Check for updates

Journal of Institute of Control, Robotics and Systems (2025) 31(5):496-501
https://doi.org/10.5302/J.ICROS.2025.25.0026 ISSN:1976-5622 eISSN:2233-4335

인간-로봇 핸드오버를 위한 생성형 AI 활용 연구

Using Generative AI for Human-robot Handover

박 무 열, 박 규 민*

(Mu Yeol Park[1] and Kyu Min Park[1,*])

[1]Department of Artificial Intelligence and Robotics, Sejong University

Abstract: Recent advancements in human-robot interaction (HRI) have highlighted the importance of natural and efficient human-robot handover techniques, particularly in the field of industrial and service robotics. Traditional learning-based handover methods rely on large-scale data collection, which is costly and time-consuming. In this paper, we explore the use of generative AI, specifically OpenAI's Sora, to generate synthetic video for training a handover recognition model. Based on the synthetic training data, we propose a framework that can be used for the following three key handover scenarios: recognizing H2R (human-to-robot) handover initiation, recognizing R2H (robot-to-human) handover initiation, and object searching for R2H handover. The model is based on the YOLO11n architecture, achieving high accuracy in recognizing the user's hand and object within the three handover scenarios. Experimental results demonstrate that the model trained only with synthetic data can effectively generalize to real-world handover scenarios with a precision of 0.94 and recall of 0.97. This paper highlights the potential of generative AI to change the data collection paradigm for robot learning, in the context of efficiency and expansion to various scenarios. Future work will focus on extending our approach to diverse objects, users, and handover scenarios.

Keywords: human-robot interaction, handover, generative AI, synthetic data, object recognition

I. 서론

최근 로봇과 사용자(사람) 간의 협업이 더욱 중요해짐에 따라 인간-로봇 상호작용(human-robot interaction, HRI)의 관점에서 자연스럽고 효율적인 핸드오버(handover) 기술이 요구되고 있다[1]. 특히 제조업, 물류, 서비스 로봇 분야에서 인간-로봇 협업의 수요가 급증하고 있으며, 글로벌 협동로봇 시장은 2030년까지 연평균 30% 이상의 성장이 예상된다[2].

인간-로봇 핸드오버는 로봇과 사용자 간의 물건 전달을 의미하며, 이 과정에서 로봇은 사용자의 의도를 정확하게 인식하고 이에 대응할 수 있어야 한다. 특히 핸드오버 과정에서는 물체의 크기, 형태, 무게뿐만 아니라 사용자의 자세, 움직임 속도, 접근 방향 등 다양한 변수들을 고려해야 한다. 이는 AI 등 데이터 기반 기술을 활용하는 경우 데이터 수집의 복잡성을 증가시킨다[3][4]. 하지만 기존의 핸드오버 연구 결과는 대부분 학습된 상황 및 사용자에 대해서만 최적의 성능을 보이며, 각 상황과 사용자에 맞춰 실제 영상(또는 이미지) 데이터를 활용한 학습이 필요하다.

이에 따라 자연스럽고 효율적인 핸드오버 기술을 구현하기 위해서는 대량의 데이터를 수집하고 레이블링하는 데 많은 시간과 비용이 필요하며, 현재로서는 다양한 핸드오버 상황, 사용자의 사용 패턴 등을 충분히 반영하지 못한다는 한계가 있다. 이러한 데이터 수집의 어려움은 로봇 개발 및 배포 과정에서 상당한 병목 현상을 야기하고 있다.

최근 생성형 AI (generative AI)의 발전은 이러한 문제를 일부 해결할 새로운 가능성을 제시하고 있다. 생성형 AI를 이용하면 현실과 유사한 합성(synthetic) 데이터를 대량으로 생성할 수 있으며, 다양한 환경과 조건을 고려한 학습 데이터 구축이 가능하다. 특히 OpenAI의 Sora는 기존 이미지 생성 AI와 달리 시간에 따른 연속적인 동작을 자연스럽게 표현할 수 있어, 핸드오버와 같은 동적인 상호작용을 담은 학습 데이터 생성에 적합하다. 본 연구에서는 Sora를 활용하여 생성한 합성 영상 데이터를 기반으로 인간-로봇 핸드오버 인식 모델을 학습하고, 실제 영상에서도 이 모델이 효과적으로 작동하는지 검증하고자 한다.

본 연구에서는 세부적인 인간-로봇 핸드오버 시나리오를 정의하고, Sora를 통해 생성한 데이터를 학습 데이터로 활용하는 방안을 제안한다. 특히, 영상 내에서 손과 물체의 조합을 인식하여 다음과 같은 세 가지 주요 상황을 모델이 판단할 수 있도록 설계한다. 1) 손과 물체가 함께 인식되는 경우, 사용자가 로봇에게 물체를 전달하고자 하는 H2R (human-to-robot) 핸드오버 상황임을 판단할 수 있다. 2) 사용자의 손만 인식되는 경우, 사용자가 로봇으로부터 물건을 받고자 하는 R2H (robot-to-human) 핸드오버 상황임을 판단할 수 있다. 3) 추가로 R2H 핸드오버를 위해서는 전달 대상이 되는 물체를 탐색

*Corresponding Author

Manuscript received February 6, 2025; revised March 24, 2025; accepted April 7, 2025

박무열: 세종대학교 AI로봇학과 석사과정(parkanduf@naver.com, ORCID 0009-0005-3216-7440)
박규민: 세종대학교 AI로봇학과 조교수(kyuminpark@sejong.ac.kr, ORCID 0000-0003-1986-7520)

※ 본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성 지원(IITP-2025-RS-2024-00436528, 50%)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2024-00457702, 50%)을 받아 수행된 연구임.

Copyright© ICROS 2025

하는 과정이 있는데, 이는 물체만 인식되는 상황에 해당한다. 이러한 접근 방식은 기존의 데이터 수집 방식과 비교하여 데이터 생성의 효율성, 다양한 시나리오 확장 가능성 등의 장점이 있으며, 실제 환경에서의 활용 가능성을 높일 수 있다.

본 연구의 주요 목표는 합성 영상 데이터만을 이용하여 학습한 모델이 핸드오버를 위한 실제 상황(영상)에서도 객체를 효과적으로 인식할 수 있는지 검증하는 것이다. 본 연구는 단순히 데이터 증강이 아닌, 실제 데이터 수집 과정 자체를 생성형 AI로 대체할 가능성을 검증한다는 점에서 기존 연구들과 차별화된다. 이러한 접근 방식이 성공적으로 검증된다면, 로봇 학습을 위한 데이터 수집 패러다임에 변화를 줄 수 있을 것으로 기대한다.

본 논문의 구성은 다음과 같다. 2장에서는 인간-로봇 핸드오버, 생성형 AI 기반 데이터 관련 연구를 검토하고 본 연구의 차별성을 설명한다. 3장에서는 본 연구의 데이터 구축 과정과 모델 학습 방법을 설명하고, 4장에서는 실험 결과를 제시하고 모델 성능을 분석한다. 마지막으로 5장에서는 결론과 향후 연구 방향을 논의한다.

II. 관련 연구

1. 인간-로봇 핸드오버 연구

인간-로봇 상호작용에서 핸드오버 기술은 협동로봇의 핵심 요소 중 하나로 연구되고 있다[1]. 핸드오버는 단순한 물체 전달을 넘어, 사용자의 의도 파악, 물체의 안정적인 전달, 상황 변화에 따른 적응성 등을 요구하는 복잡한 과정이다. 기존 연구에서는 주로 RGB 카메라, 깊이 센서, 힘/토크 센서 등을 활용하여 핸드오버를 인식하고 로봇이 반응하도록 설계되었다.

[5]에서는 로봇에서 인간으로 물체를 전달하는 핸드오버 과정에서의 접촉 패턴을 분석하고 3D 접촉 지도를 활용하여 최적의 핸드오버 전략을 설계하는 연구가 수행되었다. 사용자가 로봇으로부터 물체를 더 자연스럽고 편하게 받을 수 있도록 로봇의 그립 위치와 전달 포즈를 조정하는 알고리즘을 제안하였으며, 다양한 일상 물체에 대한 실험을 통해 검증하였다. 하지만 이러한 연구들은 데이터 수집 및 처리 과정에 높은 비용이 소요된다.

[6]에서는 핸드오버 시 사용자의 손 자세를 인식하기 위해 딥러닝 모델을 적용하였으며, 제스처 인식을 활용하여 인간이 로봇에게 물체를 전달하려는 의도를 효과적으로 파악하는 방법을 연구하였다. 이들은 RGB 이미지를 기반으로 신체 키포인트(key points)를 추출하고, 이를 특징(feature) 벡터로 변환하여 로봇이 자연스러운 핸드오버를 수행할 수 있도록 설계하였다. 그러나 사용자의 개별적인 차이를 반영하기 어려우며, 해당 연구 또한 다양한 실제 데이터의 확보가 필요하다.

이처럼 기존 연구들은 대부분 실제 데이터 수집에 의존하며, 학습된 특정 환경에서만 높은 성능을 보이는 경향이 있다. 본 연구에서는 생성형 AI를 활용하여 학습 데이터를 생성함으로써 이러한 문제를 해결하는 방안을 제시하고자 한다.

2. 생성형 AI 기반 핸드오버 데이터 구축

최근 딥러닝 기반의 컴퓨터 비전 모델이 발전함에 따라, 데이터 증강 및 합성 데이터 생성이 활발히 연구되고 있다. 특히 GAN (Generative Adversarial Network)과 Diffusion 모델을 활용한 데이터 생성 기술이 주목받고 있으며, 이는 다양한 환경에서 학습 데이터의 부족 문제를 해결하는 데 유용하게 활용되고 있다.

[7]에서는 GAN을 활용하여 핸드오버에서의 손 자세 인식 모델을 개선하는 연구를 수행하였다. 그들은 제한된 손 데이터 세트를 보완하기 위해 RGB 이미지를 기반으로 깊이 이미지를 합성하여 3D 손 자세를 추정하는 DGGAN (Depth-image Guided GAN)을 개발하였으며, 이를 통해 다양한 각도와 조명 조건에서의 합성 데이터를 생성하여 모델의 일반화 성능을 향상했다.

본 연구 또한 합성 데이터를 이용해 학습한 모델이 실제 데이터에서도 성능을 발휘할 수 있는지를 검증한다는 점에서 위 연구와 유사한 접근 방식을 가진다. 생성형 AI로 완전히 대체된 데이터 수집 방식이 핸드오버와 같은 복잡한 상호작용 상황에서도 효과적으로 적용될 수 있음을 확인하고자 한다.

3. 생성형 AI를 활용한 합성 데이터 연구

기존 연구에서는 주로 GAN과 Diffusion 모델을 활용한 이미지 생성이 이루어졌으나, 영상 단위의 생성에는 한계가 있었다. 하지만 OpenAI의 Sora와 같은 영상 생성 AI의 발전으로, 시간에 따른 연속적인 동작을 반영한 데이터 생성이 가능해졌다.

[8]에서는 텍스트 프롬프트를 활용한 이미지 생성 연구를 조사하고, 다양한 생성 모델을 비교·분석하여 합성된 이미지의 품질을 평가하였다. 또한 성능 평가 지표와 데이터 세트를 분석하여 관련 향후 연구 방향을 제시하였다. 본 연구에서는 세 가지 인간-로봇 핸드오버 시나리오를 정의하고, Sora를 활용하여 생성한 합성 데이터를 모델 학습에 활용하는 접근 방식을 제안한다.

[9]에서는 Sora로 생성된 합성 영상과 실제 영상 간의 시각적, 운동적, 기하학적 특성을 비교·분석하였다. 합성 영상만을 이용해 3D CNN (Convolutional Neural Network) 기반 탐지기를 학습시킨 후, 실제 영상에 대한 추가 학습 없이도 영상의 진위(합성과 실제를 구분)를 약 80%의 정확도로 판별할 수 있음을 보였다. 프레임 간 일관성 부족, 움직임의 부자연스러움, 깊이 정보의 불안정성 등이 주요 탐지 단서로 작용한다고 분석하였다.

[10]에서는 Sora 기반 합성 영상에서의 객체 인식 문제를 분석하고, YOLOv8 구조에 기반한 개선된 모델을 제안하였다. 기존 YOLOv8 모델의 완전 연결층을 제거하고, 앵커 박스(anchor box) 기반 위치 예측, 정제된 데이터 세트, 양자화 대응 구조를 도입해 모델을 재학습시켜, 기존 모델이 인식에 실패했던 캐릭터 및 모호한 객체들을 정확히 탐지하고 클래스 분류 정확도 또한 향상하였다.

4. 본 연구의 차별성

기존 인간-로봇 핸드오버 연구들은 실제 데이터 수집을 통해 학습 데이터를 구축하였으며, 생성형 AI 기반 데이터 관련 연구들은 일반적인 객체 탐지를 중점적으로 다루거나 Sora로 생성된 합성 영상을 실제 영상과 비교하여 영상 자체의 문제점을 분석하는 것에 집중하였다. 본 연구는 Sora를 통해 생성된 합성 영상 데이터를 세부적인 주요 핸드오버 상황 인식 학습에 활용하는 방식을 제안함으로써 더욱 복합적인 상황에서의 생성형 AI 기반 데이터 활용법을 제시한다.

III. 데이터 구축 및 모델 학습 방법

1. 손과 물체의 조합에 따른 데이터 구축 과정

1.1 손과 물체가 함께 인식되는 경우(H2R)

H2R 핸드오버는 사용자가 로봇에게 물건을 전달하는 과정이기에 일반적으로 사용자의 손과 물체가 함께 인식된다. 이 과정에서는 모델을 통해 사용자의 손과 물체를 각각 인식해야 하므로, 학습 데이터로 손이 물체를 쥐고 있는 영상이 필요하다. 데이터 생성을 위한 프롬프트는 “사람 손으로 사과를 쥐고 있는 모습을 위쪽에서 적당한 거리를 두고 찍은 영상 보여줘. 또한, 가만히 있는 게 아닌 다양한 각도에서 볼 수 있게 손목을 조금씩 돌려줘”이며, 이를 통해 5초 길이의 영상을 생성한다. 이렇게 생성된 합성 영상 데이터에서 프레임별 이미지를 추출하여 합성 이미지 데이터를 얻어낼 수 있다. 모든 영상은 30fps로 생성되어 각 영상당 150장의 합성 이미지 데이터가 추출된다. 손과 물체가 함께 있는 합성 영상 데이터는 총 4개를 생성하며, 최종적으로 추출되는 학습용 합성 이미지 데이터는 총 600장이다(그림 1 참조).

이미지 데이터 세트 생성 후, 모델 학습을 위해서는 이미지에 대응되는 레이블 파일 또한 필요하므로 우선 이미지와 동일한 파일명을 가진 레이블 파일을 만든다. 레이블 파일은 학습 이미지에서 바운딩 박스를 생성하여 각 객체의 클래스와 YOLO 포맷의 바운딩 박스 좌표 정보를 저장하는 파일을 의미한다. YOLO 포맷의 바운딩 박스 좌표 정보는 다음과 같다: 1) 바운딩 박스 중심의 X 좌표, 2) 바운딩 박스 중심의 Y 좌표, 3) 바운딩 박스의 너비, 4) 바운딩 박스의 높이. X 좌표와 박스 너비의 경우 이미지 너비에 대한 비율로, Y 좌표와 박스 높이의 경우 이미지 높이에 대한 비율로 표현되며, 모든 비율은 0 이상 1 이하의 값으로 표현된다. 각 이미지에 대해 사과(클래스 1)와 손(클래스 0)에 해당하는 바운딩 박스 좌표 정보를 함께 저장하고 있는 레이블 파일을 만든다(총 600개).

1.2 사용자의 손만 인식되는 경우(R2H)

사용자의 손만 인식되는 경우, 사용자가 로봇으로부터 물건을 전달받고자 함을 알 수 있다. 즉, R2H 핸드오버 상황임을 판단할 수 있다. 이를 위해서는 모델이 사용자의 손을 인식해야 하며, 이에 따라 사람 손만 있는 영상을 학습 데이터에 포함한다. 데이터 생성을 위한 프롬프트는 “사람의 손바닥과 손등을 돌려가면서 잘 보여주는 영상 보여줘. 이때 손이 전체적으로 잘 보이도록 적당한 거리를 두고 손가락은 5개로 사람

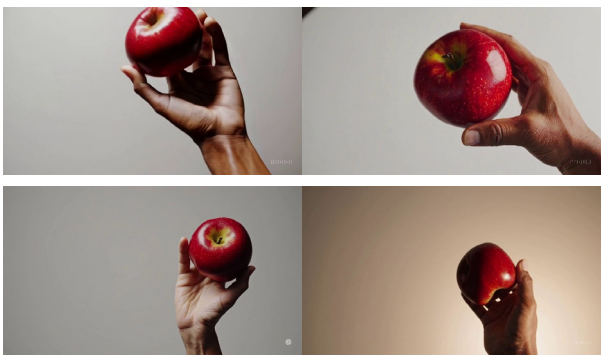


그림 1. 손이 물체를 쥐고 있는 합성 데이터.

Fig. 1. Synthetic data of a hand holding an object.

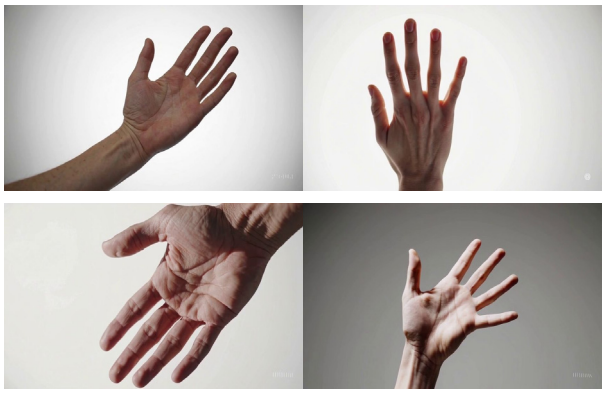


그림 2. 손만 있는 합성 데이터.

Fig. 2. Synthetic data with human hand only.

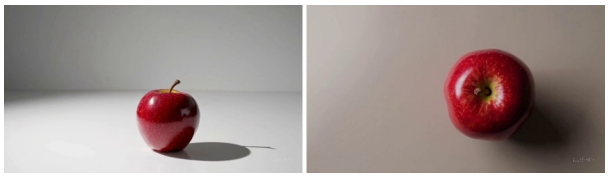


그림 3. 물체만 있는 합성 데이터.

Fig. 3. Synthetic data with object only.

손과 똑같은 모습으로 나타내줘”이며, 이를 통해 5초의 합성 영상 데이터를 생성한다. 손만 있는 합성 영상 데이터 또한 4개를 생성하고, 이를 통해 추출되는 학습용 합성 이미지는 총 600장이다(그림 2 참고). 각 이미지에 대응되는 레이블 파일 600개를 만들고, 손(클래스 0) 위치에 해당하는 바운딩 박스 정보를 저장한다.

1.3 물체만 인식되는 경우(R2H)

R2H 핸드오버를 위해서는 로봇이 사용자에게 전달해야 하는 물체(음성 명령 등을 통해 설정됨)를 주변 환경에서 탐색하는 과정이 필요하다. 모델은 물체를 정확히 인식해야 하며, 이를 학습하기 위해 물체만 있는 영상을 생성하여 학습 데이터에 포함한다. 본 연구에서는 우선 한 가지의 물체를 설정하여 학습 및 테스트를 진행하며, 비교적 특징이 뚜렷한 빨간색 사과로 설정한다. 데이터 생성을 위한 프롬프트는 “빨간 사과를 윗면, 옆면에서 찍은 영상 만들어줘”이다. 빠른 학습을 위해 총 2개의 합성 영상 데이터만을 생성하며, 이를 통해 학습용 합성 이미지 데이터를 총 300장 추출한다(그림 3 참고). 각 이미지에 대응되는 레이블 파일 300개는 객체 클래스 1과 물체의 위치에 해당하는 바운딩 박스 정보를 저장한다.

손이 물체를 쥐고 있는 데이터, 손만 있는 데이터, 물체만 있는 데이터를 모두 합쳐 모델 학습을 위한 전체 데이터 세트를 구축한다.

2. 구축된 데이터를 통한 모델 학습 방법

앞서 구축한 학습 데이터를 이용해 단일 YOLO 모델을 학습시켜 영상에서 손과 사과를 인식하도록 한다. 본 연구에서 활용하는 YOLO 모델은 YOLO11n으로, 해당 모델을 선택한 이유는 모델의 크기가 작아 연산량이 적으면서도 경량화된 환경에서 높은 정확도, 빠른 추론 속도 등 뛰어난 성능을 보여주기 때문이다.

학습 데이터는 종류별로(사과+손, 손, 사과) 80:20 비율로

학습(training) 및 검증(validation) 데이터로 나눈다. 모델 학습 파라미터의 경우 1) epoch 수는 150, 2) 이미지 입력 크기는 더 높은 해상도를 설정해 세부적인 객체 특징을 학습하기 위해 960×960, 3) 배치 크기(batch size)는 16, 4) 초기 및 최종 학습률(learning rate)은 각각 0.005와 0.05, 5) 학습률 감소 방식은 cosine annealing, 6) optimizer는 일반화 성능 향상과 빠른 수렴, 그래디언트 크기에 덜 민감한 안정적 학습을 위해 AdamW 최적화, 7) 가중치 감소(weight decay)는 5e-4, 8) 검증 성능이 향상되지 않는 상태가 30 epoch 연속으로 지속되면 학습을 중단하기 위해 patience 파라미터를 30으로 설정한다. 또한 두 개의 서로 다른 이미지와 그에 대한 레이블을 선형 결합해 새로운 학습 샘플을 생성하는 mixup과 이미지의 크기를 epoch마다 매번 조금씩 다르게 바꿔주는 multi_scale을 활성화하여 데이터 증강을 통해 더욱 다양한 데이터 학습을 유도한다.

IV. 실험 결과

1. 학습 과정 및 모델 성능 분석

모델 학습 과정에서는 손실값과 주요 성능 지표를 분석한다. 그림 4는 epoch에 따른 학습 및 검증 데이터 세트에서의 box loss, classification loss, distribution focal loss (DFL loss) 변화를 나타낸다. 초기 손실값이 높은 상태에서 점진적으로 감소하여 수렴하는 경향을 보이고, 이를 통해 모델이 점진적으로 최적화되고 있음을 알 수 있다. 또한 정밀도, 재현율, mAP50 및 mAP50-95 지표가 일정 수준 이상에서 안정적으로 수렴하는 것을 확인할 수 있다. 이렇게 최종 학습된 모델의 성능을 요약한 결과를 표 1에서 확인할 수 있다.

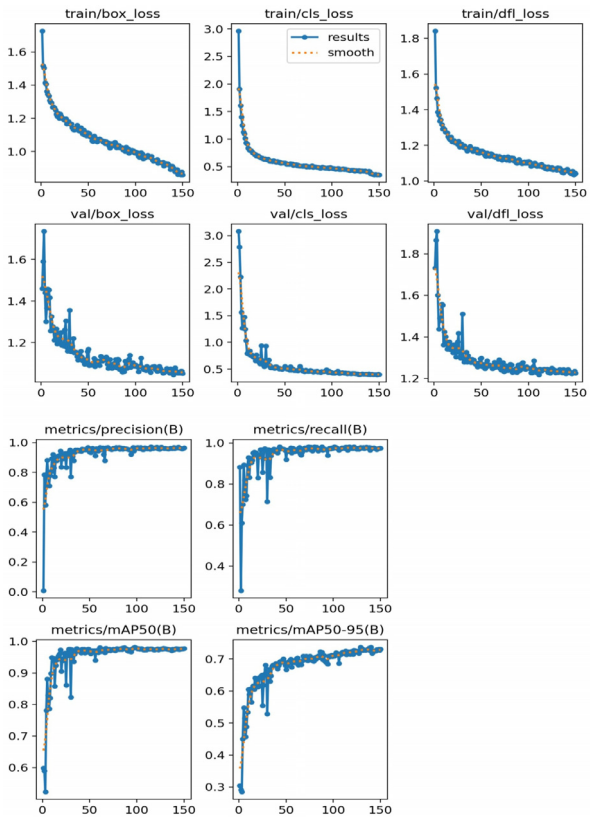


그림 4. Epoch에 따른 손실값과 주요 지표 변화.

Fig 4. Loss values and key metrics according to epochs.

표 1. 학습된 YOLO11n 모델의 성능 요약.

Table 1. Performance summary of the trained YOLO11n model.

	Value
Precision	0.95
Recall	0.99
mAP50	0.97
mAP50-95	0.73

2. 혼동 행렬 분석

혼동 행렬에서는 실제 클래스(apple, hand, background)와 예측된 클래스 간의 혼동 정도를 확인할 수 있다(그림 5 참조). 사과와 손은 높은 정확도로 분류되나, 배경이 사과로 잘못 인식된 사례가 존재하는 것을 확인할 수 있다. 배경이 사과로 잘못 인식된 비율은 약 13%, 사과가 배경으로 잘못 인식된 비율은 약 1.6%이다. 반면 손은 정확하게 분류되는 것을 확인할 수 있다. 배경이 사과로 잘못 인식된 비율이 높은 이유는 학습 데이터를 생성할 때 빛, 그림자, 테두리, 깊이 정보 등이 현실과 다르게 생성되는 경우가 있어, 실제 영상에서 낯선 배경이 나오거나 사과의 색상이 일부 존재할 때 사과로 잘못 인식되는 경우가 있기 때문이다. 이는 생성형 AI 기반 합성 데이터의 현실성 부족이 모델의 일반화 성능에 영향을 줄 수 있음을 보여주는 사례이다.

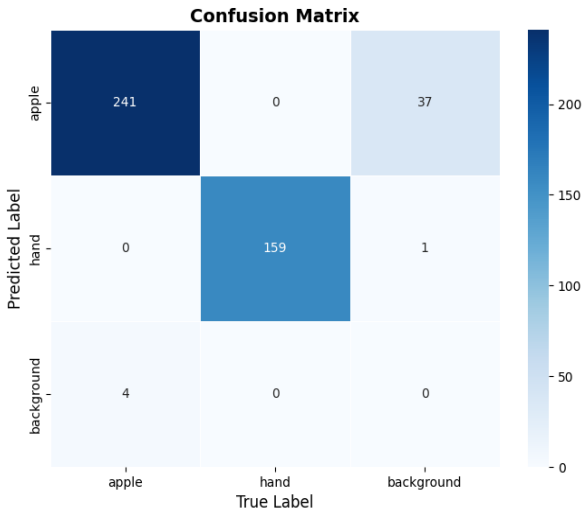


그림 5. 모델의 일반 혼동 행렬.

Fig. 5. General confusion matrix of the model.

3. 정밀도-재현율 곡선 분석

그림 6의 정밀도-재현율 곡선은 임계값(threahold) 0~1에 따른 모델의 사과, 손, 사과 및 손 인식에 대한 정밀도와 재현율 사이의 관계를 그래프로 나타낸 것이며, 범례의 숫자는 각 클래스의 mAP50을 나타낸다. 임계값이 높은 경우 탐지된 객체 수가 적어 재현율이 낮으며 모델 탐지 객체가 정확해 정밀도는 높다(왼쪽 위). 반대로 임계값이 낮은 경우 탐지된 객체 수가 많아 재현율이 높고 모델 탐지 객체가 많은 오류를 포함하고 있어 정밀도가 낮다(오른쪽 아래). 이렇듯 그래프가 우상단에 가까울수록(즉, 곡선 아래의 면적이 1에 가까울수록) 모델이 좋은 성능을 나타낸다고 볼 수 있으며, 앞서 학습시킨 모델이 전반적으로 우수한 성능을 보이는 것을 확인할 수 있다.

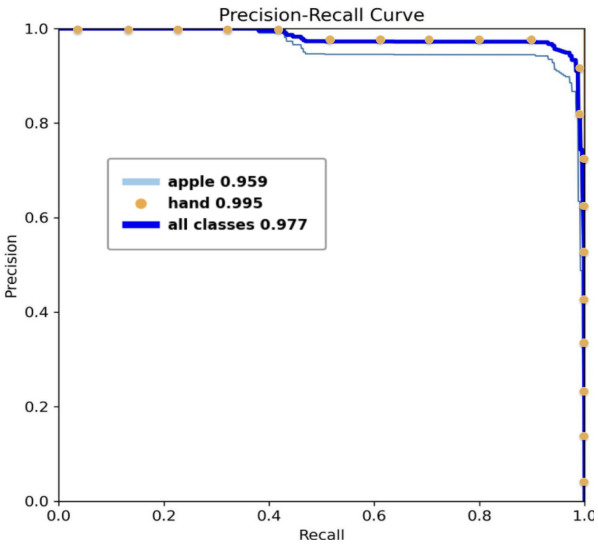


그림 6. 정밀도-재현율 곡선과 각 클래스의 mAP50.

Fig. 6. Precision-recall curve and mAP50 of the classes.

4. 모델의 실제 예측 결과

학습시킨 YOLO 모델에 실제 사과 영상, 실제 사람 손 영상, 실제로 사과를 쥐고 있는 손 영상을 입력으로 넣어 생성형 AI 합성 데이터로 학습한 모델의 실제 데이터에 대한 테스트를 진행한다. 그 결과, 표 2와 그림 7에서 확인할 수 있듯이 전체적인 정밀도와 재현율은 0.94 이상으로 높은 수준을 기록하였고, 특히 apple 클래스는 매우 높은 탐지 성능을 보였다. 반면 hand 클래스의 mAP50-95는 다른 클래스에 비해 낮게 측정되었는데, 이는 손 탐지 시 다른 클래스 탐지에 비해 잘못된 탐지 혹은 객체를 탐지하지 못하는 경우가 많다는 것을 의미한다.

표 2. 실제 영상 데이터에 대한 모델 테스트 결과.

Table 2. Model test results for real video data.

	All	Hand	Apple
Precision	0.94	0.93	0.95
Recall	0.97	0.94	0.99
mAP50	0.97	0.96	0.98
mAP50-95	0.73	0.66	0.80

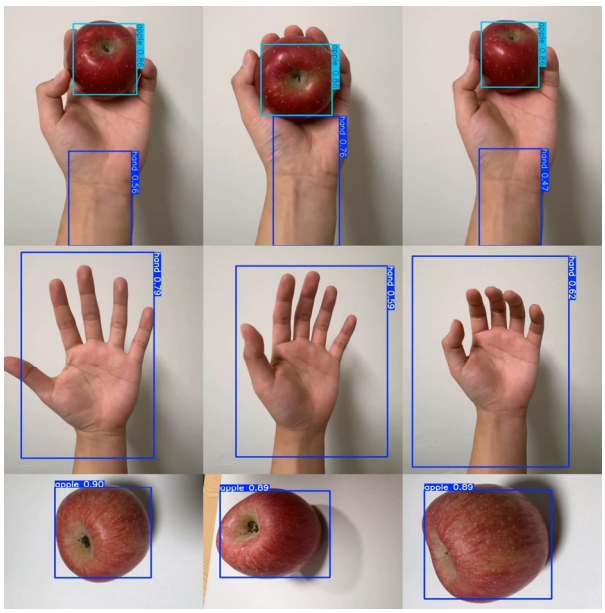


그림 7. 실제 영상 데이터에 대한 모델 테스트 예시.

Fig. 7. Model test examples for real video data.

표 1에서와 마찬가지로 mAP50과 mAP50-95 사이에 차이가 있는 것을 확인할 수 있는데, IoU의 기준이 높아질수록 성능이 떨어지는 경향이 있음을 의미한다. 즉, 안정적인 탐지 성능을 보여주며 객체의 존재 여부는 잘 감지하는 반면 바운딩 박스의 위치가 정확히 맞지 않는 경우가 있다는 것을 알 수 있다. 이는 박스 위치의 오차, 크기 차이가 존재할 수 있음을 의미하며 후처리를 통해 더 정밀한 탐지가 가능하도록 개선할 여지가 있다.

실제 데이터 학습 기반 연구와의 성능 비교를 위해, 비록 핸드오버 연구는 아니지만 동일한 모델(YOLO11n)을 활용하여 실제 항공 이미지 데이터를 학습해 선박을 탐지하는 [11]과의 객체 탐지 성능을 비교한다. [11]의 경우 mAP50 97.6%, mAP50-95 90.1%의 성능을 보였으며, 이는 본 연구 결과 대비 mAP50은 유사한 성능, mAP50-95는 약 17.1% 높은 성능이다. 본 연구 결과는 실제 데이터 학습 결과에 비해 탐지 성능은 다소 떨어지지만, 학습 데이터 수집에서의 시간과 비용이 감소한다는 점에서 이점을 가진다.

V. 결론

본 연구에서는 생성형 AI Sora를 활용하여 다양한 핸드오버 상황을 반영한 합성 영상 데이터를 생성하고, 이를 기반으로 인간-로봇 핸드오버 상황 인식 모델을 학습하는 새로운 방법을 제안하였다. YOLO11n 모델을 사용해 적은 연산량으로 손과 물체의 인식을 효율적으로 수행하였으며, 합성 데이터만을 이용해 학습시킨 모델이 실제 영상에서도 효과적으로 물체와 손을 인식하며 높은 성능을 기록하는 것을 확인하였다. 이를 통해 실제 데이터 수집 과정 없이도 신뢰할 수 있는 학습 데이터를 확보하여 데이터 수집에 필요한 시간과 비용 문제를 해결할 가능성을 검증하였다.

본 연구는 단일 물체(사과)에 대해서만 실험을 수행했다는 점, 추후 더욱 다양한 핸드오버 상황이 발생했을 때 낮은 일반화 성능을 보일 가능성이 있다는 점, 단일 사용자와 유사한 손 모양을 가진 데이터만을 활용하여 모델을 학습시켰다는 점에서 한계점이 존재한다. 따라서 향후 연구에서는 다양한 물체와 손(장갑 착용, 피부색 차이 등), 핸드오버 상황을 포함한 합성 데이터 세트를 구축하고, few-shot learning [12], meta-learning [13] 기법 등을 적용해 생성형 AI 기반 핸드오버 인식 모델이 실제 상호작용 환경에서 높은 신뢰성과 효율성을 보장하도록 발전시킬 계획이다.

REFERENCES

[1] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić, “Object handovers: A review for robotics,” *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1855-1873, Dec. 2021.

doi: https://doi.org/10.1109/tro.2021.3075365

[2] H. Bae, "A study on the legislation for the promotion of the robot industry -With cooperative robots at the center-," *Sogang Journal of Law and Business (in Korean)*, vol. 14, no. 1, pp. 35-55, Apr. 2024.
doi: https://doi.org/10.35505/sjlb.2024.4.14.1.35

[3] H. S. Kim, H. Y. Ra, A. R. Kim, and S. Y. Kim, "A study on construction of transfer learning-based dataset for object detection using CARLA," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 2, pp. 175-182, Feb. 2024.
doi: https://doi.org/10.5302/J.ICROS.2024.23.0204

[4] S.-H. Kim, "Development of virtual reality data collection technology for deep neural network-based landing point recognition methods," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 29, no. 5, pp. 392-397, May 2023.
doi: https://doi.org/10.5302/J.ICROS.2023.23.0016

[5] Z. Wang, Z. Liu, N. Ouporov, and S. Song, "ContactHandover: Contact-guided robot-to-human object handover," *Proc. of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9916-9923, Oct. 2024.
doi: https://doi.org/10.1109/IROS58592.2024.10801777

[6] J. Kwan, C. Tan, and A. Cosgun, "Gesture recognition for initiating human-to-robot handovers," *RO-MAN Workshop in Active Vision and Perception in Human-Robot Collaboration*, Aug. 2020.
doi: https://doi.org/10.48550/arXiv.2007.09945

[7] L. Chen, S.-Y. Lin, Y. Xie, Y.-Y. Lin, W. Fan, and X. Xie, "DGGAN: Depth-image guided generative adversarial networks for disentangling RGB and depth images in 3D hand pose estimation," *Proc. of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 400-408, Mar. 2020.
doi: https://doi.org/10.1109/wacv45572.2020.9093380

[8] S. K. Alhabeeb and A. A. Al-Shargabi, "Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction," *IEEE Access*, vol. 12, pp. 24412-24427, Feb. 2024.
doi: https://doi.org/10.1109/ACCESS.2024.3365043

[9] C. Chang, Z. Liu, X. Lyu, and X. Qi, "What matters in detecting AI-generated videos like Sora?," arXiv preprint arXiv:2406.19568, Jun. 2024.
doi: https://doi.org/10.48550/arXiv.2406.19568

[10] Z. Zhou, D. Gu, Y. Shi, H. Zhou, K. Chen, H. Qu, and H. Ren, "Improving object detecting by structuring and training YOLO model," *Proc. of the 5th International Conference on Computer Information and Big Data Applications (CIBDA)*, pp. 659-664, Jul. 2024
doi: https://doi.org/10.1145/3671151.3671268

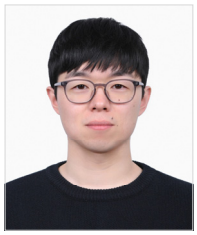
[11] J. Huang, K. Wang, Y. Hou, and J. Wang, "LW-YOLO11: A lightweight arbitrary-oriented ship detection method based on improved YOLO11," *Sensors*, vol. 25, no. 1, Dec. 2024.
doi: https://doi.org/10.3390/s25010065

[12] Y. Wu, S. Chanda, M. Hosseinzadeh, Z. Liu, and Y. Wang, "Few-shot learning of compact models via task-specific meta distillation," *Proc. of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6254-6263, Jan. 2023.
doi: https://doi.org/10.1109/WACV56688.2023.00620

[13] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artificial Intelligence Review*, vol. 18, no. 2, pp. 77-95, Jun. 2002.
doi: https://doi.org/10.1023/A:1019956318069

박 무 열

2025년 세종대 지능기전공학부 학사. 2025년~현재 세종대 AI로봇학과 석사과정. 관심분야는 인간-로봇 상호작용, 인공지능, 로봇 제어.



박 규 민

2016년 서울대 전기정보공학부 학사. 2022년 서울대 기계항공공학부 박사. 2022년~2023년 KIST 지능로봇연구단 박사후연구원. 2023년~현재 세종대 AI로봇학과 조교수. 관심분야는 인간-로봇 상호작용, 로보틱스를 위한 인공지능 응용, 로봇 모델링, 동작계획 및 제어.