논문 2024-19-10

작물 수확 자동화를 위한 시각 언어 모델 기반의 환경적응형 과수 검출 기술

(Domain Adaptive Fruit Detection Method based on a Vision-Language Model for Harvest Automation)

남 창 우, 송 지 민, 진 용 식, 이 상 준*
(Changwoo Nam, Jimin Song, Yongsik Jin, Sang Jun Lee)

Abstract : Recently, mobile manipulators have been utilized in agriculture industry for weed removal and harvest automation. This paper proposes a domain adaptive fruit detection method for harvest automation, by utilizing OWL-ViT model which is an open-vocabulary object detection model. The vision-language model can detect objects based on text prompt, and therefore, it can be extended to detect objects of undefined categories. In the development of deep learning models for real-world problems, constructing a large-scale labeled dataset is a time-consuming task and heavily relies on human effort. To reduce the labor-intensive workload, we utilized a large-scale public dataset as a source domain data and employed a domain adaptation method. Adversarial learning was conducted between a domain discriminator and feature extractor to reduce the gap between the distribution of feature vectors from the source domain and our target domain data. We collected a target domain dataset in a real-like environment and conducted experiments to demonstrate the effectiveness of the proposed method. In experiments, the domain adaptation method improved the AP50 metric from 38.88% to 78.59% for detecting objects within the range of 2m, and we achieved 81.7% of manipulation success rate.

Keywords : Fruit detection, Deep learning, Vision-language model, Domain adaptation, Adversarial learning

Ⅰ. 서 론

인구고령화로 인하여 다양한 산업문제가 발생하고 있으며, 특히 농업분야에서는 인력난으로 인한 생산량 감소가 주요한 이슈 중 하나이다. 이는 농촌 지역에서의 소득감소와 성장률 정체 등을 야기하며 국내 농업의 지속가능성에 대한 우려를 증폭시키고 있다. 이러한 어려움에 대응하고 지속 가능한 농업을 구축하기 위한 다양한 대안이 제안되고 있으며, 최근에는 스마트 팜 기술이 큰 주목을 받고 있다. 특히, 스마트 팜에서 인공지능 기술을 활용한 혁신적인 시도가 이뤄지고 있으며, 이러한 인공지능 기술은 농업 분야에서 다양한 문제들을 해결하고 농작업을 효율적으로 최적화할 수 있는 가능성을 제시하고 있다. 예를 들어, 토양 센서를 통해 획득한 질소량이나 수분량 수치데이터를 기반으로 인공지능을 이용하여 농업 토양 적합성에 대한 평가가 가능하다 [1]. 사물인터넷과 저비용 및 소형 센서를 활용하여 센서 네트워크를 구축하고 인공지능과 결합하여 농업 토양 적합도를 평가한다. 또한, 로봇이나 드론에 장착된 카메라를 이용하여 작물의 잎이나 모양을 찍은 이미지를 인식하고 잡초, 식물 질병 등을 감지하고 대응함으로써 농작물의 수확량과 품질 향상이 가능하다 [2, 3].

하지만 기존의 센서 데이터 분석 및 영상인식 방법들은 농작업 보조기술에 해당하여 여전히 사람의 노동력을 필요로 하며 작물 수확을 자동화하는 데에는 한계가 있다. 현재 대부분 농가에서의 작물 수확 방식은 인간 노동자에게 의존하며 이는 노동 집약적이고 시간 소모적이다. 한편, 제조 분야에서는 공정자동화를 위한 무인 로봇 기술이 개발되는 추세이다. 농업 분야에서도 모바일 매니퓰레이터 기술을 적용하여 작물 수확 과정을 자동화하고 효율성을 개선하는 것이 가능하며, 관련된 연구가 제안된 바 있다 [4]. 이러한 작물 수확 자동화 기술은 카메라를 통해 이미지를 획득하고 이미지 내에서 작물을 식별하며, 매니퓰레이터를 이용하여 관심 작물의 수확이 가능하다.

딥러닝 기술은 이를 위한 요소 기술로서 컴퓨터가 마치 사람처럼 스스로 학습할 수 있도록 하는 인공 신경망 기반 기계 학습 기술이다. 이미지 인식 및 처리 분야에서 사용되는 딥러닝 모델에는 대표적으로 합성곱 신경망과 비전 트랜스포머를 들 수 있다. 합성곱 신경망은 인간의 시신경 구조

*Corresponding Author (sj.lee@jbnu.ac.kr)
Received: Dec. 25, 2023, Revised: Jan. 23, 2024, Accepted: Feb. 5, 2024.
C. Nam: Jeonbuk National University (M.S. Student)
J. Song: Jeonbuk National University (M.S. Student)
Y. Jin: ETRI (Researcher)
S. J. Lee: Jeonbuk National University (Assist. Prof.)
※ This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [23ZD1130, Regional Industry ICT Convergence Technology Advancement and Support Project in Daegu-GyeongBuk (Robot)]. 이 연구는 2023년도 산업통상자원부 및 산업기술평가관리원 (KEIT) 연구비 지원에 의한 연구임 (20023305).

 2024 Apr. 19(2) 73-81
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2024.19.2.73

를 모방하여 인간이 시각 정보를 처리하는 것에 영감을 받아 제안된 구조로 2차원 이미지 데이터 처리에 적합한 구조이다. 과거부터 현재까지 활발하게 연구가 진행되고, 다양한 분야에서 적용되고 있으며 특히 이미지 처리 분야에서 우수한 성능을 보인다. 비전 트랜스포머는 자연어 처리 분야에서 사용되던 트랜스포머 구조를 이미지 처리 분야에 응용함으로써 합성곱 신경망을 이용한 네트워크보다 좋은 성능을 보여주고 있다 [5].

컴퓨터 비전 분야에서 딥러닝 기술의 발전이 이루어지며 이미 그 성능은 인간의 시각 능력을 능가했다는 평가를 받는다. 하지만 여전히 대부분의 딥러닝 모델은 실제 작업에 적용되기 위해 레이블 데이터가 포함된 학습 데이터를 필요로 하며, 별도의 학습 데이터를 구축하는 작업은 많은 비용을 요구한다. 특히 객체 경계 상자, 분할 마스크와 같은 레이블 데이터를 제작하는 작업은 여전히 인간의 노동을 필요로 하고 있는 실정이다. 따라서, 적용하고자 하는 도메인에서 추가적인 레이블 데이터의 획득 없이 우수한 성능을 낼 수 있는 딥러닝 모델에 대한 연구가 요구된다.

본 논문은 학술대회 발표논문 [6]을 확장한 논문으로서, 무인 이동 로봇의 수확 자동화를 위한 환경적응형 과수 검출 및 매니퓰레이션 기술을 제안한다. 기존 연구 [4]와 비교하여, 객체 검출을 위하여 사전훈련된 시각 언어 기반 딥러닝 모델인 OWL-ViT를 활용하여 실제로 딥러닝 모델이 적용될 현장에서 검출 대상 객체가 추가될 때마다 모델의 재훈련이 요구되는 단점을 보완하였다. 또한 훈련된 딥러닝 모델을 그대로 사용하는 기존 연구와 다르게, 실제 환경에서 모델의 검출 정확도를 향상할 수 있도록 적대적 학습 기반의 도메인 적응 과정을 구성하였다. 도메인 적응 과정을 통해 레이블 데이터 없이, 과수의 이미지 데이터만을 이용하여 실제 환경에서 딥러닝 모델이 환경에 적응할 수 있도록 하였으며, 실제 데이터 셋을 수집하여 이를 검증하였다. 또한 도메인 적응이 완료된 모델과 매니퓰레이션 기술을 결합하여 매니퓰레이터가 검출된 과수에 접근하는 시스템을 구현하고 그 정확도를 측정하였다.

Ⅱ. 관련 연구

1. 객체 검출 모델

객체 검출 분야에서는 다양한 구조 및 훈련 기법을 바탕으로 딥러닝 기술들이 제안되는 추세이다. YOLO는 합성곱 신경망 기반의 객체 검출 알고리즘으로, 기존의 객체 후보군 검출과 분류로 이루어지는 2단계 검출 기술과 달리 단일 단계의 검출기 모델을 구성하여 높은 성능과 빠른 처리 속도를 달성하였다 [7]. 이미지 전체의 시각적 특징을 공유하여 객체를 검출함으로써 기존 알고리즘 대비 배경 이미지를 객체로 탐지하는 오검출을 최소화하고, 객체의 일반화된 표현을 학습하여 여러 도메인에서 좋은 성능을 보이는 장점이 있다. DETR은 합성곱 신경망으로부터 이미지 특징을 추출하고, 이미지 특징을 트랜스포머 구조에 입력함으로써 객체의 경계 상자와 클래스 분류를 예측한다 [8]. 객체 검출 분야에서 최초로 트랜스포머 구조를 적용하였으며. 기존 알고리즘들과 달리 이분 매칭 (bipartite matching)을 활용하여 복잡한 후처리 과정을 제거함으로써 계산 비용을 줄이고 우수한 성능을 달성하였다.

2. 시각-언어 모델

최근 언어와 시각 정보 사이의 관계를 학습하고 이해하며 이를 바탕으로 문제를 해결하는 시각-언어 (vision-language) 모델이 제안되고 있다. 이전의 딥러닝 모델들은 사전에 정의된 종류의 객체들을 검출하도록 훈련되어왔고, 이러한 지도학습 방식은 다른 종류의 객체를 검출하기 위해 추가적인 레이블 데이터가 필요하다는 단점이 있다. 이에 대한 대안으로 시각-언어 모델인 CLIP은 텍스트 인코더와 이미지 인코더를 이용하여 텍스트와 이미지에서 임베딩이라고 불리는 특징벡터를 추출하고 두 임베딩 사이의 유사도를 학습에 이용한다 [9]. CLIP은 이미지, 그리고 해당 이미지와 연관된 텍스트로 구성된 상관성 쌍 (positive pair)의 유사도는 최대화하고 비상관성 쌍 (negative pair)의 유사도는 최소화하는 대조 학습 (contrastive learning)을 통해 학습된다. 4억 장의 이미지-텍스트 쌍을 바탕으로 학습하여 zero-shot 이미지 분류에서 기존 알고리즘 대비 우수한 성능을 달성하였다. CLIP을 기반 네트워크로 사용하여 텍스트를 기반으로 객체 검출을 수행하는 개방형 어휘 탐지 모델인 OWL-ViT가 제안된 바 있다 [10]. OWL-ViT는 기존 CLIP 모델의 이미지, 텍스트 특징 추출기에 경량화 된 객체 분류 및 지역화 네트워크를 결합한 구조로 구성되어 있다. OWL-ViT는 텍스트를 기반으로 이미지에서 사전에 정의되지 않은 객체를 검출하는 제로샷 검출이 가능하다. 따라서, 학습 데이터에 존재하는 종류의 객체들만 탐지할 수 있는 일반적인 객체 검출 모델과 비교했을 때, 사전에 정의되지 않은 종류의 객체도 검출 가능하다는 특징이 있다.

3. 공개 데이터 셋

과수 재배 분야에서도 딥러닝 기술을 접목하기 위한 데이터 구축 및 알고리즘 개발이 진행되고 있다. 사과 검출 및 분할 학습을 위해 객체 경계 상자 및 분할 마스크를 비롯한 레이블 데이터와 사과 이미지를 포함한 공개 데이터 셋이 제안된 바 있다 [11]. Galaxy S4 기기의 카메라를 이용해 촬영된 1280×720 사이즈의 사과 이미지 1,001장과 41,325개의 주석으로 이루어진 라벨링 데이터를 제공한다. 국내 데이터 셋 제공 기관인 AI Hub (https://www.aihub.or.kr/)에서는 사과 품종별 이미지와 객체 검출, 분할을 위한 라벨링 데이터를 포함한 학습용 공개 데이터 셋을 제공한다 [12]. 당도 측정 데이터, 촬영 각도, 토양 센서 데이터 등 다양한 데이터를 포함하고 있으며, 총 535,691장의 이미지가 포함되어 있다. 영상에서 과수 검출을 위한 딥러닝 기법으로서 Chu et al.에서는 억제 마스크 R-CNN 기반 딥러닝 기반 사

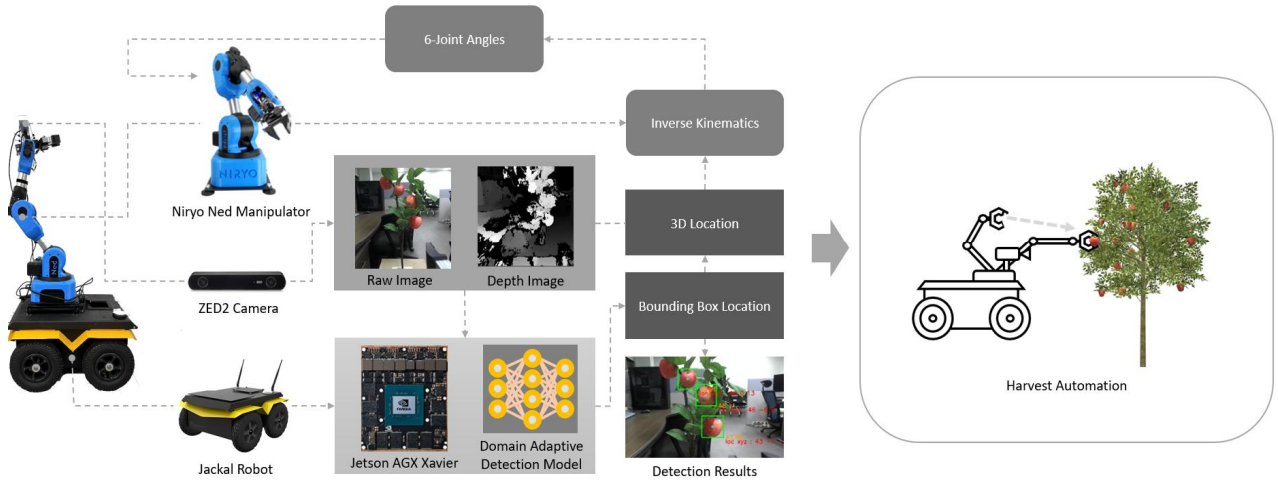

그림 1. 작물 수확 자동화를 위한 시스템 구성
Fig. 1. Overview of the manipulation system for harvest automation

과 감지 프레임워크를 제안한다 [13]. 기존 Mask R-CNN 구조에 추가된 특징 억제 네트워크는 이미지 내의 사과가 아닌 배경 영역을 필터링하여 사과 검출 정확도 향상을 돕는다. 또한 나뭇가지나 잎과 같은 장애물에 의해 가려진 과수를 검출하거나 다양한 조명 조건 내에서 과수를 검출하는 것이 가능함을 보였다.

4. 도메인 적응

일반적으로 딥러닝 모델들은 학습 데이터와 동일한 특징 공간과 분포를 가진 테스트 데이터에서 우수한 성능을 보인다. 이는 학습된 딥러닝 모델을 실제 작업에 적용시켰을 때 성능 하락의 경우로 이어질 수 있다. 실제 데이터와 학습 데이터 사이의 분포 차이로 인해 딥러닝 모델의 성능의 실제로 적용하고자 하는 도메인에서 하락하는 문제가 발생한다. 이를 해결하기 위해 기존 도메인 정보를 바탕으로 새로운 도메인에 적응하여 모델의 성능을 유지, 향상할 수 있는 도메인 적응형 딥러닝 기술에 대한 연구가 진행되고 있다. Ganin et al.에서는 학습 데이터 도메인과 실제 데이터 도메인 사이의 특징 맵핑을 통해 분류기가 실제 도메인에서도 효과적으로 동작하도록 하는 도메인 적응 방법인 DANN을 제안하였다 [14]. 경사 반전 층 (gradient reversal layer)을 도입하여 학습 과정의 오차 역전파 과정에서 미분계수에 음의 상수를 곱하여 학습함으로써 모델이 클래스 분류에 쓰이는 특징을 효과적으로 추출할 수 있도록 하였다. Tzeng et al.에서는 GAN의 손실 함수를 응용한 적대적 손실 함수를 도입하여 도메인 적응을 진행하였다 [15]. 소스 및 타겟 도메인을 구별하는 분류기와 타겟 도메인에서 추출한 특징 분포가 소스 도메인의 분포와 유사하도록 특징벡터를 매핑하는 CNN을 동시에 학습하여, 소스 도메인에서 먼저 훈련된 딥러닝 모델이 타겟 도메인에도 적응할 수 있도록 보조하는 적대적 학습 기법이다.

Ⅲ. 제안 방법

1. 데이터 셋 구성

도메인 적응을 위한 소스 도메인 데이터 셋으로 AI Hub의 공개 데이터 셋인 전북 장수 사과 당도 품질 데이터를 사용하였다. 총 571.63GB 용량의 학습 데이터로 구성되어 있으며, 객체 경계 상자 및 분할 마스크를 비롯한 레이블 데이터와 사과 품종별 이미지를 포함하고 있다. 해당 데이터 셋은 SM-G965N, Canon EOS 600D 등 다양한 촬영 장비로 수집되었으며, 해당 데이터 셋에서 품종별 분포를 확인하였다. 가장 많은 분포를 차지하는 후지 품종의 실제 과수원 환경에서 촬영된 데이터를 소스 도메인 데이터 셋으로 활용하였다.

실제 도메인을 가정한 환경에서 도메인 적응과 객체 검출 모델의 성능 검증을 위해 타겟 도메인 데이터 셋을 직접 수집하였다. ZED2 스테레오 카메라가 탑재된 자율 주행 로봇 Jackal을 사용하였으며, 카메라는 지상으로부터 1m 높이가 되도록 로봇에 설치되었다. 엣지 컴퓨팅을 위한 Jetson AGX Xavier가 탑재되어 로봇 제어, 이미지 처리 및 수집 과정을 수행하였다. Jackal 로봇을 이용하여 현실적인 과수 수확 로봇 환경에서 데이터를 수집할 수 있도록 하였으며, 사과 가지 모형과 로봇 간 0.3m부터 3.0m 거리 구간에서 0.3m 간격으로 성능 검증 데이터를 확보하였다. 매니퓰레이터의 구동 범위를 고려하고 Jackal 로봇이 과수 근처에 충분히 도달한 상황을 가정하여, 확보한 데이터 셋에서 0.3m와 0.9m 구간의 사이에 해당하는 일부 데이터를 도메인 적응을 위한 타겟 도메인 데이터 셋으로 활용하였다.

2. 작물 수확 자동화 시스템 구성

과수 검출 알고리즘과 매니퓰레이터를 이용한 작물 수확 자동화 시스템의 개요는 그림 1과 같다. 모바일 로봇에 장착된 ZED2 스테레오 카메라를 통해 RGB 이미지와 뎁스 정보로 구성된 이미지를 획득하고 임베디드 컴퓨터인 Jetson

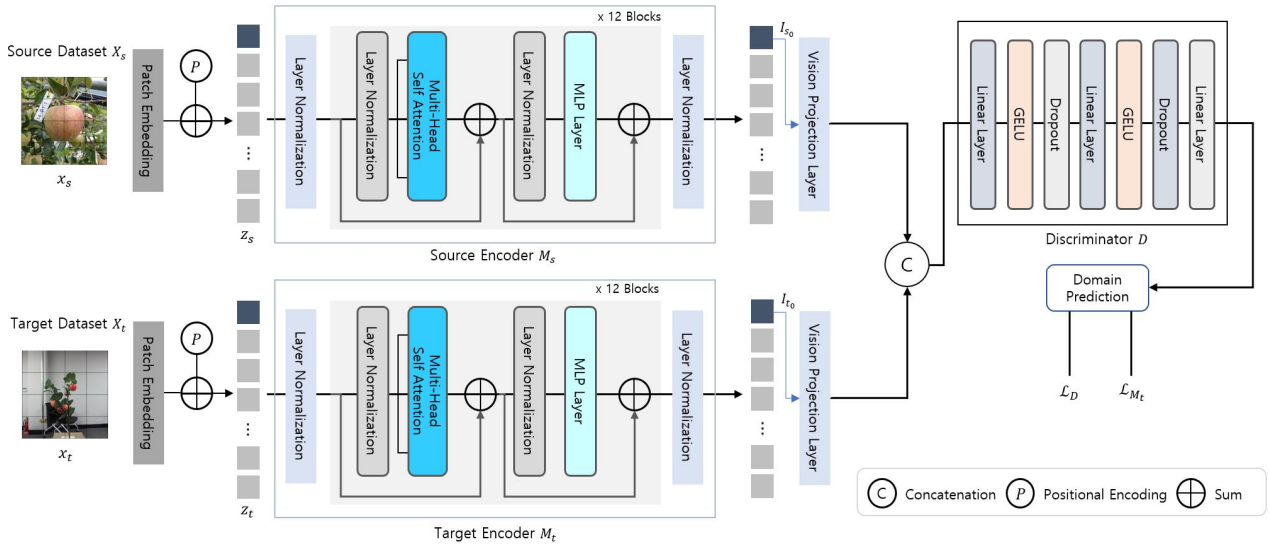

그림 2. 적대적 손실 기반의 도메인 적응과정

Fig. 2. Domain adaptation process based on an adversarial loss function

AGX Xavier에서 이미지 처리가 수행된다. 딥러닝 모델을 통해 과수를 검출하고 검출 결과와 뎁스 정보를 종합하여 해당 과수의 3차원 위치를 추정한다. 카메라 좌표계에서 추정된 과수의 위치 정보를 매니퓰레이터 기준 좌표계로 변환하고, 역기구학 (inverse kinematics)을 기반으로 계산한 관절 각도를 활용하여 로봇팔이 과수에 접근 할 수 있도록 제어한다.

3. 적대적 도메인 적응을 이용한 과수 검출 알고리즘

본 논문에서 제안하는 환경적응형 기술은 객체 검출을 위한 시각-언어 기반 모델인 OWL-ViT와 모델의 검출 성능을 향상하기 위한 적대적 학습과정으로 구성되며, 그림 2는 이와 같은 딥러닝 모델의 도메인 적응 과정을 보여준다. 도메인 적응 과정은 합성곱 신경망 기반 분류 모델의 정확도를 향상하는 기존 연구의 구조를 기반으로 하며 [15], 이를 OWL-ViT의 트랜스포머 구조에 적용할 수 있도록 변형하였다. OWL-ViT의 사전 학습된 비전 트랜스포머 인코더를 소스 인코더와 타겟 인코더로 적용하였다. 비교적 선명하고 객체의 구분이 명확한 공개 데이터 셋을 소스 도메인으로, 그렇지 않은 실제 환경에서의 데이터 셋을 타겟 도메인으로 활용하였다. 이를 통해 사전 훈련된 소스 인코더가 소스 도메인의 데이터에서 과수의 특징을 추출하는 것처럼, 타겟 인코더가 비교적 특징을 추출하기 어려운 타겟 도메인의 데이터에서도 과수의 특징을 잘 추출할 수 있도록 하였다. 각각의 도메인과 연결된 인코더는 이미지를 입력받아 이미지 정보를 담고 있는 임베딩을 출력하는 특징 추출기의 역할을 수행한다. 이미지 임베딩의 첫 번째 토큰인 클래스 토큰은 이미지의 전체적인 특징을 함축한 정보를 담고 있으며, 판별자 (discriminator)의 입력으로 사용된다. 도메인 적응 과정에서, 판별자는 이미지 임베딩의 출처 도메인을 잘 구분할 수 있도록 훈련되며, 타겟 인코더는 판별자의 도메인 구분을 방해하는 방향으로 학습한다. 이는 타겟 인코더가 소스 도메인의 특징을 모사하도록 학습되는 것을 의미하며, 이를 통해 타겟 도메인의 이미지에서도 과수 검출로 이어질 수 있는 충분한 과수의 특징을 추출할 수 있게 된다.

그림 2에서 X_s, X_t는 소스 도메인 데이터 셋과 타겟 도메인 데이터 셋을, x_s, x_t는 해당 데이터 셋의 이미지 샘플을 의미한다. 768×768×3의 크기를 가지는 3차원 RGB 이미지는 패치 임베딩 층에서 패치 크기 값인 32를 커널 크기와 stride로 가지며 출력 차원의 수가 768인 2차원 컨볼루션 레이어를 통과하여 24×24×768의 크기를 가지는 특징맵으로 변환된다. 특징맵은 트랜스포머의 입력 형식에 적합하도록 576×768의 크기를 가지는 1차원 패치들의 시퀀스로 평탄화되며, 평탄화된 시퀀스에 이미지 전체 표현을 나타낼 수 있는 학습 가능한 토큰을 추가하고, 이미지 패치의 위치 정보 소실을 방지하기 위한 포지션 임베딩과 합하여 577×768의 크기를 가지는 입력 임베딩인 z_s와 z_t가 구성된다.

각 도메인의 인코더는 전, 후 계층 정규화 층과 12개의 비전 트랜스포머 인코더 블록으로 구성되며, 각 블록은 다시 계층 정규화, 멀티헤드 셀프 어텐션과 MLP 층으로 구성된다. 입력 층 정규화가 수행된 다음, 각 블록을 지나며 셀프 어텐션 연산이 이루어지고, 이를 통해 이미지 내에 존재하는 객체에 대한 정보를 담고 있는 577×768 크기의 이미지 임베딩으로 변환된다. 12개의 인코더 블록을 지나며 변환된 이미지 임베딩은 다시 층 정규화가 수행되며, 각 도메인의 encoder가 입력 임베딩 z_s, z_t를 입력받아 이미지 임베딩 I_s, I_t을 맵핑하는 함수를 다음과 같이 정의한다.

$$M_s(z_s) = I_s,\ M_t(z_t) = I_t. \tag{1}$$

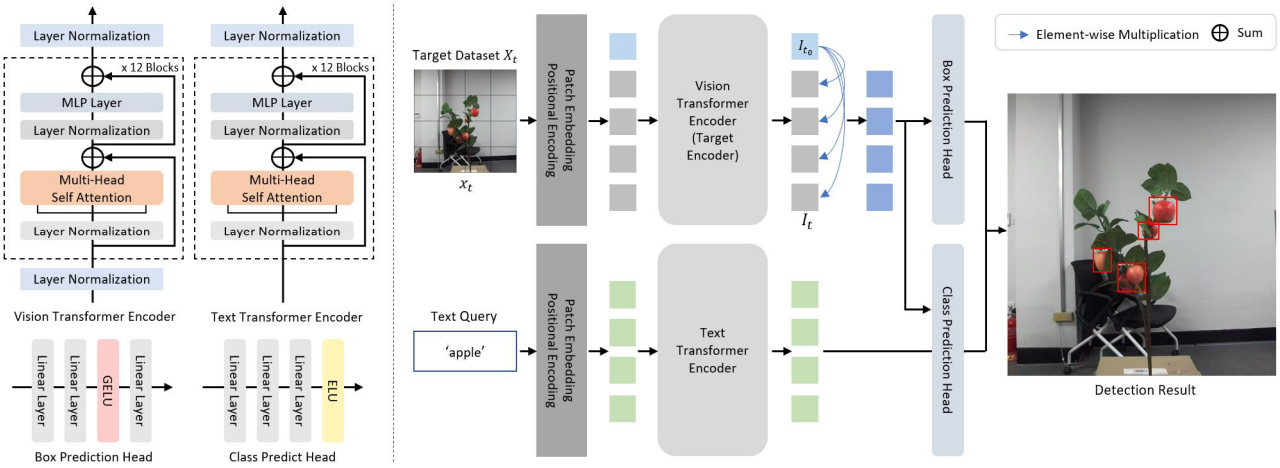

그림 3. OWL-ViT의 개방형 어휘 탐지
Fig. 3. Open-vocabulary object detection method based on OWL-ViT model

그림 2에서 I_{s_0}, I_{t_0}는 각 인코더의 출력인 I_s, I_t의 첫 번째 토큰으로 1×768의 크기를 가지며, 인코더에 입력되기 전 단계에서 추가된 학습 가능한 클래스 토큰을 나타낸다. 판별자는 이미지 전체의 특징 표현을 담고 있는 클래스 토큰인 I_{s_0}, I_{t_0}를 입력으로 받아, 각 토큰이 어느 도메인의 특징 정보인지 추론하는 이진 분류를 수행한다. 판별자의 도메인 예측 값과 소스 도메인, 타겟 도메인의 도메인 라벨 1, 0을 비교하여 판별자가 도메인을 잘 예측할 수 있도록 학습된다. Z_s, Z_t를 각각 소스 도메인 데이터 셋과 타겟 도메인 데이터 셋의 입력 임베딩 집합으로 정의하며, 판별자의 학습을 위한 손실 L_D는 다음의 수학식으로 정의된다.

$$L_D(Z_s, Z_t, M_s, M_t) = -E_{z_s \sim Z_s}[\log D(M_s(z_s))] - E_{z_t \sim Z_t}[\log(1 - D(M_t(z_t)))]. \quad (2)$$

타겟 인코더는 입력 임베딩 z_t가 맵핑 함수를 통과한 $M_t(z_t)$인 I_t의 클래스 토큰 I_{t_0}을 판별자가 입력받았을 때 소스 도메인의 클래스 토큰으로 예측하도록 학습이 진행된다. 타겟 도메인 이미지 임베딩의 클래스 토큰인 I_{t_0}에는 타겟 도메인의 도메인 라벨인 0이 아닌 소스 도메인의 도메인 라벨인 1을 부여하고 판별자의 예측값 $D(I_{t_0})$와 비교한다. 타겟 인코더의 학습을 위한 손실 L_{M_t}는 다음의 수학식으로 정의된다.

$$L_{M_t}(Z_t, D) = -E_{z_t \sim Z_t}[\log D(M_t(z_t))]. \quad (3)$$

따라서, 타겟 도메인에 적응된 타겟 인코더는 두 손실함수 L_D와 L_{M_t}의 최소화를 통하여 얻을 수 있다.

적대적 도메인 적응과정은 판별자의 학습과 타겟 인코더의 학습을 일정 반복마다 교차로 진행하였다. 학습 시작 시, 소스 인코더의 가중치를 고정하여 오차 역전파을 통한 가중치 업데이트가 되지 않도록 설정하였다. 판별자의 학습과정에서, 타겟 인코더의 가중치를 고정하고 판별자의 가중치는 학습되도록 하였다. 타겟 인코더의 학습 과정에서는 반대로 타겟 인코더의 가중치를 학습되도록 하고, 판별자의 가중치를 고정하였다. 학습이 교차되는 반복 횟수는 5로 설정하여 5개의 이미지 샘플마다 타겟 인코더와 판별자가 교차로 학습되도록 하였다. 도메인 적응이 완료된 뒤 타겟 인코더의 가중치를 OWL-ViT의 비전 트랜스포머 인코더로 전이하여 성능 평가를 진행하였다. 그림 3은 도메인 적응 후 OWL-ViT의 객체 검출 과정을 나타내었다. 타겟 인코더는 실제 데이터 셋 X_t의 이미지를 입력받아 이미지 임베딩 I_t을 출력한다. 이미지 임베딩에서 이미지 전체 정보를 담고 있는 클래스 토큰 I_{t_0}이 분리되고, I_{t_0}는 I_{t_0}를 제외한 모든 토큰들과 요소 별 곱셈이 가능하도록 변형된다. I_{t_0}가 곱해진 토큰들을 이미지 임베딩으로 재정의하며 텍스트 임베딩과 함께 클래스 예측 헤드 (class prediction head)에서 객체의 클래스 분류와 박스 영역 헤드 (box prediction head)에서 객체의 위치를 특정하며 최종적으로 객체 검출을 수행한다.

4. 검출 정보를 이용한 과수의 위치 추정

그림 1의 작물 수확 자동화 시스템에서, OWL-ViT는 카메라로부터 획득된 이미지를 입력받아 이미지 내 과수를 식별하며, 식별된 과수의 클래스 정보와 경계 상자 정보를 반환한다. 반환된 경계 상자 정보는 직사각형 경계 상자의 좌상점 픽셀 좌표인 w_l, h_l과 우하점 픽셀 좌표 w_r, h_r로 이루어진다. 이를 통해 경계 상자 중심의 픽셀 좌표 c_w, c_h를 구할 수 있으며 다음의 수식과 같다.

$$c_w = \frac{w_l + w_r}{2}, \quad c_h = \frac{h_l + h_r}{2}. \quad (4)$$

경계 상자 중심의 픽셀 좌표를 검출된 과수의 중심으로 가정한다. ZED2 카메라의 자체 라이브러리를 활용하여 뎁스맵을 추정하고, 추정된 뎁스맵과 이미지 내 과수의 중심 좌표를 이용하여 카메라와 과수 간 실제 거리와 카메라 기준 과수의 실제 위치 좌표를 획득한다.

표 1. 실제 데이터 셋에서의 도메인 적응 전·후 비교실험 결과
Table 1. Performance comparison results on real dataset before and after domain adaptation

	Before domain adaptation				After domain adaptation			
Distance	Precision	Recall	AP50	AP50-95	Precision	Recall	AP50	AP50-95
0.3m	0.8630	0.577	0.7131	0.5393	0.8275	0.9718	0.9621	0.5733
0.6m	0.8110	0.4308	0.5772	0.3570	0.8788	0.8923	0.8805	0.4046
0.9m	0.6686	0.2097	0.3530	0.1837	0.8207	0.9516	0.9376	0.3900
1.2m	0.8704	0.2576	0.4735	0.2620	0.8346	0.7647	0.8231	0.2535
1.5m	0.6736	0.0469	0.1247	0.0872	0.5821	0.6311	0.6318	0.1742
1.8m	0.8290	0.0323	0.0916	0.0502	0.5272	0.6129	0.4808	0.1159
Average	0.7859	0.2590	0.3888	0.2465	0.7451	0.8040	0.7859	0.3185

표 2. 원거리 데이터에서의 비교실험 결과
Table 2. Performance comparison results on long-distance real data before and after domain adaptation

	Before domain adaptation		After domain adaptation	
Distance	AP50	AP50-95	AP50	AP50-95
2.1m	0	0	0.5822	0.1229
2.4m	0	0	0.4011	0.0957
2.7m	0	0	0.3059	0.0659
3.0m	0	0	0.0890	0.0135
Average	0	0	0.3446	0.0745

5. 과수의 위치 정보를 활용한 매니퓰레이터 조정

ZED2 카메라 기준 좌표계에서의 실제 위치 좌표를 매니퓰레이터 기준 좌표계로 변환하기 위해, ZED2 카메라와 매니퓰레이터 로봇팔에 장착되어 있는 카메라 사이의 외부파라미터를 계산하였다. 카메라 캘리브레이션을 통해 ZED2 카메라에서의 3차원 과수 위치를 매니퓰레이터 기준 좌표계에서 나타낼 수 있도록 하였다. 매니퓰레이터의 소프트웨어를 사용하여 과수의 실제 위치로부터 역기구학을 통해 로봇팔의 말단 부분이 목표 지점까지 이동하기 위한 매니퓰레이터의 6개 관절의 각도를 획득한다. 획득된 6개 관절의 각도를 통해 매니퓰레이터의 로봇팔이 과수에 접근할 수 있도록 제어하였다.

Ⅳ. 실험 결과

딥러닝 모델의 최적화 과정에서는 두 가지 최적화 기법을 적용하였다. 판별자의 학습 과정에서는 Adam [16]을, 타겟 인코더의 학습에는 ASGD 기법 [17]을 사용하고, 100 epoch까지 학습을 진행하고 epoch마다 실제 데이터 셋 전체에 대한 검출 성능을 확인하여 가장 좋은 모델과 도메인 적응 전 모델의 성능을 비교하였다.

객체 검출 모델이 추론한 경계 상자의 정확도는 실제 경계 상자와의 IoU (Intersection of Union)를 통해 판단이 가능하다. IoU는 모델이 추론한 경계 상자 B_{gt}와 라벨 데이터의 경계 상자 B_p간 중첩되는 부분의 면적을 계산하고 중첩된 면적을 합집합의 면적으로 나눠준다. IoU는 다음과 같은 수식으로 정의된다.

$$IoU = \frac{B_{gt} \cap B_p}{B_{gt} \cup B_p}. \quad (5)$$

본 논문에서는 IoU의 계산 값이 0.5 이상이면 제대로 검출 되었다고 판단하고 (true positive), 그 반대의 경우 잘못된 검출 (false positive)이라 판단하였다. Precision은 정밀도를 의미하며 딥러닝 모델이 positive로 검출한 결과의 정확도를 나타낸다. Recall은 재현율을 의미하며, 딥러닝 모델이 실제 positive의 경우를 얼마나 잘 검출하는지 그 정도를 나타낸다. Precision과 recall은 trade-off 관계에 있어 서로 반비례 관계를 가진다. 따라서 검출 모델의 성능을 평가하는 방법으로 precision과 recall을 모두 고려한 Precision-Recall Curve를 이용한다. Confidence threshold 값을 변화시켜 그에 따른 precision 및 recall 값의 변화를 Precision-Recall Curve로 나타내고, 곡선의 아래 면적을 계산한 Average Precision (AP)을 통해 알고리즘의 정량적 성능을 비교하였다. TP를 판단하는 기준인 IoU 값을 0.5부터 0.95까지 0.05마다 다르게 적용하여 계산한 AP의 전체 평균 AP50-95를 성능 비교 지표에 추가하였다.

본 논문에서 제안한 적대적 도메인 적응을 통하여 학습된 비전 트랜스포머 인코더의 가중치를 OWL-ViT의 구조에 전이하여 텍스트 쿼리 기반 객체 탐지 성능을 측정하였으며 표 1과 같이 나타내었다. 도메인 적응 후 모델은 적응 전 모델과 비교했을 때, 모든 거리의 데이터에서 더 높은 AP50 수치를 보였다. 또한 1.2m 거리를 제외한 모든 데이터에서 더 높은 AP50-95수치를 보였다. 도메인 적응 후 모델은 적응 전 모델과 비교하였을 때, 정밀도가 다소 감소하였지만, 재현율이 큰 폭으로 상승하였다. 이는 모델이 도메인 적응을 통해 수집한 데이터 셋의 이미지에서도 충분히 과수 특징을 추출하고 인식함으로써 실제 과수에 해당하는 객체 검출이 가능함을 의미한다. 표 2는 2.1m에서 3.0m까지의 거리

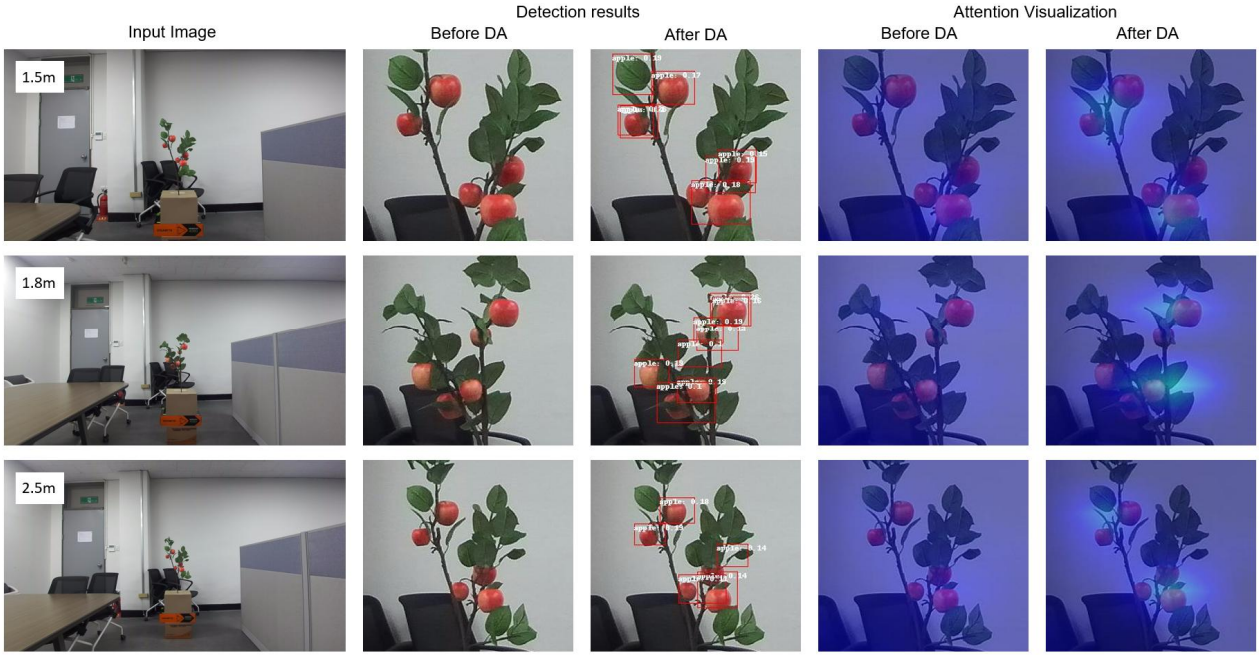

그림 4. 도메인 적응 전·후 검출결과 비교 및 어텐션 시각화
Fig. 4. Comparison of detection results and visualization of attention before and after domain adaptation (DA)

의 실제 데이터 셋에서 도메인 적응을 적용하기 전 후 모델의 비교실험 결과를 보여준다. 도메인 적응 전 모델은 2m 이상의 비교적 먼 거리에 있는 사과들을 전혀 검출하지 못했으나, 도메인 적응 후, 모델은 2m 이상 거리 데이터에서 일정 수준의 사과 검출이 가능하였다. 모든 거리의 데이터를 종합하였을 때, 도메인 적응 후 모델의 성능 수치가 전반적으로 향상된 것을 확인하였다. 정량적 실험결과를 바탕으로, 도메인 적응 과정이 기존 딥러닝 모델의 성능 향상에 도움이 된다는 사실을 간접적으로 확인하였다.

그림 4.는 OWL-ViT 모델의 도메인 적응 전과 적응 후의 검출 결과를 시각화하여 비교한 것이다. 1.5m, 1.8m, 2.1m 거리의 샘플 이미지에 대해서, 모델의 추론을 수행하고 결과 이미지를 확대하여 그림 4와 같이 나타냈다. 타겟 인코더를 구성하는 12개의 인코더 블럭의 어텐션 맵을 이용하는 Attention Rollout을 통해 모델의 검출 결과 기반을 확인하였다 [18]. 12개 블록의 멀티헤드 별 어텐션 맵의 평균을 구하고 작은 값을 0으로 만들며, 이를 차례대로 곱하며 정규화를 수행하여 최종적인 어텐션 맵을 획득한다. 이미지 전체에 대한 표현을 나타내는 첫 번째 토큰인 클래스 토큰에 대한 나머지 토큰들의 어텐션을 추출하여 이를 입력 이미지에 투영하였다. 도메인 적응 전 모델의 경우 샘플 이미지에서 전혀 사과를 검출하지 못하였다. 실제 객체에 해당하는 것 중 모델이 검출한 비율인 재현율 관점에서, 성능이 현저히 낮다는 사실을 알 수 있다. 또한 어텐션 시각화를 통해 모델이 사과의 특징에 집중하지 못하는 것을 확인하였다. 도메인 적응 후 모델의 경우 샘플 이미지에서 충분히 과수의 특징을 추출하고 인식함으로써 실제 과수에 해당하는 객체 검출이 가능해졌음을 알 수 있고, 이것이 재현율 관점에서 성능 상승으로 이어졌음을 표 1을 통해서 확인할 수 있었다. 또한 모델의 객체 검출 판단 근거를 특징 추출 부분에 해당하는 타겟 인코더의 어텐션 맵을 실제 이미지에 투영하여 확인함으로써 알 수 있었다. 도메인 적응을 통해서, 모델의 원거리 객체 검출 능력 향상이 가능함을 정성적인 측면에서 확인하였다. 도메인 적응 전 모델은 실제 과수에 해당하는 객체들을 전혀 검출하지 못하고 있지만, 도메인 적응 후 모델의 경우에도 원거리 사과 검출에는 여전히 높은 검출 정확도를 보장할 수 없다는 한계가 있었다.

매니퓰레이터로 사용된 Niryo NED robot의 구동 범위인 440mm를 고려하여 과수가 구동 범위 내에 존재함을 가정하고 작물 수확 자동화 시스템의 정확도를 측정하였다. 매니퓰레이터 로봇팔의 말단 부분이 가장 가까운 과수에 접근하여 과수 표면에 닿았을 경우를 성공으로 간주하고, 실험 횟수와 성공 횟수의 비율을 정확도로 나타내었다. 매니퓰레이터의 구동 범위를 고려하여 약 400mm 이내의 근거리 접근 상황을 가정하였으나, 뎁스 추정 라이브러리의 한계로 위치 추정이 불가능한 사각 지대가 존재하였다. 객체 검출은 비교적 정확히 수행되었으나, 위치 추정에 어려움이 있어 역기구학을 이용한 매니퓰레이터 제어가 힘든 경우가 다소 있었다. 총 60 번의 매니퓰레이터 제어 실험 중 49번이 성공하였으며, 81.7%의 성공률로 매니퓰레이터의 말단 부분이 과수 위치에 접근 가능함을 실험을 통해 확인하였다.

V. 결 론

본 논문에서는 작물 수확 자동화를 위한 환경적응형 과수 검출 알고리즘을 제안하고, 그 성능을 실제 데이터 셋에서 검증하였다. 시각-언어 기반의 객체 검출 알고리즘인 OWL-ViT의 이미지 특징 추출을 담당하는 비전 트랜스포머 인코더에 적대적 손실 기반의 도메인 적응을 적용하였다. 이를 통해 OWL-ViT는 비교적 실제 환경과 유사한 타겟 도메인의 이미지에서도 과수의 특징을 잘 검출할 수 있도록 학습되고, 실제로 실험 결과를 통해 이것이 가능해졌음을 확인할 수 있었다. 또한 검출 알고리즘의 추론 결과를 기반으로 과수의 실제 위치로 매니퓰레이터를 조정할 수 있음을 실험적으로 확인하였다. 본 논문에서 사용된 OWL-ViT는 레이블 데이터를 포함한 학습 데이터 기반의 재훈련 없이 텍스트 입력만으로 기존에 정의되지 않은 객체 검출이 가능하다는 장점이 있었다. 그러나 실제 도메인의 데이터에서 모델의 검출 정확도를 보장할 수 없는 한계가 있었다. 실제 도메인에서의 검출 정확도를 향상하기 위해, 레이블 데이터가 필요 없는 도메인 적응 기반의 딥러닝 모델 학습 방법을 연구하였다. 도메인 적응 모델의 성능을 검증하기 위해 실제 데이터 셋을 수집하였으며, AP50 및 AP50-95를 포함한 4가지 성능 지표에서 도메인 적응 전 모델과 비교하였다. 객체 검출 모델의 정량적 성능 지표인 AP50 측면에서, 도메인 적응 전 모델에 비해 도메인 적응 후 모델은 2m 이내의 구간에서 약 39.71%의 성능향상을 보였다. 또한 2m 이상의 거리 구간에서 도메인 적응 전 모델은 전혀 과수를 검출하지 못한 것에 비해, 도메인 적응 후 모델은 약 34.46%의 성능 향상을 보이며 어느 정도 과수 검출이 가능해졌음을 확인할 수 있었다. 이를 통해 레이블 데이터가 없는 소량의 이미지 데이터를 바탕으로 도메인 적응을 통해 모델의 성능 향상이 가능함을 확인하였다. 이는 실제 다양한 작업 환경에서, 본 논문에서 제안하는 알고리즘이 간단하고, 편리하게 다양한 과수 검출작업에 적용될 수 있음을 간접적으로 보여준다. 그러나 본 논문에서 수집한 데이터 셋은 실제 다양한 환경을 고려하지 못했으며, 데이터 셋의 양과 다양성이 부족하다는 한계점이 있다. 또한 근거리 대비 원거리의 과수 검출 정확도 측면에서 모델의 성능 개선이 필요하다는 점을 확인할 수 있었다. 다양한 도메인에 대한 강건한 적응을 위해 다양한 품종, 거리, 환경을 고려한 실제 데이터 셋을 확장하여 구성하고, 본 논문에서 제안하는 알고리즘의 안정성을 향상할 수 있는 방안에 대해 지속적인 연구가 필요하다.

References

[1] D. R. Vincent, N. Deepa, D. Elavarasan, K. Srinivasan, S. H. Chauhdary, C. Iwendi, "Sensors Driven AI-based Agriculture Recommendation Model for Assessing Land Suitability," Sensors, Vol. 19, No. 17, pp. 3667, 2019.

[2] M. D. Bah, A. Hafiane, R. Canals, "Deep Learning with Unsupervised Data Labeling for Weed Detection in Line Crops in UAV Images," Sensors, Vol. 10, No. 11, pp. 1690, 2018.

[3] L. Li, S. Zhang, B. Wang, "Plant Disease Detection and Classification by Deep Learning-A Review," IEEE Access, Vol. 9, pp. 56683-56698, 2021.

[4] Y. Onishi, T. Yoshida, H. Kurita, T. Fukao, H. Arihara, A. Iwai, "An Automated Fruit Harvesting Robot by Using Deep Learning," Robomech Journal, Vol. 6, No. 1, pp. 1-8, 2019.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv:2010.11929, 2021.

[6] 남창우, 송지민, 진용식, 이상준, "작물 수확 자동화를 위한 환경적응형 과수 검출 알고리즘," 2023 대한임베디드공학회 추계학술대회, 제주, 2023.

[7] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Oonce: Unified, Real-time Object Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end Object Detection with Transformers," Proceedings of the European Conference on Computer Vision, pp. 213-229, 2020.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," International Conference on Machine Learning, PMLR, pp. 8748-8763, 2021.

[10] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, N. Houlsby, "Simple Open-Vocabulary Object Detection with Vision Transformers," European Conference on Computer Vision, pp. 728-755, 2022.

[11] N. Häni, P. Roy, V. Isler, "MinneApple: a Benchmark Dataset for Apple Detection and Segmentation," IEEE Robotics and Automation Letters, Vol. 5, No. 2, pp. 852-858, 2020.

[12] Sugar Content-Quality Data of Apple in Jeonbuk Jangsu, online available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=490

[13] P. Chu, Z. Li, K. Lammers, R. Lu, X. Liu, "Deep Learning-based Apple Detection Using a Suppression Mask R-CNN," Pattern Recognition Letters, Vol. 147, pp. 206-211, 2021.

[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, "Domain-Adversarial Training of Neural Networks," The Journal of Machine Learning Research, Vol. 17, No. 59, pp. 1-35, 2016.

[15] E. Tzeng, J. Hoffman, J. Saenko, C. Chen, "Adversarial Discriminative Domain Adaptation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7167-7176, 2017.

[16] D. P. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980, 2014.
[17] B. T. Polyak, A. B. Juditsky, "Acceleration of Stochastic Approximation by Averaging," SIAM Journal on Control and Optimization, Vol. 30, No. 4, pp. 838-855, 1992.
[18] S. Abnar, W. Zuidema, "Quantifying Attention Flow in Transformers," arXiv:2005.00928, 2020.

Changwoo Nam (남 창 우)

2022 Division of Electronic Engineering from Jeonbuk National University (B.S.)
2022~Division of Electronics and Information Engineering from Jeonbuk National University (M.S.)

Field of Interests: Artificial intelligence, Computer vision, Deep learning, Robotics
Email: nmcgvv00@jbnu.ac.kr

Jimin Song (송 지 민)

2022 Division of Electronic Engineering from Jeonbuk National University (B.S.)
2022~Division of Electronics and Information Engineering from Jeonbuk National University (M.S.)

Field of Interests: Artificial intelligence, Computer vision, Deep learning, Robotics
Email: jimin_song@jbnu.ac.kr

Yongsik Jin (진 용 식)

2014 School of Electronic and Electrical Engineering from Daegu University (B.S.)
2017 School of Electronic engineering from Kyungpook National University (M.S.)
2022 School of Electronic engineering from Kyungpook National University (Ph.D.)

2019~Electronics and Telecommunications Research Institute (Researcher)
Field of Interests: Artificial intelligence, Robotics
Email: yongsik@etri.re.kr

Sang Jun Lee (이 상 준)

2011 Electrical Engineering from POSTECH (B.S.)
2018 Electrical Engineering from POSTECH (Ph.D.)

Career:
2018~2020 Samsung Advanced Institute of Technology (Senior Researcher)
2020~ Jeonbuk National University (Assistant Professor)
Field of Interesst: Artificial intelligence, Computer vision, Deep learning, Robotics
Email: sj.lee@jbnu.ac.kr