Journal of KIIT. Vol. 24, No. 7, pp. 139-148, Jul. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.7.139

실내 자율 이동 로봇의 안전 주행을 위한 강건한 제로샷 기반 주행 가능 영역 분할 알고리즘

최윤정*, 이세훈**[1], 김태훈**[2], 안준호**[3]

A Robust Zero-Shot Traversable Area Segmentation Algorithm for Safe Navigation of Indoor Mobile Robots

Yunjeong Choi*, Sehun Lee**[1], Taehoon Kim**[2], and Junho Ahn**[3]

This work was supported by Kyonggi University's Graduate Research Assistantship 2026, by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2024-00436954), by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation)(2021-0-01393), by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2026-25489409)

요 약

실내 자율 이동 로봇의 안전한 주행에는 정확한 바닥(주행 가능 영역) 탐지가 필수적이다. 그러나 LiDAR 기반 방식은 비용이 높아 보급형 플랫폼에 적용하기 어렵고, 학습 기반 분할 모델은 새로운 실내 환경에서 일반화 성능이 크게 저하된다. 파운데이션 모델을 활용한 이미지 분할은 경계 정밀도가 높지만, 의미론적 판단이 없고, 자기지도학습 기반 특징 추출 모델은 의미 이해는 가능하지만, 경계 정밀도가 낮다. 본 논문은 두 모델의 약점이 서로의 강점으로 보완된다는 점에 착안하여, 자기지도학습 기반 특징 유사도로 바닥 후보를 추정하고 탐욕적 선택 알고리즘으로 최종 마스크를 구성하는 제로샷 기반 방법을 제안한다. 실내 데이터를 대상으로 5개 비교 모델과의 정량 평가를 수행한 결과, 제안 방법은 Recall 0.939, F1 0.865로 비교 모델 중 최고 성능을 기록하였다.

Abstract

Accurate floor detection is essential for safe navigation of indoor autonomous mobile robots. LiDAR-based systems are too costly for low-cost platforms, and learning-based models fail in unseen environments. We tested foundation segmentation and self-supervised feature extraction models independently: the former flooded predictions with non-floor regions, while the latter missed precise boundaries. The weakness of each was exactly what the other could provide. We built a zero-shot method around this: self-supervised feature similarity identifies floor candidates, and a greedy selection algorithm picks the final mask from foundation model proposals. Against five baselines on diverse indoor images, our method achieves Recall 0.939 and F1 0.865, outperforming all compared models.

Keywords

zero-shot floor detection, indoor navigation, segmentation, self-supervised feature extraction, foundation model

* 경기대학교 인공지능전공
- ORCID: https://orcid.org/0009-0000-6396-0189
** 경기대학교 컴퓨터과학과(**[3] 교신저자)
- ORCID[1]: https://orcid.org/0009-0006-0293-7544
- ORCID[2]: https://orcid.org/0009-0004-0494-3177
- ORCID[3]: https://orcid.org/0000-0003-2916-6410

· Received: May 18, 2026, Revised: Jun. 26, 2026, Accepted: Jun. 29, 2026
· Corresponding Author: Junho Ahn
Dept. of AI Computer Science and Engineering, Kyonggi University, 154-42 Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea
Tel.: +82-31-249-9731, Email: jha@kyonggi.ac.kr

Ⅰ. 서 론

실내 공간은 오피스, 상업 시설, 병원, 교육 기관 등 다양한 형태로 분화되고 있으며, 그 인테리어 구성 방식 또한 기업이나 기관의 성격에 따라 복잡도가 높아지고 있다. 이러한 환경에서 자율 이동 로봇(AMR, Autonomous Mobile Robot)은 물류 배송, 서빙, 안내 등 다양한 서비스 역할을 수행하며 빠르게 확산되고 있다. 국제로봇연맹(IFR)의 보고에 따르면, 2023년 전 세계 전문 서비스 로봇 판매량은 전년 대비 30% 증가한 205,000대를 기록하였으며, 그중 절반 이상인 113,000대가 운송·물류 분야에 배치된 이동 로봇이다[1]. IFR은 실내 환경에서의 이동 자동화를 해당 세그먼트의 가장 중요한 적용 분야로 명시하고 있어[2], 실내 AMR의 확산은 당분간 지속될 것으로 전망된다. 자율 이동 로봇이 실내 공간을 안전하게 주행하기 위해서는 바닥과 장애물 영역을 정확히 구별하는 능력이 필수적이다[3].

실내 주행 로봇은 LiDAR 센서를 이용하여 3차원 정보를 획득한다. 하지만 실내 환경에서는 반사성 표면이 많이 존재하며, 유리 너머의 장애물을 잘못 감지하는 한계를 가진다[4]. Patruno et al.도 LiDAR가 200mm 이하 장애물을 탐지하지 못하거나 반사성 표면에서 잘못 측정하는 물리적 한계를 가진다는 점을 지적했다[5]. 또한 LiDAR는 보급형 플랫폼에 적용하기에는 비용이 많이 든다는 현실적인 한계가 있다. 이러한 한계를 극복하기 위해 상대적으로 저비용인 단안 RGB 카메라를 활용한 딥러닝 기반 방법이 대안으로 제안되었다[6]. 그중 학습 기반 분할 모델의 경우, 학습되지 않은 새로운 환경에서 성능이 크게 저하되어 일반화에 한계를 갖는다[7]. 이는 바닥의 재질, 색상, 조명 조건이 공간에 따라 상이함에도 모델이 이를 충분히 학습하지 못하기 때문이다. 이러한 모델을 사용하는 연구에서도 특정 실내 환경 데이터로만 훈련된 분할 모델은 일반화를 향후 과제로 남기는 사례도 있다[8].

이에 본 연구에서는 어떠한 학습 데이터나 파인튜닝 없이 단일 RGB 이미지만으로 실내 바닥 영역을 분할하는 제로샷 기반 방법을 제안한다. 이를 위해 파운데이션 모델을 활용한 이미지 분할과 자기지도학습 기반 특징 추출 모델을 단독으로 적용하는 예비 실험을 수행하였으며, 실험을 통해 파운데이션 모델을 활용한 이미지 분할은 의미론적 판단 부재로 인한 과분할 문제가, 자기지도학습 기반 특징 추출 모델은 바닥 경계 정밀도 문제가 발생하는 것을 확인하였다. 두 모델의 단점을 보완하기 위해, 두 모델을 결합하는 방식으로 진행하였다.

본 논문의 주요 기여는 다음과 같다. 첫째, LiDAR나 별도의 학습 데이터 없이 보급형 로봇 플랫폼에서도 환경별 재학습 없이 적용 가능한 실내 바닥 탐지용 제로샷 알고리즘을 제안한다. 둘째, 자기지도학습 특징 추출 모델의 특징 유사도 기반 SAP(Single Anchor Prior)와 파운데이션 모델(Foundation model)의 전체 분할 결과에 CGS(Coverage-Greedy Selection)을 결합한 알고리즘을 설계하였다. SAP는 레이블 없이 바닥 후보 영역을 추정하고, CGS는 바닥 Prior와 정합성이 높은 마스크를 체계적으로 선택하고 합산하며, 최종 마스크에 단안 깊이 추정 모델을 오버레이하여 주행 가능 경로의 원근 구조를 시각화한다. 셋째, 단안 RGB 카메라가 탑재된 이동 로봇을 이용해 수집한 실내 데이터와 다양한 환경 이미지를 대상으로 5개 비교 모델과 정량 평가를 수행한 결과, 제안 방법은 Recall 0.939, F1 0.865를 달성하여 비교 대상 모델 중 최고 성능을 기록하였다.

이후 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 검토하고, 3장에서는 제안 방법을 기술한다. 4장은 실험 설계와 결과이며, 5장에서 결론과 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

2.1 이미지 분할

이미지 분할(Image segmentation)은 이미지에서 각 픽셀에 의미론적 레이블을 부여하거나 객체의 경계를 구분하는 컴퓨터 비전의 핵심 과제로, 장면 이해, 의료영상 분석, 로봇 지각 등 다양한 분야에서 활용된다[9]. 분할 방식은 크게 시맨틱 분할, 인스턴스 분할, 파놉틱(Panoptic) 분할로 분류된다. 시맨틱

분할은 픽셀 단위로 클래스를 부여하지만, 같은 클래스에 속하는 여러 객체를 개별적으로 구분하지는 못한다. 인스턴스 분할은 개별 객체를 따로 분리하지만, 배경 영역은 처리하지 않는다[10]. 파놉틱 분할은 두 방식의 한계를 극복한 방법으로, 전경과 배경 모두에 클래스 레이블과 인스턴스 식별자를 동시에 부여한다[11]. 하지만 세 방식 모두 대규모 레이블 데이터에 대한 의존도가 높고, 학습에 포함되지 않은 환경에서는 일반화 성능이 크게 떨어진다. 특정 실내 공간에 한정된 로봇은 이러한 한계를 감수할 수 있으나, 다양한 환경에 적용될 경우 심각한 성능 저하가 발생할 수 있다.

학습 기반 모델의 일반화 한계에 대한 대안으로 본 연구에서 주목한 것은 파운데이션 모델을 활용한 접근이다. 파운데이션 모델은 대규모 데이터로 사전 학습된 범용 모델을 그대로 활용하기 때문에, 특정 환경 데이터나 추가 파인튜닝 없이도 새로운 환경에서 분할이 가능하다는 점에서 제로샷 적용에 적합하다. 이 분야의 대표적 연구로 Kirillov et al.이 제안한 SAM(Segment Anything Model)이 있다[12]. SAM은 포인트, 바운딩 박스, 마스크 등 다양한 프롬프트를 입력으로 받아 해당 영역의 분할 마스크를 출력하며, 1,100만 장의 이미지와 10억 개 이상의 마스크로 구성된 SA-1B로 학습되어 다양한 객체를 정밀하게 분할한다. 이에 정확도와 속도가 향상되고 이미지와 비디오를 통합 처리할 수 있는 SAM2를 채택하였다[13].

특히 SAM 계열의 AMG(Automatic Mask Generation) 기능에 주목하였다. AMG는 내부적으로 균등 격자 포인트를 생성해 후보 마스크를 구성하고 신뢰도 기반 필터링을 거쳐 프롬프트 없이 이미지 전체를 자동으로 분할한다. 그러나 실내 데이터에 적용하였을 때 바닥, 벽, 가구 등 모든 영역이 구분 없이 과분할되는 문제를 직접 확인하였다. AMG는 어디를 분할할 것인지 판단하지만 무엇이 바닥인지는 알 수 없으므로, 바닥 영역만을 선택적으로 추출하기 위한 의미론적 판단 기준이 추가로 필요하다는 한계가 있다.

2.2 자기지도학습 기반 특징 추출 모델

자기지도학습(Self-supervised learning)은 레이블이 없는 대규모 데이터로부터 유용한 시각 표현을 스스로 학습하는 방법론이다. Jing 과 Tian[14]에 따르면, 자기지도학습은 비지도학습의 일종으로서 레이블 없이 이미지나 비디오로부터 범용적인 시각 특징을 학습하며, 다운스트림 태스크에서 지도학습에 필적하는 성능을 달성한다. 자기지도학습 기반 특징 추출 모델은 이렇게 학습된 인코더를 이용해 입력 이미지에서 의미론적 특징 벡터를 추출하는 구조를 말하며, 추출된 특징은 분류, 검출, 분할 등 다양한 태스크에 파인튜닝 없이 또는 최소한의 적응만으로 활용될 수 있다. 대표적인 방법론으로는 데이터 증강을 통한 대조 학습 방식이 있으며, 동일 이미지에서 생성된 서로 다른 뷰(Positive pair)를 가깝게, 다른 이미지의 뷰(Negative pair)를 멀게 학습하는 방식으로 표현을 정제한다[15].

ViT(Vision Transformer)는 이미지를 고정 크기 패치로 나누어 트랜스포머로 처리하는 구조다[16]. ViT는 지도학습 기반 이미지 분류용으로 제안되었지만, 이후 자기지도학습 분야에서도 핵심 아키텍처로 활용되었다. Caron et al.은 ViT를 활용해 레이블 없이도 의미론적 분할 특성이 자연스럽게 나타남을 보인 DINO를 제안하였고[17], Oquab et al.[18]은 이를 더 발전시킨 DINOv2를 제안하였다. DINOv2는 정제된 대규모 데이터셋(LVD-142M)에서 교사-학생 자기지도 방식으로 학습되어 파인튜닝 없이도 분류·검출·분할 등 다양한 비전 태스크에서 활용할 수 있는 범용 특징을 추출해낸다. DINOv2의 패치 토큰은 입력 이미지를 14×14픽셀 단위로 분할하여 각 패치의 시각적 특징을 인코딩한다. 이때 의미론적으로 유사한 영역 간에 높은 특징 유사도가 나타나며, 이 점이 본 연구에서 해당 모델을 선택한 핵심 근거이다. 하지만 DINOv2에서는 특징 맵에 불필요한 아티팩트가 나타나는 문제가 있었다. Darcet et al.[19]은 그 원인을 분석하고, 레지스터 토큰을 추가하여 DINOv2의 아티팩트 문제를 해결하는 방법을 제안하였다. 이후 Siméoni et al.[20]은 규모 확장과 Gram anchoring 기법을 도입한 DINOv3를 발표하였으며, 이를 통해 밀집 특징 맵 품질과 밀집 예측 성능을 향상시켰다.

그럼에도 불구하고 DINOv2와 DINOv3 모두 패치 단위 특징 추출이라는 구조적 특성으로 인해 특징 맵의 공간 해상도와 원본 이미지 간의 해상도 차이가 있다. 이로 인해 바닥과 벽이 인접하거나 바닥과 가구의 경계가 가까운 영역에서, 특징 유사도만으로는 픽셀 수준의 정밀한 경계를 추정하지 못한다. 이를 극복하기 위해, DINO 방식과 높은 경계 정밀도를 가진 파운데이션 모델을 활용한 이미지 분할을 결합하고자 한다.

III. 제안 방안

3.1 제안된 알고리즘

본 연구에서 제안하는 방법은 SAM2와 DINOv3 두 계열 모델의 한계를 상호 보완적으로 결합한다. 의미론적 판단은 DINOv3, 경계 결정은 SAM2로 역할을 분리한다. DINOv3는 패치 단위 특징 유사도를 통해 바닥 앵커 기반의 Seed Mask를 생성하고, SAM2는 경계가 픽셀 수준으로 정밀한 후보 마스크를 생성하며, CGS가 어느 후보를 선택할지 결정한다. 이렇게 함으로써 DINOv3가 가진 의미론적 이해 능력과 SAM2가 가진 픽셀 수준의 경계 정밀도를 동시에 활용한다.

그림 1은 제안하는 제로샷 기반 바닥 탐지를 위한 결합 모델의 전체 구조를 나타낸다. SAP와 CGS는 사전 학습된 두 파운데이션 모델인 특징 추출용 DINOv3 백본과 분할용 SAM2의 출력인 패치 특징과 분할 후보를 입력으로 받는 후처리(Post-processing) 알고리즘으로, 학습 가능한 파라미터나 가중치 갱신 없이 동결된 두 모델의 추론 결과만을 결합한다. 제안 방법에서는 두 브랜치가 병렬로 처리된 후 CGS에서 통합되어 최종 마스크가 출력된다. DINOv3 브랜치에서는 DINOv3 ViT-L/16 모델로 입력 이미지의 패치 특징 벡터를 추출한다. 이후 SAP를 적용하여 이미지 하단 앵커 기반의 바닥 유사도 맵(Heat Mask)과 이를 이진화한 Seed Mask를 생성한다. SAM2 브랜치에서는 SAM2 AMG를 통해 입력 이미지 전체에 대한 분할 후보 집합 $R = \{M_1, M_2, \ldots, M_n\}$를 생성한다. 각 후보는 SAM2 내부의 IoU 예측 점수와 안정성 점수를 함께 가지며, 경계가 픽셀 수준으로 정밀하다. DINOv3 브랜치에서 생성된 Seed Mask를 기준으로, 분할 후보 집합 R에서 Seed Mask를 가장 효율적으로 커버하는 후보들을 탐욕적으로 선택한다. 마지막으로, 선택된 후보들의 합집합을 최종 바닥 마스크 $\hat{M}$으로 결정하고, Depth Anything V2로 깊이 정보를 오버레이하여 주행 경로를 시각화한다[21]. 따라서 제안 방법은 별도의 학습 없이 추론 시에만 동작하는 완전한 제로샷 방식이다.

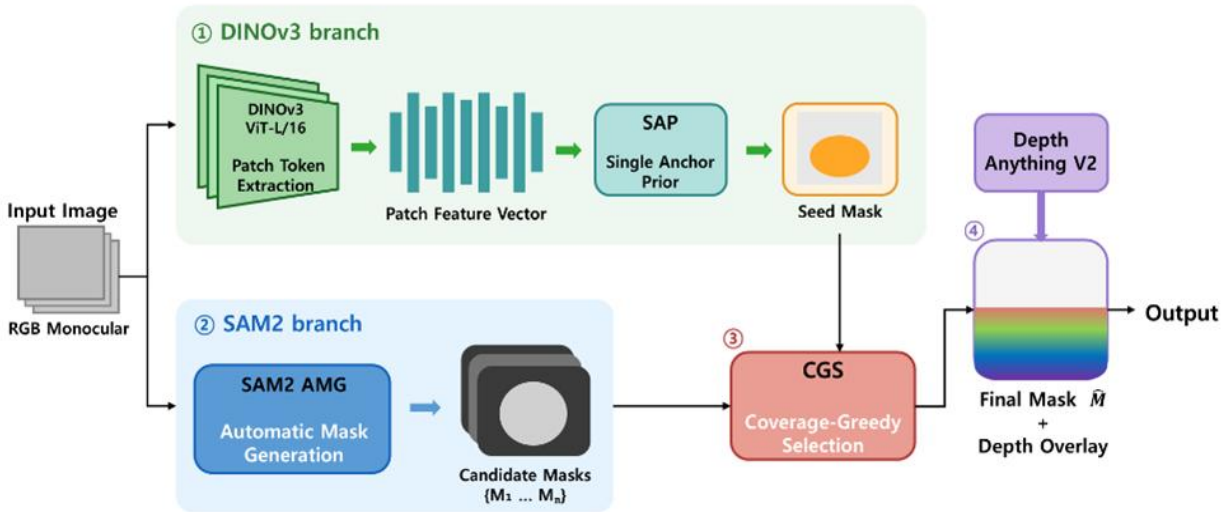

그림 1. 제안하는 바닥 탐지 방법의 전체 구조
Fig. 1. Overall architecture of the proposed floor detection method

3.2 SAP(Single Anchor Prior)

로봇이 주행하는 시점에서 촬영된 이미지는 카메라가 전방을 향하기 때문에, 이미지 하단부에 바닥이 나타날 가능성이 높다는 기하학적 가정을 활용한다. 입력 이미지를 사전 학습된 백본 모델에 통과시켜 각 패치의 특징 벡터 t_i를 추출한 뒤, 이미지 하단 25% 영역에 해당하는 패치 인덱스 집합 A의 특징 벡터를 평균하여 바닥 영역의 대표 특징 벡터인 앵커 프로토타입 $\hat{f}_{anchor}$를 식 (1)과 같이 구성한다.

$$f_{anchor} = \frac{1}{|A|}\sum_{i\in A} t_i,\ \hat{f}_{anchor} = \frac{f_{anchor}}{\| f_{anchor} \|} \quad (1)$$

이후 이미지 전체의 각 패치 특징 벡터 $\hat{t_i}$와 앵커 프로토타입 사이의 유사도 h_i를 식 (2)와 같이 계산하여 바닥 유사도 맵을 생성한다. 이때 두 특징 벡터 간의 유사도 척도로 코사인 유사도를 적용한다. 코사인 유사도를 사용하는 이유는 조명 변화나 색상 차이와 무관하게 특징 벡터의 방향 유사성을 측정할 수 있기 때문이다.

$$h_i = \cos(\hat{t_i}, \hat{f}_{anchor}) = \hat{t_i}^{\top}\hat{f}_{anchor} \quad (2)$$

이 유사도 맵을 원본 이미지 해상도로 보간한 결과를 Heat Mask라 한다. Heat Mask에서 값이 높은 픽셀일수록 앵커 영역과 특징이 유사한 바닥 후보에 해당한다. Heat Mask를 이진화하여 바닥임이 확실한 고신뢰 영역을 Seed Mask로 추출하며, Seed Mask는 이후 CGS 단계에서 SAM2 후보 선택의 기준으로 사용된다.

3.3 SAM2 CGS(Coverage-Greedy Selection)

SAM2 AMG를 통해 입력 이미지 전체에 대한 분할 후보 집합 $R=\{M_1, M_2, \dots, M_n\}$을 생성한다. 각 후보는 SAM2 내부의 IoU 예측 점수 σ_i와 안정성 점수를 함께 가지며, 경계가 픽셀 수준으로 정밀하다. 그러나 어느 후보가 실제 바닥에 해당하는지는 SAM2 스스로 판단할 수 없으므로, 이전 단계에서 생성한 M_{seed}를 기준으로 후보를 선별해야 한다. 단순히 SAM2 내부 점수가 높은 상위 k개를 합산하는 방식(top-k union)에서는 바닥과 무관한 고점수 후보가 포함될 수 있다. 이를 해소하기 위해 현재 누적 합집합이 Seed Mask를 얼마나 덮고 있는지(Coverage)를 기준으로 후보를 탐욕적으로 반복 선택하는 CGS를 채택하였다.

각 후보 M_i에 대해 히트맵 평균 $\overline{h_i}$, 상위 분위수 $h_i^{(q)}$, 시드 커버리지 $cov(M_i, M_{seed})$, 하단 연결성 $conn_{\perp}(M_i)$, 상단 점유 패널티 $occ_{\top}(M_i)$, 면적 Prior $p_{area}(M_i)$, SAM2 신뢰도 σ_i를 종합한 복합 점수 s_i를 식 (3)과 같이 산출한다.

$$\begin{aligned} s_i = {} & w_1\overline{h}_i + w_2 h_i^{(q)} + w_3\, cov(M_i, M_{seed}) \\ & + w_4 conn_{\perp}(M_i) + w_5\, occ_{\top}(M_i) \\ & + w_6\, p_{area}(M_i) + w_7\sigma_i \end{aligned} \quad (3)$$

CGS는 식 (4)와 같이 현재 누적 합집합이 M_{seed}를 덮는 커버리지를 최대화하는 후보를 탐욕적으로 반복 선택한다. 커버리지가 사전 정의된 임계값 Γ에 도달하거나 유의미한 증가가 없을 때까지 반복하며, 최대 K개의 후보를 선택한다. 최종 바닥 마스크 $\hat{M}$은 선택된 후보들의 합집합으로 결정된다.

$$S^* = \arg\max_{S\subseteq R} cov\left(\bigcup_{j\in S} M_j, M_{seed}\right) \quad (4)$$

바닥의 분할 후보들이 공간적으로 인접하고 유사한 특성을 공유하므로, 탐욕적 선택만으로도 전역 최적에 근접한 결과를 효과적으로 산출할 수 있다. 또한 복수의 바닥 영역이 공간적으로 분리된 경우에도 반복 선택을 통해 모두 포함할 수 있다.

Ⅳ. 실험 결과

4.1 실험 환경

실험에 사용된 데이터셋은 그림 2와 같이 단안 RGB 카메라를 탑재한 이동 로봇으로 직접 촬영한

실내 이미지와 인터넷에서 수집한 이미지로 구성된다. 표 1과 같이 이미지는 복도, 사무실, 회의실 등 다양한 유형의 실내 공간을 포함하며, 바닥의 재질과 무늬, 조명 조건도 다양하다. 제안 방법은 추가적인 학습 없이 동작하는 제로샷 방식이므로, 수집된 이미지 약 3,000장을 테스트 데이터로 활용하였다. 수집된 데이터셋을 대상으로 SAP의 앵커 영역 비율 및 Seed Mask 이진화 임계값, CGS의 커버리지 임계값 Γ와 최대 선택 후보 수 K를 포함한 하이퍼파라미터를 반복 실험을 통해 경험적으로 결정하였다. 이 중 픽셀 단위 바닥 영역을 직접 라벨링하여 GT 마스크를 생성하였고, 이를 기반으로 최종 정량 평가를 수행하였다.

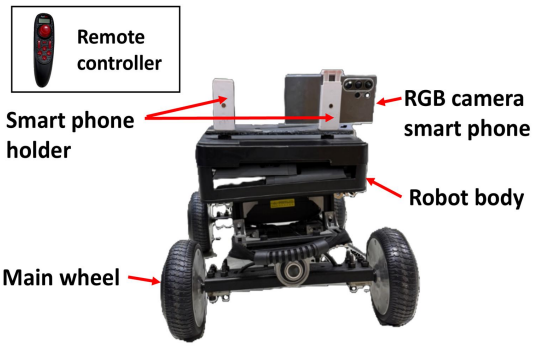

그림 2. 실험에 사용된 주행 로봇
Fig. 2. Overview of the experimental mobile robot

표 1. 데이터셋 구성
Table 1. Composition of the dataset

Category	Item	Ratio(%)
Collection method	Self-captured	27%
	Web-collected	73%
Indoor space type	Corridors, offices, meeting rooms, etc.	
Total	Total(test set)	100%

실험은 NVIDIA GeForce RTX 4070ti GPU를 탑재한 컴퓨터에서 Python 3.11, PyTorch 2.5.1, CUDA 12.1, cuDNN 9.1 환경에서 수행되었다. DINOv3는 ViT-L/16 모델을, SAM2는 sam2_hiera_large 모델을, Depth Anything V2는 ViT-L 모델을 사용하였다. DINOv2(패치 크기 14)와 DINOv3(패치 크기 16)의 패치 크기를 고려하여, DINO 기반 실험에서 입력 이미지는 896×896픽셀로 리사이즈하여 사용하였다.

평가 지표로는 IoU, mIoU, mDice, Global Dice, Precision, Recall, F1을 사용하였다. IoU, mIoU, mDice, Global Dice는 GT 마스크와 예측 마스크 간의 픽셀 수준 중첩 정도를 측정하여 바닥 경계의 분할 정밀도를 평가하는 지표이다. IoU는 두 영역의 교집합을 합집합으로 나눈 값을, Dice는 두 영역의 크기 합에 대한 중첩 영역의 2배 비율을 나타내며, mIoU와 mDice는 이미지별 평균값, Global Dice는 전체 픽셀을 통합하여 산출한 Dice 값이다. 또한 로봇 주행 안전성 관점에서 주행 가능 면적의 확보 여부를 평가하기 위해 이미지 단위 면적 비율 기반의 Precision, Recall, F1을 추가로 정의하였다. 예측 바닥 픽셀 수와 GT 픽셀 수의 비율이 80~120%이면 TP, 120% 초과이면 FP, 80% 미만이면 FN으로 판정한다. 특히 Recall은 바닥 영역의 누락 정도를 반영하므로 주행 가능 경로 확보와 밀접하게 관련되며, F1은 Precision과 Recall의 균형을 통해 과탐지와 미탐지를 함께 고려하는 핵심 지표로 사용하였다.

4.2 비교 모델 및 실험 설계

제안 방법의 유효성을 검증하기 위해 제로샷 조건 하에서 동작하는 5개의 비교 모델을 설계하였다. 비교 모델은 SAM2 단독 계열과 DINOv3/DINOv2 단독 계열로 구분되며, 제안 방법(모델 6)과의 대조를 통해 각 구성 요소의 기여도를 확인한다. 모든 모델은 학습 데이터, 파인튜닝, GT 마스크를 사용하지 않는 제로샷 조건에서 평가된다.

모델 1은 SAM2의 AMG를 통해 생성된 모든 후보 마스크를 합산하여 최종 예측으로 사용한다. 의미론적 판단 없이 이미지 전체 분할 결과를 그대로 활용하는 가장 단순한 기준 모델이다. 모델 2는 AMG로 생성된 후보 마스크 중 SAM2 자체 신뢰도 점수를 기준으로 상위 k개를 선택하여 합산한다. 모델 1 대비 노이즈 마스크를 일부 제거하지만 의미론적 기준 없이 신뢰도만으로 선택한다. 모델 3은 이미지 하단 일정 비율 영역(Bottom band)과 겹치는 SAM2 후보 마스크만을 선택하여 합산한다. 바닥이 이미지 하단에 위치한다는 위치 기반 휴리스틱을 활용한 접근법이다.

모델 4는 DINOv2의 패치 토큰을 활용하며, 이미지 하단 25% 영역의 패치 토큰을 앵커로 삼아 유사도 기반 히트맵을 생성한 뒤 k-means 클러스터링으로 바닥 후보 영역을 추정한다. 별도의 재조정(Retune) 없이 DINOv2 특징만을 단독으로 사용한다. 모델 5는 모델 4와 동일한 과정에서 백본을 DINOv2에서 DINOv3로 교체한 모델로, 아티팩트가 제거된 DINOv3의 정제된 패치 토큰을 활용한다. 모델 4와 5는 SAM2를 사용하지 않으므로 픽셀 수준 경계 정밀도가 패치 해상도에 제한된다.

모델 6은 본 연구에서 제안하는 방법으로, SAP와 CGS를 결합한다. 구체적으로 DINOv3의 패치 토큰에서 이미지 하단 단일 영역을 앵커로 하는 SAP를 통해 바닥 유사도 히트맵을 생성하고 Seed Mask를 추정한다.

이후 SAM2 AMG로 생성된 전체 후보 마스크 각각에 대해 히트맵 평균, 시드 커버리지, 하단 연결성, 면적 Prior, SAM2 신뢰도를 종합한 복합 점수를 산출하는 CGS를 적용하여 바닥 Prior와 정합성이 높은 후보 마스크를 선택·합산한다. Depth Anything V2[21]는 최종 마스크 위에 오버레이하여 주행 가능 경로의 원근 구조를 시각화하는 데만 활용되며 탐지 로직에는 관여하지 않는다.

4.3 정량 평가 결과

표 2는 6개 모델의 정량 평가 결과이며, 그림 3은 시각화 결과다. SAM2 단독 계열(모델 1~3)에서 가장 먼저 확인한 것은 Precision의 전반적인 취약함이었다. 의미론적 판단 없이 AMG 전체를 합산한 모델 1은 IoU 0.283, Precision·Recall·F1 모두 0.000을 기록하였다. 예측 마스크가 GT 범위를 크게 벗어났기 때문으로, 의미론적 판단 기준 없이 분할하면 바닥 탐지에 활용할 수 없음을 보여주는 결과였다. 신뢰도 상위 k개의 마스크를 고른 모델 2는 IoU 0.486으로 개선되었지만 Precision은 0.155에 머물렀다. SAM2 내부 신뢰도는 바닥과 비바닥을 구분하는 기준이 될 수 없음을 이 수치에서 확인할 수 있다. 모델 3은 이미지 하단 위치 기반 휴리스틱을 적용하여 Recall 0.983으로 높은 재현율을 달성하였으나, Precision은 0.498에 그쳤다. 이미지 하단에 위치한다고 해서 전부 바닥은 아니기 때문이다.

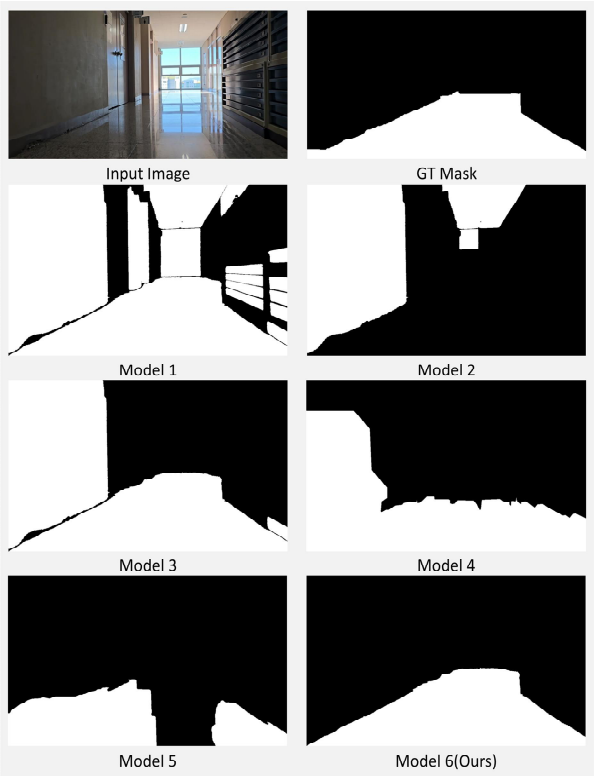

그림 3. 비교 실험 시각화 결과
Fig. 3. Comparison experiment visualization results

표 2. 모델의 전체 정량 평가 결과
Table 2. Overall quantitative evaluation results

Models		IoU	mIoU	mDice	Global dice	Precision	Recall	F1 score
SAM2	model1	0.283	0.326	0.480	0.441	0	0	0
	model2	0.486	0.558	0.677	0.654	0.155	0.757	0.257
	model3	0.619	0.717	0.804	0.765	0.498	0.983	0.661
DINO	model4(v2)	0.635	0.670	0.786	0.777	0.574	0.659	0.613
	model5(v3)	0.724	0.721	0.826	0.840	0.640	0.552	0.593
model6(Ours)		0.831	0.862	0.913	0.908	0.802	0.939	0.865

DINOv2를 도입한 모델 4에서 의미론적 특징의 효과를 수치로 확인할 수 있었다. Precision이 0.574로 SAM2 단독 계열보다 향상되었으며, DINO 특징이 바닥과 비바닥을 실제로 구분하는 데 유효하게 작용한다는 것을 보여주는 결과였다. 백본을 DINOv3로 교체한 모델 5에서는 IoU 0.724, Precision 0.640으로 더 향상되었다. 아티팩트가 없는 DINOv3의 깨끗한 특징이 바닥 후보를 더 정확하게 추정한다는 것을 직접 확인한 결과다. 다만 모델 5의 Recall은 0.552에 그쳤는데, 패치 단위 해상도 한계로 바닥 경계 근처 픽셀을 정교하게 추정하지 못한 것이다. 이러한 한계를 극복하기 위해 본 연구의 제안 방법은 SAM2를 결합하였다.

제안 방법인 모델 6은 비교 모델 중 가장 높은 성능을 기록하였다. IoU는 0.831로 모델 5(0.724) 대비 0.107, F1은 0.865로 모델 5(0.593) 대비 0.272만큼 개선되었다. 결과에서 주목할 점은 Precision 0.802와 Recall 0.939가 동시에 높게 나타난 점이다. 이는 SAP가 바닥과 유사도가 낮은 후보를 제거하고, CGS가 그중 실제 바닥과 겹치는 후보만 선별한 결과로 판단된다. 그림 4의 시각화에서도 이전 모델들에서 나타났던 과분할과 경계 불명확 문제가 사라졌음을 직접 확인하였다.

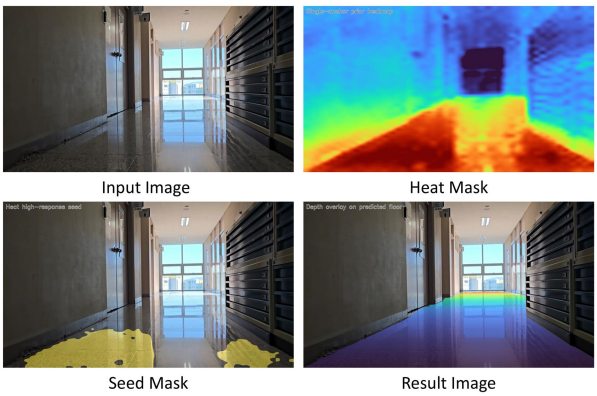

그림 4. 바닥 탐지 결과 시각화
Fig. 4. Visualization of floor detection results

V. 결론 및 향후 과제

본 논문에서는 LiDAR나 별도의 학습 데이터 없이 단일 RGB 이미지만으로 다양한 실내 환경에서 바닥을 탐지하는 제로샷 방법을 제안하였다. 실험을 통해 SAM2는 경계를 정밀하게 잡지만 무엇이 바닥인지 알지 못하고, DINOv3는 바닥의 의미를 이해하지만, 경계가 거칠다는 한계를 확인하였다. 이 두 모델의 상보적 관계를 설계에 반영하여, DINOv3의 패치 특징 유사도를 활용한 SAP로 바닥 후보 영역을 먼저 추정하고, 그 결과를 기준으로 SAM2 분할 후보 중 바닥과 정합성이 높은 마스크를 탐욕적으로 선택하는 CGS를 적용하는 구조로 결합하였다. 정량 평가에서 제안 방법은 mIoU 0.862, Recall 0.939, F1 0.865로 5개의 비교 모델 중 최고 성능을 달성하였으며, 두 모델을 결합하였을 때 각각의 한계가 실질적으로 해소된다는 것을 실험적으로 입증하였다.

향후 연구에서는 이미지 하단부에 바닥이 존재한다는 기하학적 가정에 대한 의존을 극복하기 위해, 고정된 위치 기반 앵커가 아니라 장면 이해(Scene understanding) 기반의 의미론적 판단을 통해 바닥 후보 영역을 추정하는 방식을 모색할 계획이다. 또한 깊이 추정 기법의 평면성 정보를 Prior 생성에 보조적으로 통합하여, 위치 가정에 대한 의존도를 낮추고 다양한 실내 환경에서의 일반화 성능을 향상시키고자 한다.

References

[1] IFR, "World Robotics 2024: Service Robots", International Federation of Robotics, https://ifr.org/img/worldrobotics/Press_Conference_2024.pdf. [accessed: May 01, 2026].

[2] IFR, "Executive Summary World Robotics 2025 – Service Robots", International Federation of Robotics, https://ifr.org/img/worldrobotics/Executive_Summary_WR_2025_Service_Robots.pdf. [accessed: May 01, 2026].

[3] R. Alqobali, M. Alshmrani, R. Alnasser, A. Rashidi, T. Alhmiedat, and O. M. Alia, "A survey on robot semantic navigation systems for indoor environments", Applied Sciences, Vol. 14, No. 1, Art. no. 89, Jan. 2024. https://doi.org/10.3390/app14010089.

[4] Y. Li, X. Zhao, and S. Schwertfeger, "Detection and utilization of reflections in LiDAR scans through plane optimization and plane SLAM", Sensors, Vol. 24, No. 15, Art. no. 4794, Jul. 2024. https://doi.org/10.3390/s24154794.

[5] C. Patruno, V. Renò, M. Nitti, N. Mosca, M. di Summa, and E. Stella, "Vision-based omnidirectional indoor robots for autonomous navigation and localization in manufacturing industry", Heliyon, Vol. 10, No. 4, Art. no. e26042, Feb. 2024. https://doi.org/10.1016/j.heliyon.2024.e26042.

[6] J. Rao, H. Bian, X. Xu, and J. Chen, "Autonomous visual navigation system based on a single camera for floor-sweeping robot", Applied Sciences, Vol. 13, No. 3, Art. no. 1562, Jan. 2023. https://doi.org/10.3390/app13031562.

[7] W. Ren, Y. Tang, Q. Sun, C. Zhao, and Q.-L. Han, "Visual semantic segmentation based on few/zero-shot learning: An overview", IEEE/CAA Journal of Automatica Sinica, Vol. 11, No. 5, pp. 1106-1126, May 2024. https://doi.org/10.1109/JAS.2023.123207.

[8] D. Teso-Fz-Betoño, E. Zulueta, A. Sánchez-Chica, U. Fernandez-Gamiz, and A. Saenz-Aguirre, "Semantic segmentation to develop an indoor navigation system for an autonomous mobile robot", Mathematics, Vol. 8, No. 5, Art. no. 855, May 2020. https://doi.org/10.3390/math8050855.

[9] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 44, No. 7, pp. 3523-3542, Jul. 2022. https://doi.org/10.1109/TPAMI.2021.3059968.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Venice, Italy, pp. 2980-2988, Oct. 2017. https://doi.org/10.1109/ICCV.2017.322.

[11] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, pp. 9396-9405, Jun. 2019. https://doi.org/10.1109/CVPR.2019.00963.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything", Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Paris, France, pp. 3992-4003, Oct. 2023. https://doi.org/10.1109/ICCV51070.2023.00371.

[13] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos", Proc. Int. Conf. Learn. Represent. (ICLR), Singapore, pp. 28085-28128, Apr. 2025. https://doi.org/10.48550/arXiv.2408.00714.

[14] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 43, No. 11, pp. 4037-4058, Nov. 2021. https://doi.org/10.1109/TPAMI.2020.2992393.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations", Proc. Int. Conf. Mach. Learn. (ICML), Online, pp. 1597-1607, Jul. 2020. https://doi.org/10.48550/arXiv.2002.05709.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale", Proc. Int. Conf. Learn. Represent. (ICLR), Online, May 2021. https://doi.org/10.48550/arXiv.2010.11929.

[17] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers", Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, pp. 9630-9640, Oct. 2021.

https://doi.org/10.1109/ICCV48922.2021.00951.

[18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision", Trans. Mach. Learn. Res., pp. 1-31, 2024. https://doi.org/10.48550/arXiv.2304.07193.

[19] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers", Proc. Int. Conf. Learn. Represent. (ICLR), Vienna, Austria, pp. 2632-2652, May 2024. https://doi.org/10.48550/arXiv.2309.16588.

[20] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "Dinov3", arXiv preprint, arXiv:2508.10104, Aug. 2025. https://doi.org/10.48550/arXiv.2508.10104.

[21] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2", Adv. Neural Inf. Process. Syst. (NeurIPS), Vol. 37, pp. 21875-21911, Dec. 2024. https://doi.org/10.48550/arXiv.2406.09414.

저자소개

최 윤 정 (Yunjeong Choi)

2023년 3월 ~ 현재 : 경기대학교 인공지능전공 학사과정
관심분야 : 컴퓨터 비전, 딥러닝, 인공지능

이 세 훈 (Sehun Lee)

2025년 2월 : 경기대학교 인공지능전공(학사)
2025년 3월 ~ 현재 : 경기대학교 컴퓨터과학과 석사과정
관심분야 : 컴퓨터 비전, 딥러닝, 인공지능

김 태 훈 (Taehoon Kim)

2025년 8월 : 경기대학교 인공지능전공(학사)
2025년 9월 ~ 현재 : 경기대학교 컴퓨터과학과 석사과정
관심분야 : 컴퓨터 비전, 딥러닝, 인공지능

안 준 호 (Junho Ahn)

2006년 2월 : 홍익대학교 컴퓨터공학과(공학사)
2008년 2월 : 연세대학교 컴퓨터공학과(공학석사)
2013년 12월 : University of Colorado Boulder 컴퓨터과학과(공학박사)
2013년 12월 ~ 2017년 1월 : 한국전자통신연구원(ETRI) 선임연구원
2017년 2월 ~ 2023년 2월 : 한국교통대학교 컴퓨터공학전공 부교수
2023년 3월 ~ 현재 : 경기대학교 AI컴퓨터공학부 부교수
관심분야 : 컴퓨터 비전, 딥러닝, 인공지능