사물인터넷융복합논문지 Vol. 12, No. 4, pp. 221-230, 2026 https://doi.org/10.20465/KIOTS.2026.12.4.221

비정형 실외 환경의 불규칙한 지형에서 안전한 경로 계획을 위한 제로샷 비전 알고리즘

이소원[1], 김태훈[2], 안준호[3*]
[1]경기대학교 인공지능전공 학생, [2]경기대학교 컴퓨터과학전공 석사과정, [3]경기대학교 AI컴퓨터공학부 교수

A Zero-Shot Vision Algorithm for Safe Path Planning on Irregular Terrain in Unstructured Outdoor Environments

Sowon Lee[1], Taehoon Kim[2], Junho Ahn[3*]
[1]Student, Dept. of Artificial Intelligence, Kyonggi University
[2]M.S. Student, Dept. of Computer Science, Kyonggi University
[3]Professor, Dept. of AI Computer Engineering, Kyonggi University

요 약 자율주행 로봇의 실외 운용 수요가 증가하면서, 대규모 데이터 구축과 추가 학습을 최소화하고 다양한 형태의 길에서 안전한 주행 가능 영역을 탐지하는 제로샷 기반 기술이 주목받고 있다. 저비용 단안 카메라를 활용한 제로샷 접근은 새로운 환경에도 적용할 수 있지만, 기존 기술은 경계가 불명확하고 노면 특성이 다양한 비정형 실외 환경에서 일반화 성능이 저하되는 한계가 있다. 본 연구에서는 이미지 분할, 깊이 추정, 특징 추출 알고리즘을 활용한 제로샷 기반 융합 알고리즘을 제안한다. 분할 알고리즘의 파라미터를 조정해 미세 장애물까지 포함하는 고밀도 마스크를 생성하고, 깊이 추정 모델로 픽셀별 상대적 거리를 추정한다. 특징 추출을 통해 주행 경로의 높은 활성화 마스크를 분류한 뒤, 분포 기반 코사인 유사도와 계층적 클러스터링으로 과탐지된 주행 가능 영역을 통합하여 안정적인 탐지 결과를 도출한다. 최종적으로 각 마스크에 평균 깊이 값을 할당해 공간 정보를 반영하며, 자체 구축 데이터셋에서 정확도 92.6%, ORFD 벤치마크에서 정확도 95.5%로 유효성을 검증하였다.

주제어 : 안전한 주행 경로, 장애물 탐지, 제로샷 비전, 깊이 추정, 비정형 실외 환경

Abstract As the demand for outdoor autonomous mobile robots increases, zero-shot methods that identify safe traversable areas while minimizing dataset construction and additional training are gaining attention. However, existing monocular vision-based approaches often show limited generalization in unstructured outdoor environments with ambiguous boundaries and diverse surface conditions. This study proposes a zero-shot fusion algorithm integrating image segmentation, depth estimation, and feature extraction to distinguish obstacles and hazardous regions without additional training. The segmentation parameters are adjusted to generate dense masks that include small obstacles, while a depth estimation model calculates pixel-wise relative distances. Feature extraction is then used to identify highly activated masks corresponding to traversable paths, and distribution-based cosine similarity and hierarchical clustering merge over-detected regions to improve detection stability. Finally, average depth values are assigned to each mask to incorporate spatial information. The proposed method achieved accuracies of 92.6% on a self-constructed dataset and 95.5% on the ORFD benchmark.

Key Words : Safe navigable path, obstacle detection, zero-shot vision, depth estimation, unstructured outdoor

This work was supported by Kyonggi University's Graduate Research Assistantship 2026, 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), 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), and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2026-25489409)
*교신저자 : 안준호(jha@kyonggi.ac.kr)
접수일 2026년 06월 29일 수정일 2026년 08월 13일 심사완료일 2026년 08월 18일

1. 서론

최근 자율주행 기술의 발전과 함께 등산로, 산책로, 공원과 같은 비정형 실외 환경에서 로봇의 안정적인 주행을 지원하기 위한 연구가 활발히 진행되고 있다. 이러한 환경은 포장도로와 달리 경로의 폭, 형태, 재질 및 경계가 일정하지 않으며, 나뭇가지, 풀, 돌과 같은 다양한 크기의 장애물이 불규칙하게 존재한다. 특히 캠퍼스, 공원 및 테마파크 등 반복적이거나 인력 수급이 어려운 업무 영역에서 실외 자율주행 로봇의 활용이 확대되면서 [1], 다양한 노면과 경로 형태에 대응할 수 있는 주행 환경 인식 기술의 필요성이 커지고 있다.

실외 자율주행을 위한 주행 가능 영역과 장애물 탐지에는 주로 카메라와 LiDAR 센서가 활용된다. LiDAR는 정밀한 거리 정보를 제공하지만 높은 하드웨어 비용으로 인해 소형 이동 로봇이나 생활밀착형 서비스 로봇에 적용하는 데 제약이 있다[2]. 이에 따라 카메라가 운전자의 주변 환경 가시성을 높이는 첨단 안전장치로 활용되고 있으며, 자율주행 레벨 향상에 따라 핵심 안전장치로서의 역할이 급성장하고 있다[3][4]. 그러나 단안 영상만으로는 객체와 경로의 거리를 직접 파악하기 어렵고, 경계가 불명확하거나 원거리에 위치한 영역에서는 주행 가능 여부를 안정적으로 판단하기 어렵다.

최근 범용 이미지 분할 모델과 단안 깊이 추정 모델의 발전으로 별도의 학습 없이 객체 영역과 상대적 거리 정보를 추출할 수 있게 되었다. 그러나 분할 결과는 파라미터 설정에 따라 크게 달라질 수 있으며, 그림자, 노면의 색상 변화 및 표면 무늬가 개별 객체로 분할되어 주행 가능 영역이 과도하게 나뉘는 문제가 발생한다. 또한 깊이 정보만으로는 나뭇가지, 풀, 돌과 같은 미세 장애물과 실제 주행 가능 영역을 정확히 구분하기 어렵다. 이러한 한계는 경로의 형태와 노면 조건이 일정하지 않은 비정형 실외 환경에서 더욱 두드러진다.

기존 연구는 주로 정형화된 도로 환경이나 특정 데이터셋을 기반으로 학습된 모델에 집중해 왔다. 이에 따라 학습 과정에서 관찰되지 않은 형태의 길이나 새로운 노면 조건에서는 탐지 성능이 저하될 수 있으며, 환경 변화에 따라 데이터를 추가로 구축하고 모델을 재학습해야 하는 부담이 있다. 따라서 대규모 학습 데이터와 추가 학습에 대한 의존도를 낮추면서도, 곡선형 경로, 협로, 분기점 및 불규칙한 노면에서 미세 장애물과 위험 영역을 구분하고 안전한 주행 가능 영역을 결정할 수 있는 방법이 필요하다.

이를 위해 본 연구에서는 이미지 분할, 단안 깊이 추정 및 시각 특징 분석을 결합한 제로샷 기반 융합 알고리즘을 제안한다. 제안 방법은 사전학습된 모델의 구조를 변경하거나 새로운 데이터로 재학습하지 않고, 단안 카메라 영상으로부터 분할 정보, 상대적 깊이 및 시각적 특징을 통합한다. 먼저 분할 파라미터를 조정하여 미세 장애물까지 포함하는 고밀도 마스크를 생성하고, 픽셀별 상대적 깊이와 각 마스크의 평균 깊이를 산출하여 공간 정보를 반영한다. 이후 분포 기반 코사인 유사도와 계층적 클러스터링을 활용하여 시각적으로 유사한 마스크를 선택적으로 병합함으로써, 그림자와 노면 변화로 인한 과분할을 완화하면서 실제 장애물 영역은 유지한다.

본 연구의 기여점은 다음과 같다. 첫째, 별도의 추가 학습 없이 이미지 분할, 상대적 깊이 추정 및 시각 특징 분석을 결합하여 다양한 비정형 실외 환경에서 주행 가능 영역과 장애물을 동시에 탐지한다. 둘째, 경로의 폭, 형태, 경계 및 노면 특성이 변화하는 환경에서도 미세 장애물과 위험 영역을 구분하여 안전한 주행 가능 영역을 결정한다. 셋째, 선택적 마스크 병합을 통해 그림자와 노면 변화로 인한 과탐지를 억제하면서 실제 장애물 영역을 보존한다. 넷째, 자체 구축 데이터셋과 공개 벤치마크를 활용하여 특정 환경이나 데이터셋에 대한 추가 학습 없이도 다양한 비정형 실외 환경에 적용 가능한 일반화 가능성을 검증한다.

본 논문의 구성은 다음과 같다. 제2장에서는 이미지 분할, 단안 깊이 추정, 특징 추출 및 주행 가능 영역 탐지와 관련된 선행 연구를 정리한다. 제3장에서는 제안 알고리즘의 전체 구조와 주요 구성요소를 설명한다. 제4장에서는 실험 환경과 평가 방법을 제시하고, 자체 구축 데이터셋과 공개 벤치마크를 활용하여 제안 방법의 성능과 일반화 가능성을 검증한다. 마지막으로 제5장에서는 연구 결과와 한계점 및 향후 연구 방향을 제시한다.

2. 관련 연구

객체 탐지와 의미론적 분할의 특성을 결합한 인스턴스 분할은 이미지 내 각 픽셀의 의미 정보를 바탕으로 개별 객체 영역을 식별하고 정밀한 마스크를 생성하는 기술이다. 기존 객체 탐지 방식보다 픽셀 수준의 세밀한 결과를 제공하며, 객체 간 경계를 명확히 구분할 수 있다는 장점이 있다. 트랜스포머 기반 방식은 전역 문맥 정보를 활용하여 분할 성능을 크게 향상시켰으며, Mask2Former[5]

는 마스크 어텐션을 통해 인스턴스 분할, 의미론적 분할, 전경 분할을 단일 아키텍처로 통합하였다. 그러나 이러한 학습 기반 방식들은 대규모 레이블 데이터에 대한 의존도가 높아 학습되지 않은 환경이나 객체 범주에 대한 적용이 어렵다. 이에 본 연구에서는 제로샷 방식으로 동작하는 SAM 2(Segment Anything Model 2)[6]를 활용하여 비정형 실외 환경에서 미세 장애물을 포함한 전체 영역의 후보 마스크를 자동 생성하고, 이를 주행 가능 영역과 장애물 영역 구분의 기반으로 활용한다.

단안 깊이 추정은 단일 RGB 이미지로부터 장면의 3차원 기하학적 구조를 추정하는 기술로, 스테레오 카메라나 LiDAR에 비해 하드웨어 구성이 단순하고 비용이 낮아 자율주행 및 로보틱스 분야에서 실용적인 대안으로 주목받고 있다. 생성 모델 기반 방식은 Stable Diffusion[7]과 같은 확산 모델의 시각적 사전 정보를 깊이 추정에 활용하며, Marigold[8]는 깊이 추정 태스크에 파인튜닝되어 합성 데이터만으로도 실제 장면에 대한 높은 일반화 성능을 보였다. 그러나 반복적인 확산 샘플링 과정으로 인해 추론 속도와 연산 비용 측면의 부담이 크다. 본 연구에서는 판별 모델 기반의 Depth Anything V2[9]를 활용하여 마스크 영역별 상대적 거리 정보를 추정하고, 주행 경로와 장애물 간의 공간적 관계를 정량적으로 표현하는 융합 표현 생성에 활용한다.

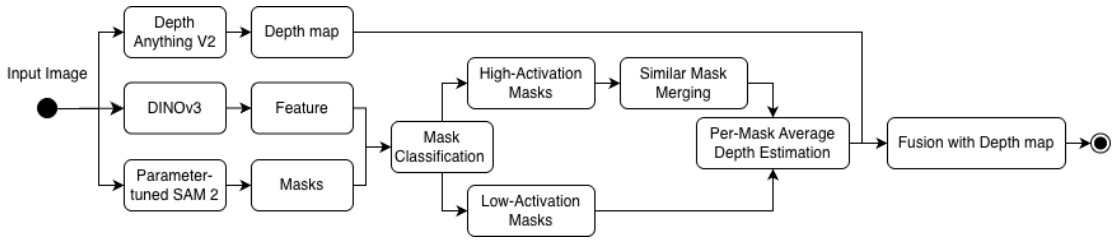
특징 추출은 입력 이미지로부터 의미론적·구조적 정보를 포함한 고차원 표현을 생성하는 과정으로, 자기지도 학습 기반 특징 추출은 레이블 없이도 강건한 시각 표현을 학습할 수 있어 도메인 변화에 유리하다. DINO(Self-Distillation with NO labels)v2[10]는 대규모 정제 데이터셋을 통해 특징 표현의 품질을 향상시켰으나, 모델 규모가 커질수록 밀집 특징의 품질이 저하되는 문제가 있었다. 이를 개선한 DINOv3[11]는 Gram anchoring 학습 단계를 통해 특징 맵의 노이즈를 줄이고 픽셀 단위 작업에서 향상된 성능을 보였다. 본 연구에서는 DINOv3를 특징 추출기로 활용하여 SAM 2로 생성된 마스크 간 특징 유사도를 분석하고, 그림자나 텍스처 변화로 인해 과도하게 분할된 영역을 의미 단위로 병합함으로써 과탐지를 억제하고 주행 가능 영역 탐지의 안정성을 높인다.

주행 가능 영역 및 장애물 탐지 연구는 객체 탐지 기반 방식과 분할 기반 방식으로 발전해 왔다. YOLO(You Only Look Once)[12] 계열 모델은 빠른 처리 속도와 높은 탐지 성능을 바탕으로 차량, 보행자 등 주요 장애물 탐지에 활용되었으며, 시맨틱 분할 기반 연구[13]는 도로와 비도로 영역을 픽셀 단위로 구분하는 데 사용되었다. 또한 차선 및 도로 경계 추출을 통한 주행 경로 탐지 연구[14]와 LiDAR 기반 3차원 포인트 클라우드 활용 연구[15]도 수행되었다. 그러나 기존 연구들은 대부분 차도나 포장도로와 같은 정형화된 환경을 대상으로 하며, 특정 데이터셋에 사전 학습된 모델에 의존하기 때문에 비정형 실외 환경에 대한 일반화 성능이 제한적이다. 최근에는 비전 파운데이션 모델을 활용한 주행 가능성 추정 연구가 활발히 제안되고 있다. V-STRONG[16]은 SAM의 ViT-H 인코더와 마스크 기반 대조 학습을 결합하여 온 트레일 및 오프 트레일 환경에서 제로샷 일반화 성능을 보였으며, ViTA[17]는 SAM 2에 학습 가능한 주행 가능성 프롬프트와 기하학적 지식 증류를 결합하여 험지 환경에서 강건한 성능을 보였다. WVN[18]은 자기지도 방식의 온라인 적응을 통해 환경별 재학습 없이 실시간 주행 가능성 추정을 수행하였고, CoNVOI[19]와 VLM-GroNav[20]는 비전-언어 모델을 활용하여 컨텍스트 인식 기반 실외 내비게이션 성능을 향상시켰다. 그러나 기존 학습 기반 방법은 사전 정의된 클래스나 픽셀 수준 분류에 의존하여 인스턴스 단위의 미세 장애물 분리와 거리 정보 부여에 한계가 있으며, RGB+LiDAR 융합 방식은 높은 성능에도 불구하고 고가의 하드웨어 의존성이 크다. 이에 반해 본 연구는 단일 RGB 입력만을 사용하여 추가 학습 없이 제로샷으로 동작하며, 인스턴스 단위 마스크 생성, 특징 기반 선택적 병합, 픽셀별 깊이 추정을 단계적으로 융합함으로써 주행 가능 영역과 미세 장애물을 동시에 탐지하고 각 마스크에 공간 거리 정보를 부여한다는 점에서 기존 방법들과 차별화된다.

3. 제안 방법

3.1 알고리즘 개요

본 연구는 비정형 실외 환경에서 단안 카메라를 이용하여 주행 가능 영역과 장애물을 동시에 탐지하는 제로샷 기반 융합 알고리즘을 제안한다. 그림 1은 본 논문에서 제안하는 비정형 실외 환경에서의 주행 가능 영역 및 장애물 탐지를 위한 전체 구조를 나타낸다. 제안하는 구조는 단일 RGB 이미지를 입력으로 받아 마스크 생성, 깊이 추정, 특징 추출 및 선택적 병합의 단계별 파이프라인으로 구성된다. 입력 이미지에 SAM 2를 적용하여 객체 단위의 마스크 집합을 생성하며, 파라미터를 조정하여 미세한 장애물까지 분할하도록 한다. 동시에 Depth Anything V2를 활용하여 각 픽셀의 상대적 거리 정보를



[Fig. 1] Algorithm State Diagram Overview

포함하는 깊이 맵을 생성함으로써 장면의 3차원적 구조를 반영한다. 이후 입력 이미지에 DINOv3를 이용하여 이미지의 패치 단위 특징을 추출하고, 특정 채널의 값을 기준으로 SAM 2에서 생성된 마스크를 높은 활성화 값 마스크와 낮은 활성화 값 마스크로 분류한다. 높은 활성화 값 마스크는 주행 가능 영역을 포함하며, 분포 기반 유사도 계산과 계층적 클러스터링을 통해 유사한 마스크들을 병합한다. 반면, 낮은 활성화 값 마스크는 병합 없이 유지하여 세부 구조를 보존한다. 이후 각 마스크 영역에 대해 평균 깊이 값을 계산하여 공간 정보를 부여하고, 이를 기존 깊이 맵과 결합함으로써 최종 결과를 시각화한다. 이를 통해 제안하는 방법은 추가 학습 없이 제로샷 추론이 가능하며, 과도한 분할로 인한 오탐지를 억제하면서도 미세 장애물에 대한 탐지 성능을 유지한다.

3.2 제로샷 기반 고밀도 마스크 생성

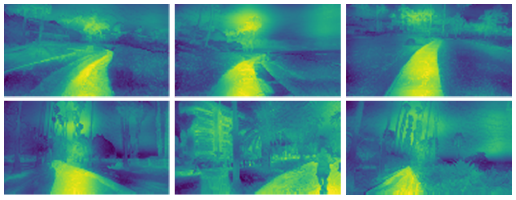
비정형 실외 환경에서는 미세 장애물이 주행 안정성에 큰 영향을 미치므로, 이를 정밀하게 탐지하는 것이 중요하다. 이를 위해 본 연구에서는 SAM 2의 자동 마스크 생성 과정에서 주요 파라미터를 조정하여 분할을 수행한다. points per side는 입력 이미지 상에서 샘플링되는 포인트의 밀도를 결정하는 파라미터로, 값이 증가할수록 더 촘촘한 위치에서 마스크가 생성된다. 이를 증가시킴으로써 작은 객체나 경계가 불명확한 영역까지 세밀하게 분할할 수 있으므로, 기본값 32에서 64로 증가시켰다. 또한 predicted IoU threshold는 생성된 마스크의 품질을 필터링하는 기준으로, 높은 값은 정밀한 마스크는 유지하지만 작은 객체가 제거되는 현상이 발생한다. 작은 객체도 포함하기 위해서 기본값 0.8에서 0.5로 감소시켜 사용하였다. 마스크의 안정성을 평가하는 지표인 stability score thresh 또한 값이 낮을수록 경계가 불완전한 마스크도 포함한다. 이를 기본값 0.95에서 0.3으로 완화함으로써 불규칙한 형태의 장애물까지 탐지할 수 있도록 하였다.

[Fig. 2] Parameter-tuned SAM 2 segmentation output

그림 2는 파라미터 조정을 적용한 SAM 2의 분할 결과를 보여준다. 상단의 원본 이미지와 비교하여 하단에 미세 장애물 및 경계 영역까지 세밀하게 분할된 고밀도 마스크가 생성됨을 확인할 수 있다. 이와 같은 파라미터 조정은 미세 장애물까지 포함하는 고밀도 마스크 생성을 가능하게 하지만, 동시에 그림자, 질감 변화, 지면의 미세 구조 등 비의미적 영역까지 분리되는 과분할이 발생할 수 있다. 따라서 본 단계는 정밀도를 높이는 동시에, 과탐지를 보완하기 위한 특징 기반 정제 과정을 수행한다.

3.3 선택적 마스크 병합

본 연구에서는 선택적 마스크 병합 기법을 제안한다. 기존 특징 기반 후처리는 각 마스크의 평균 특징 벡터를 직접 비교하는 방식을 사용하지만, 이는 마스크 크기나 픽셀 수의 절대적 차이에 영향을 받는다. 제안 기법은 마스크 내부 특징값의 히스토그램 분포 벡터를 사용하여



[Fig. 3] Feature activation maps

크기 불변 특성을 확보하고, 임계값 이상의 높은 활성화 구간에 가중치를 부여하는 가중 히스토그램 구조를 통해 주행 경로에 해당하는 특징 분포를 강조한다. 이를 통해 단순 평균 특징 비교 대비 그림자, 질감 변화로 인한 비의미적 과분할 영역을 더 효과적으로 통합할 수 있다. 먼저 DINOv3를 특징 추출기로 활용하여 입력 이미지의 패치 단위 특징을 추출하고, 다차원 특징 맵을 시각화하였다.

그림 3을 통해 특징맵의 특정 차원에서 일관되게 높은 활성화 값이 주행 경로에 나타나는 것을 알 수 있다. 이에 따라 해당 차원의 전역 최소값과 최대값을 각각 $g_{\min}$, $g_{\max}$라 할 때 분포에서 높은 활성화 값의 기준 임계값은 τ로 정의하고 높은 활성화 값 영역을 주행 경로의 특징이 활성화된 영역으로 정의한다. 마스크 M_i내부의 특징값 집합을 V_i로 정의하고 전체 특징 범위를 B개의 구간으로 분할하여 기본 히스토그램을 계산한다.

$$V_i = \{v_p \mid p \in M_i\} \tag{1}$$

이에 따라 정규화된 기본 히스토그램은 다음과 같이 정의된다. $h_i^{(b)}$는 b번째 구간의 정규화된 빈도값을 의미하며, 각 bin은 전체 범위를 균등 분할하여 정의된다. b번째 구간은 $[e_b, e_{b+1})$이며, ω_b는 구간의 중요도를 나타내는 가중치이고, ε은 0으로 나누는 문제를 방지하기 위한 작은 상수이다.

$$h_i^{(b)} = \frac{\omega_b \sum_{v \in V_i} 1(v \in [e_b, e_{b+1}))}{\sum_{k=1}^{B} \omega_k \sum_{v \in V_i} 1(v \in [e_k, e_{k+1})) + \varepsilon} \tag{2}$$

가중치 ω_b는 구간 시작값 e_b가 임계값 τ이상이면 w_h, 그렇지 않으면 1로 설정된다.

$$\omega_b = \begin{cases} w_h, & \text{if } e_b \geq \tau \\ 1, & otherwise \end{cases} \tag{3}$$

이를 통해 높은 활성화 구간에 더 큰 가중치를 부여하여, 해당 특징이 분포 벡터에 더 강하게 반영되도록 한다. 이는 경계에 걸쳐있는 마스크에만 작동하고, 분포가 명확히 나뉘어져 있는 마스크에는 재정규화로 인해 수학적 영향이 약하다.

이와 더불어, 높은 활성화 값 영역의 세부 분포를 보다 정밀하게 반영하기 위해 τ이상인 값들을 대상으로 별도의 세부 히스토그램을 구성한다. 이를 $h_{i,high}$라 할 때, 이를 위해 $V_{i,high}$를 정의하고, 구간 $[\tau, g_{\max}]$를 B_h개의 구간으로 분할하여 위와 같은 히스토그램 계산을 추가적으로 진행한다.

이후 최종 분포 벡터 d_i는 기본 히스토그램과 세부 히스토그램을 연결한다. 즉, d_i는 마스크 M_i 내부 특징값의 전반적인 분포와 높은 활성화 영역의 세부 분포를 동시에 반영하는 통합 벡터이다. 마스크 크기나 구간 빈도 절대값의 차이에 따른 영향을 줄이고 분포의 형태적 유사도에 기반한 비교가 가능하도록 L2 정규화를 적용하며, 정규화된 분포 벡터 $\tilde{d}_i$를 산출한다. 여기서 $\|d_i\|_2$는 분포 벡터의 L2 norm이며, ε은 분모가 0이 되는 것을 방지하기 위한 작은 상수이다. 이 과정을 통해 각 마스크의 분포 벡터는 단위 벡터로 변환되며, 이후 유사도 계산에서 절대적인 크기보다 분포의 형태적 유사성이 강조된다.

$$d_i = [h_i; h_{i,high}], \qquad \tilde{d}_i = \frac{d_i}{\|d_i\|_2 + \varepsilon} \tag{4}$$

정규화된 두 마스크 M_i와 M_j의 분포 벡터 $\tilde{d}_i$와 $\tilde{d}_j$ 간 유사도는 코사인 유사도를 이용하여 산출한다. 코사인 유사도는 두 벡터의 크기가 아닌 방향만을 비교하므로, 마스크 내 픽셀 수의 차이로 인한 히스토그램 빈도 절대값의 편향이 유사도 계산에 개입되지 않는다. 또한 L2 정규화로 단위 벡터화된 상태에서는 두 벡터의 내적이 코사인 유사도와 수학적으로 동치가 되며, 이를 거리로 변환한 값이 계층적 클러스터링의 거리 행렬에 직접 활용 가능하다. 정규화된 두 마스크의 분포 벡터간 유사도는 다음과 같이 산출되며, 두 마스크의 특징 분포가 유사할수록 s_{ij}는 커지고, 거리 D_{ij}는 작아진다.

$$s_{ij} = \frac{\tilde{d}_i^{\,T} \tilde{d}_j}{\|\tilde{d}_i\|_2 \|\tilde{d}_j\|_2}, \qquad D_{ij} = 1 - s_{ij} \tag{5}$$

본 연구에서는 이러한 거리 값을 기반으로 마스크 간 거리 행렬을 구성한 뒤, 계층적 클러스터링을 수행하여 유사한 마스크들을 동일 그룹으로 병합한다.

그림 4에서 주행 가능 영역에서 분할되었던 마스크들이 통합된 것을 확인할 수 있다. 반면 낮은 활성화 값 마스크는 후처리 없이 원래 상태로 유지하여 주변 장애물에 대한 정밀 탐지 성능을 보존한다. 최종적으로 병합된

[Fig. 4] Mask merging results before and after hierarchical clustering

높은 활성화 값 마스크와 낮은 활성화 값 마스크를 하나의 통합 맵으로 구성하고, 각 마스크에 Depth Anything V2로부터 추정된 평균 깊이 값을 할당한다. 이를 통해 깊이 맵과 마스크 결과를 오버레이한 시각화를 제공하며, 주행 가능 영역과 장애물의 공간적 분포를 직관적으로 확인할 수 있도록 한다. 이와 같은 선택적 병합 방식은 주행 경로에서 발생하는 그림자 및 텍스처 변화 등으로 인한 과탐지를 효과적으로 억제하면서도, 실제 장애물에 대한 정밀 탐지 성능을 유지할 수 있다는 점에서 비정형 실외 환경에 적합하다.

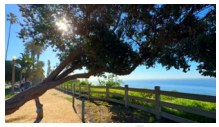
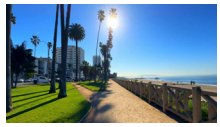
4. 실험 결과 및 분석

4.1 실험 환경

본 연구의 실험은 별도의 추가 학습 없이 제로샷 환경에서 수행되며, 활용된 개별 알고리즘 및 제안하는 융합 알고리즘의 성능을 비교 평가한다. 모든 실험은 NVIDIA GeForce RTX 4070 Ti GPU 환경에서 수행되었으며, CUDA 12.8 및 cuDNN 9.1.0을 사용하여 GPU 가속을 적용하였다. 소프트웨어 환경으로는 Python 3.11, PyTorch 2.1.0을 사용하였다. 본 연구에서 활용한 사전학습 모델은 SAM 2 (Hiera-Large), Depth Anything V2 (ViT-B), DINOv3 (ViT-B/16, LVD-142M)이며, 각각의 공식 사전학습 가중치를 그대로 사용하였다. 본 연구의 방법은 별도의 학습 없이 제로샷으로 동작하므로 추가적인 학습 설정은 적용하지 않았다. 이는 비전 기반 융합 기법의 탐지 성능 유효성 검증을 주된 목적으로 하며, 시스템 성능 지표 평가는 본 연구의 범위에 포함하지 않는다. 실시간 처리 성능은 사용되는 하드웨어 사양, 모델 경량화 기법에 따라 독립적으로 다루어지는 영역으로, 탐지 성능 중심의 기초 연구인 본 연구와는 연구 목적과 평가 범위를 달리한다.

본 연구의 실험을 위해 자체 데이터셋을 구축하였다. 기존 공개 데이터셋은 정형화된 도로 환경에 집중되어 있어 다양한 크기의 미세 장애물이 포함된 비정형 실외 환경을 충분히 반영하지 못한다는 문제가 있다. 이를 보완하기 위해 본 연구에서는 공개적으로 접근 가능한 유튜브의 4K 고해상도 자연 트레일 보행·주행 영상 7편을 데이터 출처로 활용하였다[21]-[27]. 선정한 영상은 국립공원 호숫가 산책로, 해변 보행로, 협곡 등산로, 삼나무 숲길, 온대우림 및 가을 낙엽 숲길 등 서로 다른 지형·식생과 조명 조건을 포함하며, 특정 환경에 편중되지 않은 일반화 검증용 데이터를 확보하였다. 각 영상에서 1초 간격으로 프레임을 추출한 뒤, 중복·저품질 프레임을 제거하여 최종 3,000장을 구성하였다. 데이터의 일관성을 유지하기 위해 모든 이미지는 1920×1080 해상도의 RGB 형식으로 통일하였다. 수집된 데이터셋은 장면의 특성에 따라 네 가지 클래스로 분류되었다. 각 클래스에

〈Table 1〉 Dataset Class Overview

Class	class1	class2	class3	class4
Image				
Description	Complex surroundings with dense obstacles	Simple environment with few obstacles	Obstacles directly blocking the traversable path	Branching path with 2-4 forks

대한 설명과 예시 이미지는 표 1에 제시하였다. 본 연구에서는 수집된 데이터셋을 기반으로 단독 모델 간의 성능 비교 실험을 진행하였다. 또한 공개 데이터셋 기반의 일반화 성능을 검증하기 위해 공개 데이터셋으로는 ORFD(Off-Road Freespace Detection)[28] 벤치마크를 활용하였다. 삼림, 농경지, 초원, 시골 등 다양한 오프로드 환경에서 맑음, 비, 안개, 눈과 같은 다양한 기상 및 조명 조건 하에 수집된 데이터셋으로, 각 이미지에 주행 가능 영역, 주행 불가능 영역, 접근 불가 영역이 픽셀 단위로 레이블링되어 있다. 본 연구에서는 ORFD의 공식 테스트셋을 사용하여 기존 방법들과의 비교 실험을 수행하였다.

실험은 혼동행렬을 통해 제안 알고리즘의 성능을 평가하였으며, SAM 2, Depth Anything V2와 비교를 통해 융합 전후의 성능 변화를 검토하였다. 이때 평가 지표로는 정밀도, 재현율, F1-점수, 정확도를 사용하였으며, TP, FP, FN, TN 분석을 통해 각 방법의 탐지 특성을 세부적으로 분석하였다. 또한 ORFD 기반 실험에서는 IoU 지표를 추가로 사용하였으며, 기존 방법들과의 정량적 성능 비교를 수행하였다.

4.2 자체 구축 데이터셋 실험 결과

자체 구축 데이터셋 기반으로 평가를 위해 혼동행렬을 활용하여 성능 지표를 산출하였다. 평가는 장애물 및 주행 가능 영역을 포함한 객체 단위로 수행되었으며, 각 항목은 다음과 같이 정의된다. TP는 주행 경로 및 장애물이 올바르게 탐지된 경우, FP는 장애물 또는 주행 가능 영역으로 예측되었으나 실제로는 해당하지 않는 경우로, 주행 경로가 과분할되어 분리된 영역이 장애물로 잘못 탐지되는 상황이 이에 해당한다. FN은 실제 장애물 또는 주행 가능 영역을 탐지하지 못한 미탐지 경우이며, TN은 배경 영역을 배경으로 올바르게 유지한 경우이다. 이를 기반으로 정밀도, 재현율, F1-점수, 정확도를 산출하였다.

표 2에서 확인할 수 있듯이, 전체 클래스에서 평균 0.926의 정확도와 0.875의 F1-점수를 달성하였으며, 비교 모델인 SAM 2 및 Depth Anything V2 단독 적용 대비 모든 지표에서 우수한 성능을 보인다. SAM 2 단독 적용의 경우 정밀도가 상대적으로 낮고, 이는 FP가 많아 주행 경로가 과분할된 영역을 장애물로 잘못 탐지하는 경향이 있음을 의미한다. 반면 Depth Anything V2 단독 적용의 경우 재현율이 낮아, 실제 장애물 및 주행 가능 영역을 놓치는 미탐지가 발생하는 경향을 보인다. 제안하는 방법은 두 모델의 상호보완적 특성을 융합함으로써 정밀도와 재현율 간의 균형을 효과적으로 달성하였다. 특히 장애물이 직접 주행 경로를 가로막는 클래스3과 경로가 복잡하게 분기되는 클래스4에서도 안정적인 탐지 성능을 유지하였으며, 이는 다양한 비정형 실외 환경에서 강인한 성능을 보임을 의미한다.

〈Table 2〉 Detection Performance Comparison with Baseline Methods

	Method	Acc	Pre	Rec	F1
class1	SAM 2	0.852	0.674	0.828	0.743
	DA V2*	0.845	0.722	0.671	0.696
	Ours	0.938	0.860	0.906	0.882
class2	SAM 2	0.818	0.647	0.846	0.732
	DA V2*	0.826	0.726	0.690	0.708
	Ours	0.919	0.838	0.899	0.868
class3	SAM 2	0.799	0.640	0.849	0.730
	DA V2*	0.819	0.744	0.706	0.725
	Ours	0.919	0.848	0.911	0.878
class4	SAM 2	0.832	0.648	0.832	0.728
	DA V2*	0.832	0.721	0.673	0.696
	Ours	0.927	0.843	0.901	0.871
total	SAM 2	0.825	0.652	0.838	0.733
	DA V2*	0.831	0.738	0.685	0.706
	Ours	0.926	0.847	0.905	0.875

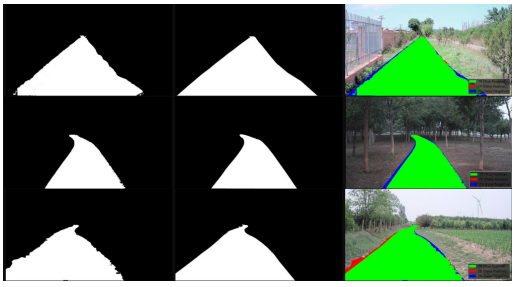
*Note: DA V2, Depth Anything V2

4.3 ORFD 데이터셋 실험 결과

일반화 성능을 검증하기 위해 공개 벤치마크인 ORFD 데이터셋을 활용한 비교 실험을 수행하였다. 비교 대상 모델은 ORFD 학습 데이터셋을 통해 학습 되었으며 RGB 단일 모달리티 기반 방법과 RGB+LiDAR 융합 기반 방법을 포함하여 성능 비교를 수행하였다. 표 3에서 확인할 수 있듯이, 본 방법은 RGB 단일 모달리티를 사용하고, 추가적인 학습을 하지 않았음에도 불구하고 지도학습 방식의 LiDAR를 추가로 활용하는 융합 기반 방법들과 대등하거나 우수한 성능을 달성하였다. 특히 정밀도 0.926과 F1-점수 0.918, IoU 0.874는 비교 모델 전체 중 가장 높은 수치로, 주행 가능 영역을 정밀하게 탐지함을 보여준다. 그림 5의 좌측은 제안 모델의 예측 마스크, 중앙은 정답 마스크, 우측은 원본 이미지에 예측 결과를 오버레이한 시각화이다. 우측 이미지의 범례에서 녹색은 정확히 탐지된 영역, 적색은 과탐지 영역, 청색은 미탐지 영역을 나타낸다.

〈Table 3〉 Quantitative Results on the ORFD Dataset

Method	Modality	Accuracy	Precision	Recall	F1-score	IoU
U-Net[29]	RGB	0.959	0.637	0.537	0.583	0.411
DLV+-MNet[30]	RGB	0.976	0.743	0.847	0.792	0.655
DLV+-R101[30]	RGB	0.980	0.781	0.871	0.824	0.700
FuseNet[31]	RGB+LiDAR	0.874	0.745	0.852	0.795	0.660
MFNet[32]	RGB+LiDAR	-	0.896	0.903	0.899	0.817
RTFNet[33]	RGB+LiDAR	-	0.842	0.967	0.900	0.818
SNE-RoadSeg[34]	RGB+LiDAR	0.938	0.867	0.927	0.896	0.812
OFF-Net[28]	RGB+LiDAR	0.945	0.866	0.943	0.903	0.823
Ours	RGB	0.955	0.926	0.910	0.918	0.874



[Fig. 5] Qualitative visualization results on the ORFD dataset

5. 결론

본 연구에서는 비정형 실외 환경에서 단안 카메라만을 활용하여 주행 가능 영역과 장애물을 동시에 탐지하는 제로샷 기반 융합 알고리즘을 제안하였다. 제안 방법은 범용 이미지 분할을 통한 고밀도 마스크 생성, 단안 영상 기반의 상대적 깊이 추정, 시각적 특징 유사도와 군집화를 활용한 선택적 마스크 병합을 단계적으로 결합하여 별도의 추가 학습 없이 동작하도록 구성하였다. 실험 결과, 제안 방법은 개별 기술을 단독으로 적용했을 때보다 주행 가능 영역의 경계를 안정적으로 구분하고, 미세 장애물을 유지하면서 과도하게 분할된 영역을 효과적으로 통합하는 것으로 나타났다. 또한 주행이 불가능한 영역을 주행 가능 영역으로 잘못 판단하는 오류를 줄여, 안전성이 중요한 실외 자율주행 환경에서 활용 가능성을 확인하였다. 공개 데이터셋과 교차 도메인 평가에서도 서로 다른 노면 형태와 기상 및 조명 조건에 대해 안정적인 탐지 성능을 유지함으로써, 특정 환경에 종속되지 않는 제로샷 일반화 가능성을 보여주었다. 이러한 결과는 고가의 거리 측정 센서나 대규모 학습 데이터에 대한 의존도를 낮추면서도, 사전에 학습되지 않은 다양한 형태의 실외 경로에서 안전한 이동 가능 영역을 탐지할 수 있음을 의미한다. 향후에는 보다 다양한 환경 조건에 대한 일반화 성능을 강화하고, 연산량 감소와 추론 속도 개선을 통해 실제 자율주행 로봇에 적용할 수 있는 실시간 탐지 시스템으로 확장할 필요가 있다.

REFERENCES

[1] Y.Shin, Robotics surpasses 200 units of autonomous delivery robot 'GAEMI'[Internet], https://zdnet.co.kr/view/?no=20251215113857

[2] Z.Wen, "Autonomous Driving Environmental Perception and Decision-Making Technology Roadmap Comparison," in Proceedings of the 4th International Conference on Computer, Electronic and Materials Engineering (ICCEME), Dalian, 2025, pp.215-227.

[3] Y.H.Jeong, "Autonomous Driving Camera Industry Trends and Standards," in 2024 Issue Report on Core Component Industry and Standard Trends for Autonomous Vehicles, Korean Standards Association (KSA), pp.68-83, 2024.

[4] S.Han, Tesla Was Dragged into 'Lidar-Radar War'[Internet], https://www.autoelectronics.co.kr/article/articleView.asp?idx=4463

[5] B.Cheng, I.Misra, A.G.Schwing, A.Kirillov, and R.Girdhar, "Masked-Attention Mask Transformer for Universal Image Segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, 2022, pp.1280-1289.

[6] N.Ravi, N.Gabeur, Y.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[Internet], https://arxiv.org/abs/2408.00714

[7] R.Rombach, A.Blattmann, D.Lorenz, P.Esser, and B.Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, 2022, pp.10684-10695.

[8] B.Ke, A.Obukhov, S.Huang, N.Metzger, R.C.Daudt, and K.Schindler, "Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, 2024, pp.9492-9502.

[9] L.Yang, B.Kang, Z.Huang, Z.Zhao, X.Xu, J.Feng, and H.Zhao, "Depth Anything V2," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, 2024.

[10] 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," arXiv:2304.07193, 2023. [Internet], https://arxiv.org/abs/2304.07193.

[11] 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:2508.10104, 2025. [Internet]. https://arxiv.org/abs/2508.10104.

[12] J.Redmon, S.Divvala, R.Girshick, and A.Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016, pp.779-788.

[13] M.Rasib, M.A.Butt, F.Riaz, A.Sulaiman, and M.Akram, "Pixel Level Segmentation Based Drivable Road Region Detection and Steering Angle Estimation Method for Autonomous Driving on Unstructured Roads," IEEE Access, Vol.9, pp.167855-167867, 2021.

[14] H.Yoo, U.Khan, and T.Tsourdos, End-to-End Deep Learning of Lane Detection and Path Prediction for Real-Time Autonomous Driving[Internet], https://arxiv.org/abs/2102.04738

[15] E.Horváth, C.Pozna, and M.Unger, "Real-Time LiDAR-Based Urban Road and Sidewalk Detection for Autonomous Vehicles," Sensors, Vol.22, No.1, p.194, 2022.

[16] W.Jung, J.Lee, and H.Myung, "V-STRONG: Visual Self-Supervised Traversability Learning for Off-Road Navigation," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Yokohama, 2024, pp.1-8.

[17] S.Hwang, J.Jung, H.Jung, J.Woo, J.Lee, J.Park, and H.Myung, ViTA: Vision Foundation Model for Traversability Estimation via Perspective-Diversified Training and Geometric Knowledge Distillation[Internet], https://arxiv.org/abs/2504.07452

[18] J.Frey, M.Mattamala, N.Chebrolu, C.Cadena, M.Fallon, and R.Siegwart, "Fast Traversability Estimation for Wild Visual Navigation," in Proceedings of Robotics: Science and Systems (RSS), Daegu, 2023.

[19] A.J.Sathyamoorthy, K.Weerakoon, T.Guan, M.Elnoor, and D.Manocha, "CoNVOI: Context-aware Navigation using Vision Language Models in Outdoor and Indoor Environments," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Yokohama, 2024, pp.13837-13844.

[20] M.Elnoor, K.Weerakoon, G.Seneviratne, R.Xian, T.Guan, M.K.M.Jaffar, V.Rajagopal, and D.Manocha, "VLM-GroNav: Robot Navigation Using Physically Grounded Vision-Language Models in Outdoor Environments," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Atlanta, 2025, pp.2391-2398.

[21] 4K Relaxation Channel, 2 HRS Virtual Walk around Emerald Lake, Yoho National Park - 4K Nature Walking Tour + Birds Chirping[Internet], https://youtu.be/U1j-eO3-aR8

[22] Strings And Wings 61, Wailea Beach Path, Maui, Hawaii, DJI Osmo 4K[Internet], https://youtu.be/SYIEFsP8Hy0

[23] 4K Relaxation Channel, Amazing Bryce Canyon Virtual Hike - 4K Footage for Fitness Equipment/Training Simulators - 1.5 HRS[Internet], https://youtu.be/yWbR6N5tYaQ

[24] Ambient Exploration, Muir Woods Early Morning, Quiet Walk in Redwood Forest | Mill Valley, California[Internet], https://youtu.be/AkG7Ogwh_Vs

[25] 4K Relaxation Channel, Walking the Autumn Forest Trails of North Cascades - 4K Virtual Hike with Relaxing Forest Sounds[Internet], https://youtu.be/e04zAMupq7E

[26] 4K Relaxation Channel, 4K Virtual Hike - Amazing Nature Scenery with Soothing Music - Sol Duc Falls Nature Trail[Internet], https://youtu.be/q1m27R0tAKM

[27] Simply Hiking, The Most Magical Rainforest Walk | Walking Ambience | Temperate Rainforest | Philosopher Falls[Internet], https://youtu.be/JkErW9Y5g-k

[28] C.Min, W.Jiang, D.Zhao, J.Xu, L.Xiao, Y.Nie, and B.Dai, "ORFD: A Dataset and Benchmark for Off-Road Freespace Detection," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Philadelphia, 2022, pp.2532-2538.

[29] O.Ronneberger, P.Fischer, and T.Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, 2015, pp.234-241.

[30] L.C.Chen, Y.Zhu, G.Papandreou, F.Schroff, and H.Adam, "Encoder-Decoder with Atrous Separable

Convolution for Semantic Image Segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), Munich, 2018, pp.801-818.

[31] C.Hazirbas, L.Ma, C.Domokos, and D.Cremers, "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture," in Proceedings of the Asian Conference on Computer Vision (ACCV), Taipei, 2016, pp.213-228.

[32] Q.Ha, K.Watanabe, T.Karasawa, Y.Ushiku, and T.Harada, "MFNet: Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scenes," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, 2017, pp.5108-5115.

[33] Y.Sun, W.Zuo, and M.Liu, "RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes," IEEE Robotics and Automation Letters, Vol.4, No.3, pp.2576-2583, 2019.

[34] R.Fan, H.Wang, P.Cai, and M.Liu, "SNE-RoadSeg: Incorporating Surface Normal Information into Semantic Segmentation for Accurate Freespace Detection," in Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, 2020, pp.340-356.

이 소 원(Sowon Lee) [준회원]

■ 2023년 3월 ~ 현재 : 경기대학교 인공지능전공 학부과정

〈관심분야〉
컴퓨터 비전, 딥러닝, 인공지능

김 태 훈(Taehoon Kim) [준회원]

■ 2025년 8월 : 경기대학교 인공지능전공 (학사)
■ 2025년 9월 ~ 현재 : 경기대학교 컴퓨터과학과 (석사과정)

〈관심분야〉
컴퓨터 비전, 딥러닝, 인공지능

안 준 호(Junho Ahn) [정회원]

■ 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, 공간지능, 인공지능