능을 보여주었다. 특히 보조 작업 손실을 전혀 포함하지 않은 (a) L_{PPO}의 경우에 비해서는 각각의 평가 지표에서 46.3%, 23.7%, 38.5%, 19.4%의 성능 향상률을 보였다.

한편, 보조 작업 손실들을 포함한 나머지 경우들을 살펴보면, 목표 관측 보조 손실만 추가한 (b) $L_{PPO}+L_{obs}$의 경우에도

Table 3. Comparison with Different Auxiliary Task Losses

Losses	Metrics(%)			
	Success	Progress	SPL	PPL
(a) L_{PPO}	41	59	26	36
(b) $L_{PPO}+L_{obs}$	44	60	27	38
(c) $L_{PPO}+L_{obs}+L_{dist}$	51	64	31	40
(d) $L_{PPO}+L_{obs}+L_{dir}$	54	67	34	41
(e) $L_{PPO}+L_{obs}+L_{dist}+L_{dir}$	60	73	36	43

(a) L_{PPO}에 비해 각 평가 지표에서 7.3%, 1.7%, 3.8%, 5.6%의 성능 향상을 보였으며, 특히 작업 성공률 Success 향상에 뚜렷한 도움이 되었음을 알 수 있다. 또한, 목표 거리 손실을 추가한 (c) $L_{PPO}+L_{obs}+L_{dist}$의 경우는 그렇지 않은 (b) $L_{PPO}+L_{obs}$에 비해 각 평가 지표에서 15.9%, 6.7%, 14.8%, 5.3%의 성능 향상을 보였다. 또한 목표 방향 손실을 추가한 (d) $L_{PPO}+L_{obs}+L_{dir}$의 경우는 그렇지 않은 (b) $L_{PPO}+L_{obs}$에 비해, 각 평가 지표에서 22.7%, 11.7%, 26%, 7.9%의 성능 향상을 보였다. 따라서 이러한 실험 결과들을 통해 목표 거리 손실 L_{dist}과 목표 방향 손실 L_{dir} 모두 모델의 성능 향상에 도움을 줄 수 있지만, 목표 방향 손실 L_{dir}과 연관된 보조 작업 학습이 목표 거리 손실 L_{dist}과 연관된 보조 작업 학습보다 상대적으로 성능 향상에 더 효과적임을 추가적으로 알 수 있었다.

네 번째 실험은 최근에 발표된 다른 MultiOn 모델들과의 비교를 통해 제안 모델의 우수성을 입증하기 위한 실험이다. 이 실험에서는 3 가지 서로 다른 맥락 특징들을 융합한 멀티모달 맥락 특징 지도를 이용하는 제안 모델 MCFMO를, 시각적 외관 단일 맥락 특징 지도만을 이용하는 ProjNeuralMap [1]와 AuxTaskMap[2], 목표 단일 맥락 특징 지도만을 이용하는 ObjRecogMap[1]과 Modular-MON[3], 점유 지도와 목표 지도를 함께 이용하는 SGoLAM[4] 등의 기존 모델들과 작업 성능을 서로 비교하였다.

Table 4의 실험 결과를 살펴보면, 본 논문의 제안 모델인 MCFMO가 비교 대상인 다른 모든 모델들에 비해 SPL과 PPL 평가 지표에서 가장 높은 성능을 보여주었다. 자세히 살펴보면, 제안 모델 MCFMO(ours)이 단일 맥락 지도를 사용하는 ObjRecogMap, ProjNeuralMap, Modular-MON, AuxTaskMap 모델들과 비교하였을 때, MultiOn 작업의 특성을 가장 잘 반영하는 평가 지표인 PPL에서 각각 126.3%, 104.8%, 65.4%, 7.5%의 성능 향상이 있었다. 또, 멀티모달 맥락 지도를 사용하는 SGoLAM에 비해서도 각 평가 지표에서 15.4%, 14.1%, 12.5%, 13.2%의 성능 향상을 보이는 것을 확인하였다. 또, PPL 평가 지표가 가장 높았던 기존 모델인 AuxTaskMap에 비해서도 모든 평가 지표에서 27.7%, 17.7%, 20%, 7.5%만큼의 성능 향상이 있었다. 이와 같은 실험 결과들을 바탕으로, 본 논문에서 제안한 MCFMO 모델의 효과와 우수성을 확인할 수 있었다.

마지막 실험은 몇 가지 대표 MultiOn 작업 사례들을 통해, 제안 모델 MCFMO의 장점과 한계점을 분석하기 위한 정성 평가 실험이다. Fig. 5부터 Fig. 7까지는 이 실험의 결과들을 나타내며, Fig. 5, 6는 MultiOn 작업의 성공 사례들을, Fig. 7은 작업 실패 사례를 보여준다. 특히, Fig. 5은 매우 신속하게 효율적으로 작업에 성공한 사례를, 반면에 Fig. 6은 최종적으로 작업에 성공하기는 했으나 배회하는 과정이 길어 작업 효율성이 낮은 사례를 나타낸다. 각 작업의 진행과정을 위에서 아래로, 왼쪽에서 오른쪽으로 시간 순서대로 열거하였다. 다음의 각 그림에서 하향식 지도(top-down view map)의 회색 부분은

Table 4. Comparison with Other Multi-Object Goal Visual Navigation(MultiOn) Models

Model	Metrics(%)			
	Success	Progress	SPL	PPL
ObjRecogMap* [1]	17	35	9	19
ProjNeuralMap* [1]	17	38	8	21
Modular-MON* [3]	50	65	21	26
SGoLAM* [4]	52	64	32	38
AuxTaskMap* [2]	47	62	30	40
MCFMO(ours)	60	73	36	43

이동 가능한 자유 공간(free space)을, 흰색 부분은 벽이나 가구 등 장애물들이 차지하고 있어 에이전트가 갈 수 없는 공간을 나타낸다. 또한, 입력(input) 부분은 에이전트의 이동에 따른 실시간 입력 RGB 영상과 깊이 영상(depth image)를 나타내며, 우측 visual은 입력 RGB-D 영상에서 추출한 시각적 외관 특징 f_{vis}을, semantic은 환경 물체의 의미적 특징 f_{sem}을, goal은 현재 관측되는 목표 특징 f_{goal}을 각각 나타낸다.

먼저 Fig. 5의 작업 사례의 경우, step 1에서 에이전트는 파란색 실린더(▮)로 표시된 현재의 목표 대신 세 번째 목표인 노란색 실린더(▮)를 먼저 관측하였다. 이때 제안 모델은 미리 관측된 세 번째 목표의 위치정보를 나중에 이용하기 위해 전역적 지도에 저장 보관한다. step 78부터 step 99까지는 첫 번째 목표(▮)를 탐지해내고 빠르게 접근하여 방문하는 과정을 보여주고, step 284에서는 검은색 실린더(▮)로 표시된 미관측 두 번째 목표를 탐색할 때, 미리 목표 관측 손실 L_{obs}을 바탕으로 학습된 제안 모델은 두 번째 목표까지 효율적인 탐색을 진행하였다. 이 과정에도 환경 물체에 대한 의미적 특징과 시각적 외관 특징을 활용함으로써, 장애물들과의 큰 충돌 없이 목표까지 이동한 것을 확인할 수 있다. 한편 step 327에서는 성공적으로 두 번째 목표까지 방문한 후, 세 번째 목표인 노란색 실린더(▮)를 찾아 방문해야 하는 상황이다. 이때 이미 이전에 관측된 세 번째 목표의 위치정보를 포함한 전역적 지도를 활용함으로써, 곧바로 세 번째 목표까지 최단 경로에 가까운 이동 경로로 접근하여 step 453에서 최종적으로 작업에 성공하는 과정을 보여준다. 이 같은 사례를 통해 제안 모델은 3가지 멀티모달 공간적 맥락 특징들과 보조 작업 손실들을 효과적으로 이용함으로써 복잡한 환경 맥락정보들을 빠르게 파악하여 효율적으로 작업에 성공할 수 있음을 다시 확인할 수 있었다.

한편, Fig. 6의 경우는 에이전트가 특정 방 내부의 시작 위치에서 출발해서 해당 방 밖에 위치한 첫 번째 목표인 노란색 실린더(▮)부터 찾아서 방문해야 하는 작업 사례이다. 제안 모델의 에이전트는 이 작업을 수행하는 동안 출입문을 찾아 방을 빠져 나가기 위해 좌충우돌하느라 상당한 지연 시간을 소모하였다. 진행된 작업과정을 자세히 살펴보면, 제안 모델의 에이전트는 초기 위치에서 시작하여 step 1부터 step 56까지

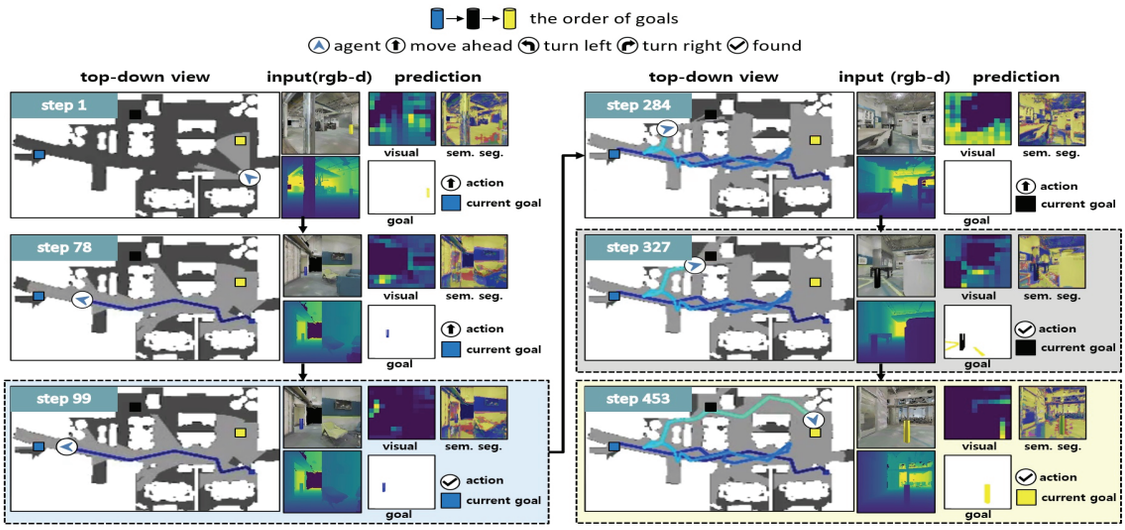

Fig. 5. Qualitative Evaluation of the Proposed Model: Case 1

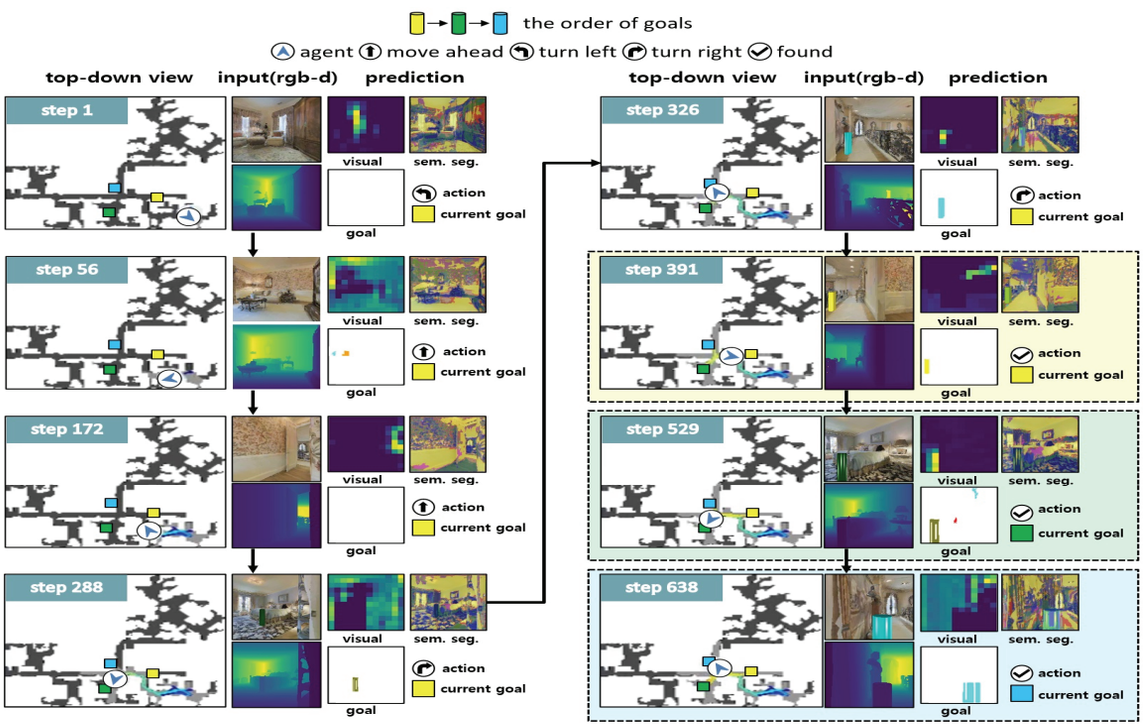

Fig. 6. Qualitative Evaluation of the Proposed Model: Case 2

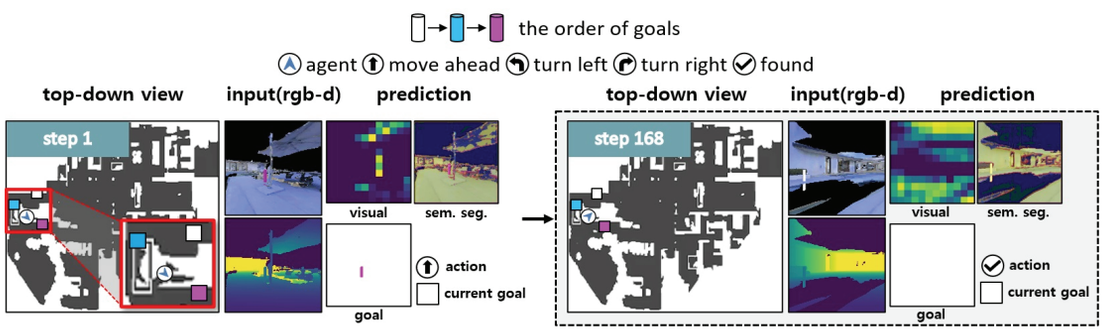

Fig. 7. Qualitative Evaluation of the Proposed Model: Case 3

현재 방 안에 첫 번째 목표 실린더()가 없음을 확인 후, 마침내 출입문을 찾아 방을 탈출하는 step 172까지 긴 시간동안 방 안을 배회하는 모습을 보여주었다. 이후 step 391에서 첫 번째 목표 실린더()를 성공적으로 방문하게 될 때까지 도중에 step 288과 step 326 각각에서 두 번째(), 세 번째() 목표 실린더를 미리 관측하여 이들의 위치정보를 전역적 지도에 저장해둠으로써, 이 정보를 이용해 추가 탐색없이 곧바로 step 529과 step 638에서 이 후속 목표들을 신속하게 방문할 수 있었다. 이 사례에서 보듯이, 현재의 제안 모델 MCFMO는 복잡한 폐쇄형 공간에서 좁은 출입구를 찾아 탈출하는데는 적지 않은 어려움을 노출시켰다. 이러한 제안 모델의 문제점을 극복하기 위해서는 열린 출입문을 좀 더 효과적으로 탐지해내도록 인식 기능을 보강하거나, 목표 탐지를 촉진하기 위해 더 넓은 공간으로 이동을 유도하는 정책 학습에 관한 연구가 향후에 추가적으로 필요할 것으로 판단된다.

한편, Fig. 7은 시각 인식 범위의 한계로 인해, 제안 모델의 에이전트가 초기 위치 주변에 밀집된 에이전트의 카메라보다 낮은 높이를 가지는 근접 장애물들로부터 벗어나지 못해서 결국 목표까지 이동 작업에 실패한 사례를 나타낸다. 본 연구에서 이용하는 MultiOn 작업 시뮬레이션 플랫폼에서는 에이전트의 내장 카메라가 바닥에서 0.88m의 높이에 위치하며, 시야각(Field Of View, FOV)은 상하 79도로 제한되어 있다. 따라서 이 범위를 벗어난 낮은 근접 장애물들이 있는 경우에는 입력 영상만을 통해서는 에이전트가 이들을 식별해낼 수 없다. Fig. 7의 step 1의 작업 초기 상황이 바로 이 경우에 해당하여, 실제로는 주변에 존재하는 근접 장애물들을 에이전트 자신은 감지해내지 못함으로써 첫 번째 목표인 흰색 실린더()를 찾기 위해 통과할 수 없는 장애물 쪽으로 끊임없이 탈출을 시도하다가 결국 작업 실패로 종결되었다. 이러한 현상은 제안 모델의 맥락 이해와 행동 추론 능력의 문제라기보다는 특수한 장애물들이 포함된 복잡한 환경 지형과 에이전트의 인식 범위를 제한하는 MultiOn 작업 시뮬레이션 플랫폼에서 그 원인을 찾을 수 있을 것 같다. 따라서 이와 같은 문제를 극복하기 위해서는 현재 MultiOn 작업 시뮬레이션 플랫폼이 허용하는 에이전트의 행동들 중에 이동 행동들 외에 고개를 들거나 숙이는 신규 행동도 추가함으로써, 에이전트가 인식할 수 있는 상하 시각 범위를 확대해주는 방안도 고려해볼 수 있을 것 같다.

5. 결 론

본 논문에서는 다중 물체 목표 시각적 탐색 이동(MultiOn) 작업을 위한 새로운 에이전트 모델 MCFMO(Multimodal Context Fusion for MultiOn tasks)를 제안하였다. 제안 모델은 에이전트가 심도있는 공간적 맥락 이해가 가능하도록 입력 RGB-D 영상에서 추출하는 시각적 외관 특징외에 환경 물체의 의미적 특징, 목표 물체 특징을 함께 포함하는 멀티모달 맥락 지도를 행동 선택에 이용한다. 또, 멀티모달 맥락정보 간의 관계를 충분히 활용하기 위해 점-단위 합성곱 신경망을 이용하여 3가지 서로 이질적인 특징들을 하나의 전역 지도로 효과적으로 융합한다. 이 밖에도 효율적인 이동 정책 학습을 유도하기 위해 목표의 방향과 거리를 예측하는 보조 작업 학습 모듈을 추가로 채용한다. 본 논문에서는 벤치마크 실내 환경 데이터 집합인 Matterport3D와 3차원 시뮬레이터인 Habitat를 이용하여 진행한 다양한 실험들을 통해, 제안 모델의 우수성을 입증하였다.

앞서 정성적 평가 실험에서 언급한 것과 같이 현재의 제안 모델은 출입구를 관측하기 어려운 폐쇄 공간을 빠져나오려고 하거나, 시야 범위 내에 들어오지 못하는 근접 장애물들이 존재할 때 작업이 지연되거나 작업에 실패하는 경우도 발생한다. 이러한 제안 모델의 한계점을 극복하고 개선하기 위해서 폐쇄 공간의 출입구를 좀 더 효과적으로 탐지해내도록 인식 기능을 보강하거나, 목표 탐지를 위해 더 넓은 공간으로 이동을 유도하는 정책 학습법을 개발하거나, 에이전트의 가시 범위를 확장해주기 위해 MultiOn 작업 시뮬레이션 플랫폼을 수정해주는 등의 추가 연구를 향후에 진행할 예정이다. 이와 함께, 경험하지 못한(unseen) 환경에도 효과적으로 MultiOn 작업을 수행할 수 있도록, CLIP과 같이 사전 학습된 거대 비전-언어(vision-language) 모델을 활용하거나 개방형-어휘 의미적 분할(open-vocabulary semantic segmentation) 모델을 활용하는 등 제안 모델을 확장하는 향후 연구도 계획하고 있다.

References

[1] S. Wani, S. Patel, U. Jain, A. Chang, and M. Savva, “Multion: Benchmarking semantic map memory using multi-object navigation,” *Advances in Neural Information Processing Systems(NeurIPS)*, Vol.33, pp.9700-9712, 2020.

[2] P. Marza, L. Matignon, O. Simonin, and C. Wolf, “Teaching agents how to map: Spatial reasoning for multi-object navigation,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, pp.1725-1732, 2022.

[3] S. Raychaudhuri, T. Campari, U. Jain, M. Savva, and A. X. Chang, “Reduce, reuse, recycle: Modular multi-object navigation,” *arXiv preprint arXiv:2304.03696*, 2023.

[4] J. Kim, E. S. Lee, M. Lee, D. Zhang, and Y. M. Kim, “Sgolam: Simultaneous goal localization and mapping for multi-object goal navigation,” *arXiv preprint arXiv:2110.07171*, 2021.

[5] P. Chen, D. Ji, K. Lin, W. Hu, W. Huang, T. Li, M. Tan and C. Gan, “Learning active camera for multi-object navigation,” *Advances in Neural Information Processing Systems(NeurIPS)*, Vol.35, pp.28670-28682, 2022.

[6] N. Savinov, A. Dosovitskiy, and V. Koltun, "Semi-parametric topological memory for navigation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, 2018.

[7] K. Chen, J. K. Chen, J. Chuang, M. Vázquez, and S. Savarese, "Topological planning with transformers for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, Nashville, pp.11276-11286, 2021.

[8] D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural topological slam for visual navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, Seattle, pp.12875-12884, 2020.

[9] N. Kim, O. Kwon, H. Yoo, Y. Choi, J. Park, and S. Oh, "Topological semantic graph memory for image-goal navigation," in *Proceedings of the 6th Conference on Robot Learning (PMLR)*, Auckland, pp.393-402, 2023.

[10] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, Venice, 2017, pp.2616-2625.

[11] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems(NeurIPS)*, Vol.33, pp.4247-4258, 2020.

[12] P. Chen, D. Ji, K. Lin, R. Zeng, T. Li, M. Tan, and C. Gan, "Weakly-supervised multi-granularity map learning for vision-and-language navigation," *Advances in Neural Information Processing Systems(NeurIPS)*, Vol.35, pp.38149-38161, 2022.

[13] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, New Orleans, pp.18890-18900, 2022.

[14] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese, "Scene memory transformer for embodied agents in long-horizon tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. Long Beach, pp.538-547, 2019.

[15] B. Mayo, T. Hazan, and A. Tal, "Visual navigation with spatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, pp.16898-16907, 2021.

[16] A. Mousavian, A. Toshev, M. Fišer, J. Košecká, A. Wahid, and J. Davidson, "Visual representations for semantic target driven navigation," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, Montreal, pp.8846-8852, 2019.

최 정 현

https://orcid.org/0000-0002-9202-8785
e-mail : wjdgus21@kyonggi.ac.kr
2022년 경기대 컴퓨터공학부 졸업.
2022년 ~ 현 재 경기대학교 컴퓨과학과 석사과정
관심분야 : 인공지능, 로봇지능, 기계 학습

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, 로봇지능