



Mask R-CNN을 활용한 고추 객체 탐지 연구

전경일^a, 김동성^{a,*}

Pepper and Object Detection Using Mask R-CNN

Kyoung-ill Jeon^a, Dong-sung Kim^{a,*}

^a Artificial Intelligence IT Convergence, Soongsil University

ARTICLE INFO

Article history:

Received	15	May	2026
Revised	20	July	2026
Accepted	25	July	2026

Keywords:

Agricultural robot
 Robotic harvesting
 Vision-based control
 Stem detection
 RGB-D camera

ABSTRACT

Labor shortages and aging populations have accelerated the demand for automated harvesting technologies in agriculture. For robust pepper harvesting, this study proposes a Mask Region-based Convolutional Neural Network (Mask R-CNN) based vision system that can simultaneously detect red peppers, green peppers, and stems at the instance level. To address background-object confusion caused by color similarity between green peppers and leaves, a leaf class is introduced, and a four-class model is compared against the original three-class model under identical training conditions. The four-class model improved the segmentation AP for green pepper and stem classes by 3.6%p and 2.0%p, respectively, while maintaining red pepper detection performance. The proposed system integrates instance segmentation with depth information from an RGB-D camera to estimate three-dimensional cutting positions for robotic harvesting. It achieves an average end-to-end latency of approximately 200 ms and a field-harvesting success rate of 90%, thereby confirming its practical applicability for agricultural harvesting robots.

1. 서론

농업 현장은 고령화와 인력 부족으로 인해 자동화 기술의 필요성이 급격히 증가하고 있다. 특히 고추와 같은 채소류의 수확 과정은 많은 노동력을 필요로 하며, 작업 환경의 비정형성으로 인해 기계화가 어렵다. 최근 인공지능(AI)과 딥러닝 기반의 객체 탐지 기술이 농업 분야에 적극적으로 도입되고 있으며, 수확 로봇의 시각 인식 모듈 개발은 필수적 과제로 부상하고 있다. Yang 등^[1]은 합성곱 신경망을 활용하여 감귤 과실과 가지를 동시에 검출하는 연구를 수행하였으며, 이를 통해 CNN 기반 영상 인식이 과수 및 채소 수확 자동화에 활용될 가능성을 제시하였다.

영상처리 시스템은 카메라·센서 입력으로부터 작물의 탐지, 위치 추정, 상태 판단을 수행하는 기술이다. 초기에는 전통적인 영상처리 기법과 신경망을 결합한 접근이 시도되었는데, Husin 등^[2]은 영상처리 기법과 신경망 알고리즘을 결합하여 허브 잎을 인식하는 임베디드 장치를 개발함으로써, 색상·형태 기반 특징 추출과 신경망 분류기의 결합이 실용적인 농업용 인식 시스템 구현에 활용될 수 있음을 보여주었다. 그러나 이러한 초기 접근법은 조명 변화와 배경 잡음이 심한 실제 노지 환경에서는 정확도 확보에 한계가 있는 것으로 알려져 있다.

최근에는 합성곱 신경망 기반 딥러닝 모델이 도입되며 객체 탐지 성능이 크게 향상되었으며, 다양한 네트워크 구조가 농업 영상

* Corresponding author. Tel.: +82-2-820-0713

E-mail address: dongsung@ssu.ac.kr (Dong-Sung Kim).

처리에 적용되고 있다. Park과 Moon^[3]은 사전 학습된 CNN을 활용하여 다양한 환경에서의 객체 탐지 성능을 검증하였으며, 이러한 CNN 기반 접근은 이후 과수 수확 분야로 확장되었다. Stein 등^[4]은 다중 시점 기하학을 활용하여 망고의 탐지, 위치 추정 및 수확량 추정을 수행하였고, Kang과 Chen^[5]은 사과 과수원 환경에서 과실의 탐지·분할과 3차원 시각화를 통합한 시스템을 제안하였다. 이러한 선행 연구들은 CNN 기반 영상 인식이 실제 과수원·농경지 환경에서도 실용적으로 적용될 수 있음을 보여주었으나, 대부분 바운딩 박스 수준의 위치 정보 제공에 그치거나 단일 작물 클래스만을 대상으로 하여, 절단 위치 산출에 필요한 픽셀 단위의 정밀한 형상 정보와 다중 클래스(작물-줄기-잎) 동시 인식은 충분히 다루어지지 않았다.

Mask R-CNN^[6]은 Faster R-CNN^[7]에 ROI Align과 FCN Branch를 추가하여 Instance Segmentation을 지원한다. 즉, 객체의 바운딩 박스뿐만 아니라 픽셀 단위의 마스크까지 예측한다. 주요 구성 요소는 다중 스케일 특징을 추출하는 Backbone (ResNet + FPN)과 객체 후보 영역을 생성하는 Region Proposal Network (RPN), 정밀한 위치 정렬 기능을 갖춘 ROI Align, 픽셀 단위 객체 마스크를 생성하는 Mask Branch로 구성되어 있다^[6,7].

본 연구에서는 Mask R-CNN 기반의 고추 객체 탐지 모델을 제안하여, RGB 영상에서 적색과 녹색의 고추 및 줄기를 동시에 검출하고, 나아가 잎을 독립 클래스로 추가함으로써 배경과의 색상 유사성으로 인한 검출 성능 저하 문제를 개선하고자 하였다.

2. 영상처리 시스템 구현

본 연구에서 제안하는 영상처리 시스템은 수확용 로봇 플랫폼에 적용 가능한 실시간 객체 인식 및 위치 추정 모듈로, Mask R-CNN 기반 딥러닝 모델과 RGB-D 센서를 결합한 구조로 구성된다. 제안 시스템은 고추의 적색·녹색 객체뿐만 아니라 줄기를 독립적으로 인식함으로써 수확에 필요한 절단 위치 계산이 가능하도록 설계하였다.

Fig. 1은 제안한 영상처리 시스템의 전체 구조를 나타낸다. 입력 영상은 ResNet-FPN backbone을 통해 multi-scale feature로 변환되며, RPN을 통해 proposal region이 생성된다. 이후 ROI Align을 이용하여 정밀 정렬된 feature를 추출하고, 이를 classification, bounding box regression 및 mask branch에 전달하여 객체 탐지와 instance segmentation을 수행한다^[6,7].

2.1 데이터 수집 및 준비

2.1.1 데이터 수집 환경

본 연구의 데이터 수집 대상은 노지 재배 환경의 고추로 설정하

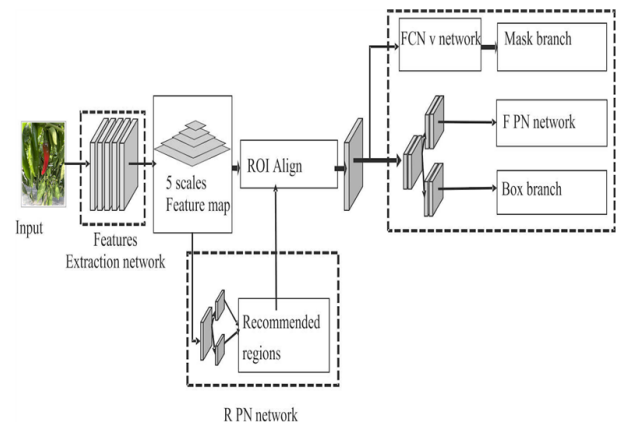


Fig. 1 Structure of the image processing system^[6,7]



Fig. 2 Test bed for growing peppers in open fields

였다. 노지 재배 테스트베드는 Fig. 2와 같이 조성되었으며, 두둑 간 간격은 1800 mm, 두둑 폭은 600 mm, 고추 식재 간격은 600 mm로 구성하였다. 이러한 배치는 수확용 로봇 플랫폼의 이동 폭과 6축 로봇 암의 작업 반경을 고려하여 설계된 것이다.

본 연구에서는 실제 수확 동작에 직접적으로 필요한 객체를 중심으로 데이터셋 클래스를 구성하였다. 초기 3-Class(Red/Green/Stem) 모델에서는 잎을 배경 객체로 간주하고, 줄기 및 고추 객체와의 경계 학습을 통해 간접적으로 특징이 반영되도록 설계하였다. 그러나 녹색 고추와 잎은 색상 및 질감이 유사하여 3-Class 모델에서는 두 객체 간 특징 공간이 충분히 분리되지 못하였으며, 이로 인해 일부 미탐지 및 과분할 현상이 관찰되었다. 이러한 한계를 개선하기 위하여, 본 연구에서는 배경으로 처리되던 잎(leaf)을 독립된 4번째 클래스(4-Class: Red/Green/Stem/Leaf)로 재정의하고, 동일한 학습 조건 하에서 3-Class 모델과 4-Class 모델의 성능을 비교하는 추가 실험을 설계하였다. 잎을 별도 클래스로 분리하여 학습시킬 경우, 녹색 고추와 배경 잎 사이의 특징 공간이 보다 명확히 구분되어 검출 성능이 개선될 것으로 예상하였으며, 이에 대한 구체적인 정량적 비교 결과는 3장 및 4.2절에서 상세히 다루었다.

Table 1은 본 연구에서 사용한 데이터 셋을 요약한 것으로, Number of images는 직접 촬영하여 수집한 원본 이미지의 총 개수를 의미하며, Number of instances는 각 이미지에서 라벨링된

Table 1 Summary of the pepper dataset used in this study

Number of images	Red+ Stem+Leaf	Green+ Stem+Leaf	Red+Green+ Stem+Leaf		Total
	1472	896	3156		5524
Item	Leaf	Red	Green	Stem	Total
Number of instances	8862	3533	2240	5773	20408
Training set (70%)	6204	2473	1568	4041	14286
Validation set (20%)	1772	707	448	1155	4082
Test set (10%)	886	353	224	577	2040
Data augmentation: x6 (Rotate $\pm 15^\circ/5^\circ$)	8862	21198	13440	34638	78138

**Fig. 3 Experiments in a flowerpot test bed****Fig. 4 Experiments on an open field test bed**

객체의 전체 개수를 의미한다. 데이터의 수를 늘리기 위하여 Red, Green, Stem의 이미지를 $\pm 15^\circ$ 범위 안에서 5° 간격으로 회전시켜 데이터수를 증강시켜 작업을 하였다. 또한 학습용, 검증용, 시험용 데이터셋의 분할은 이미지가 아닌 인스턴스 수를 기준으로 산정하였으며 학습 세트에 한해 증강을 적용하였다.

Fig. 3과 Fig. 4는 실내 환경에서의 고추 수확 테스트와 노지 상의 고추 테스트 베드에서 수확하는 모습으로서 고추 영상은 실제 재배 환경에서 RGB 카메라를 이용해 촬영되었으며, 한 이미지 내에 적색 고추, 녹색 고추, 줄기 객체가 동시에 포함되도록 구성하였다. 수집된 대표 영상 예시는 Fig. 5에 나타내었다.

**Fig. 5 Red + green + stem + leaf image****Table 2 Labelme composition of format data**

Component	Explanation
Image (JPG)	Red, Green, Stem, Leaf all in one image
JSON annotation	Labelme format (polygon format + label include)
Class distinction	"001_red_pepper", "001_green_pepper", "001_stem", "001_leaf" format

Table 3 Convert Labelme JSON format to COCO JSON format

Item	Explanation
labelme JSON $\rightarrow$ COCO JSON	Converting class, polygon, and image information
Multi-class support	Map red/green/stem by category_id
Organizing the directory structure	Need to unify into image + JSON pairs or image + mask pairs

2.1.2 데이터 라벨링 및 클래스 정의

수집된 원본 이미지에 대해 Labelme 도구를 활용하여 폴리곤 기반 마스크 라벨링을 수행하였다. 라벨링은 고추 객체를 단순히 하나의 클래스로 정의하지 않고, 실제 수확 작업에 필요한 정보를 반영하여 Red Pepper, Green Pepper, Stem, Leaf와 같이 4개의 클래스로 구분하였다.

각 클래스는 Table 2와 같이 구성되었으며, 하나의 이미지 내에 네개의 클래스가 동시에 포함될 수 있도록 설계하였다.

2.2 Labelme-COCO 변환 파이프라인

Mask R-CNN 학습을 위해서는 COCO JSON 포맷이 요구되므로, Labelme 포맷 데이터를 그대로 사용할 수 없다. 이에 따라 본 연구에서는 Labelme JSON to COCO JSON 변환을 위한 전용 파이프라인을 구축하였다.

변환 과정의 주요 구성은 Table 3에 정리되어 있으며, 클래스 자동 매핑, 폴리곤-세그멘테이션 변환, 디렉토리 구조 통합을 포함하였다.

Fig. 6과 Fig. 7은 변환 결과의 시각화 예시로서, 특히, 실제 현

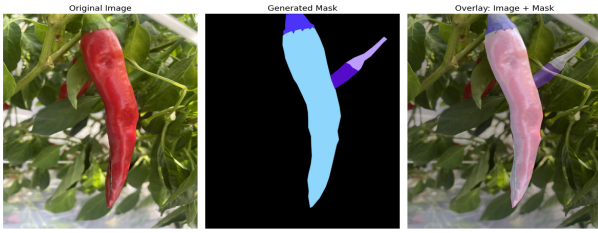


Fig. 6 Visualizing the mask of labels using Labelme



Fig. 7 Generating instance-level masks for labels using Labelme

장 데이터에서 발생할 수 있는 JSON 손상 파일을 고려하여 Safe Load 구조를 적용하였으며, 객체 크기가 일정 기준 이하인 경우 노이즈 객체로 판단하여 자동 필터링하도록 설계하였다.

2.3 모델 설계 및 학습

2.3.1 모델 구조

본 연구에서는 객체의 위치 정보뿐만 아니라 절단점 계산을 위한 형상 기반 공간 정보가 필요하므로, 단순 bounding box 기반 탐지기보다 instance segmentation 기반 구조가 요구되었다. YOLO 계열 모델은 실시간성 측면에서 장점이 있으나, 줄기와 고추가 중첩된 환경에서 픽셀 단위 형상 분리가 제한적이다. 반면 Mask R-CNN은 ROI Align과 mask branch를 통해 객체의 경계를 정밀하게 분할할 수 있어, 줄기 중심 좌표 및 절단 위치 계산에 유리하다. 따라서 본 연구에서는 수확 로봇의 절단 위치 추정을 위해 Mask R-CNN 구조를 채택하였다.

Table 4 YOLO vs Mask R-CNN comparison table

Item	YOLO	Mask R-CNN
Speed	Excellence	Commonly
Bounding box	Possible	Possible
Pixel mask	Unpossible	Possible
Stem separation	Difficulty	Possible
Cut point calculation	Limited	Excellence
Suitability of this study	Commonly	Excellence

Table 5 Detectron2 Key hyperparameters

Hyperparameter name	Value
MODEL.ROI_HEADS.SCORE_THRESH_TEST	0.5
SOLVER.MAX_ITER	30000
INPUT.MIN_SIZE_TRAIN	640, 1024
SOLVER.IMS_PER_BATCH	2
SOLVER.BASE_LR	0.00025

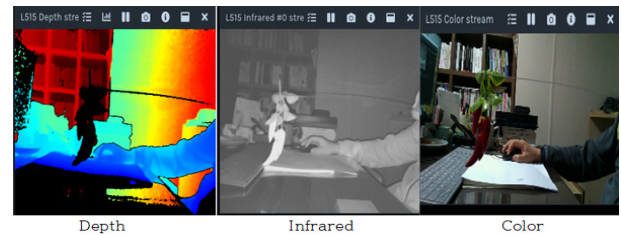


Fig. 8 Depth, infrared, and color images based on RealSense L515

Table 4는 YOLO와 Mask R-CNN을 비교한 YOLO vs Mask R-CNN 비교표이다.

2.3.2 학습 환경 및 파라미터

모델 학습은 Google Colab 환경에서 수행하였으며, Tesla T4 GPU를 사용하였다. 전이 학습을 적용하여 COCO 데이터셋으로 사전 학습된 Mask R-CNN 가중치를 초기값으로 사용하였다. 주요 학습 파라미터는 Table 5와 같다.

학습 과정에서 learning rate warm-up 및 단계적 감소 전략을 적용하였으며, 약 20,000 iteration 이후 손실 함수가 안정적으로 수렴하는 경향을 보였다.

2.4 추론 및 위치 정보 획득

학습된 모델은 별도의 PC 환경에서 추론을 수행하였으며, Detectron2의 DefaultPredictor를 활용하였다. 신뢰도 점수 (confidence score)가 0.5 이상인 객체만을 유효 탐지 대상으로 선정하였다.

Fig. 8은 RealSense L515 기반 Depth, Infrared, Color 영상으로서 추론 결과로부터 얻어진 분할 마스크 기반 중심 좌표는

Table 6 Bounding box and segmentation AP

Item	bbox	segm	Remark
AP	54.796	53.372	Mean average precision (IoU = 0.50:0.95)
AP50	74.305	74.059	Average precision at IoU = 0.50
AP75	61.426	58.775	Average precision at IoU = 0.75
APs	nan	nan	Small objects (not available)
APm	nan	nan	Medium objects (not available)
APl	54.769	53.372	Large objects

Table 7 Class-wise COCO AP comparison (4-Class, segmentation)

Class	AP	AP50	AP75
Stem	42.1	78.2	41.4
Red	86.8	97.7	95.7
Green	44.2	62.5	52.1
Leaf	40.3	57.8	46.0

OpenCV의 moments() 함수를 이용하여 계산되었으며, 해당 좌표는 RealSense L515의 Depth 정보와 정렬되어 3차원 위치로 변환되었다(x_c, y_c, z_c).

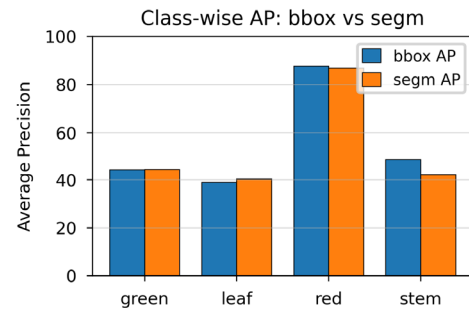
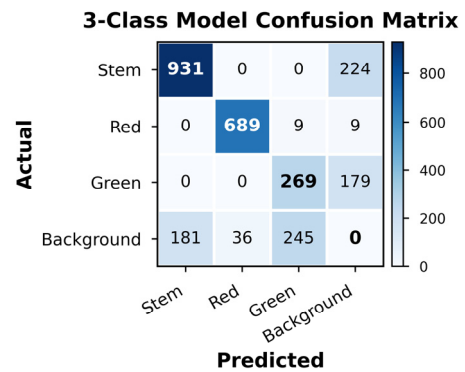
3. 영상처리 시스템 성능 평가

3.1 정량적 성능 평가

3.1.1 전체 COCO 평가 결과(AP/AP50/AP75)

제안한 Mask R-CNN 기반 영상처리 시스템의 객체 탐지 성능은 COCO 평가 기준을 적용하여 Bounding Box 및 Instance Segmentation 성능을 정량적으로 평가하였다. 전체 성능 평가는 Average Precision(AP), AP50, AP75를 사용하였으며, 클래스별 성능은 Stem, Red Pepper 및 Green Pepper, Leaf에 대하여 각각 분석하였다.

Table 6 Bounding Box와 Segmentation의 전체 COCO 평가 결과를 나타낸다. Bounding Box 기준 AP는 54.8%, Segmentation AP는 53.4%를 나타냈으며, AP50에서는 각각 74.3%와 74.1%를 기록하였다. 이는 IoU 0.50 기준에서 대부분의 객체를 안정적으로 검출할 수 있음을 의미한다. 반면 AP75에서는 Bounding Box가 61.4%, Segmentation이 58.7%로 감소하였는데, 이는 높은 IoU 조건에서는 객체의 위치와 마스크 경계가 보다 정밀하게 일치해야 하기 때문으로 판단된다. 또한 APs와 APm이 산출되지 않은 것은 본 데이터셋에 COCO 기준의 Small 및 Medium 객체가 포함되지 않았기 때문이며, 대부분의 객체가 Large 객체로 구성되어 APl만 계산되었다.

**Fig. 9 Class-wise AP comparison****Fig. 10 3-Class confusion matrix of the proposed mask R-CNN model for pepper object detection (IoU = 0.50, confidence score = 0.50)**

3.1.2 클래스별 AP 비교

Table 7은 클래스별 COCO AP 성능을 비교한 결과를 나타낸다. Red Pepper 클래스는 AP 86.8%, AP50 97.7%, AP75 95.7%로 가장 높은 성능을 나타내어 색상과 형상 특징이 명확한 객체에서 매우 우수한 인식 성능을 확인하였다. Stem 클래스는 AP 42.1%, AP50 78.2%, AP75 41.4%를 기록하였으며, 줄기의 폭이 매우 좁고 잎이나 가지와 형태적으로 유사함에도 불구하고 절단 위치 계산에 필요한 수준의 검출 성능을 확보하였다. Green Pepper 클래스는 AP 44.2%, AP50 62.5%, AP75 52.1%로 다른 클래스보다 낮은 성능을 보였으며, 이는 녹색 고추와 잎이 유사한 색상과 질감을 가지므로 객체 간 특징 구분이 어려웠기 때문으로 판단된다. Leaf 클래스는 AP 40.3%, AP50 57.8%, AP75 46.0%로 나타내었다. 이러한 결과는 이후 Fig. 11의 Confusion Matrix 및 Fig. 14의 Precision-Recall Curve에서 나타난 Green Pepper의 상대적으로 높은 미검출(False Negative) 및 Background와의 혼동 경향과도 일치하였다.

Fig. 9는 클래스별 Segmentation AP를 비교한 결과를 나타낸다. Red Pepper 클래스는 가장 높은 AP를 기록하였으며, Stem 클래스는 중간 수준의 성능을 보였다. Green Pepper 클래스는 상대적으로 낮은 AP를 나타냈는데, Leaf와의 색상 유사성 및 객체 가림에 따른 특징 추출의 어려움 때문으로 판단된다. 그러나 Stem

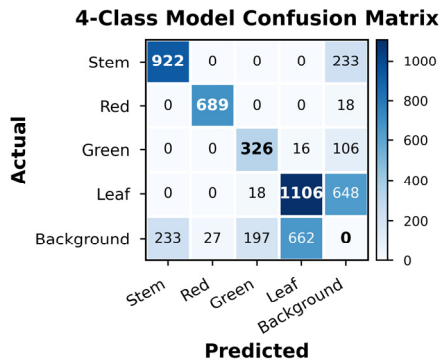


Fig. 11 4_Class confusion matrix of the proposed mask R-CNN model for pepper object detection (IoU = 0.50, confidence score = 0.50)

클래스는 절단 위치 계산에 필요한 최소 수준 이상의 공간 정보를 안정적으로 제공함으로써 실제 수확 로봇 제어에 충분한 성능을 확보하였다.

3.1.3 Confusion Matrix 기반 오분류 분석

Fig. 10과 Fig. 11은 IoU 임계값 0.5와 confidence score 0.5를 기준으로 생성한 3Class과 모델 4Class 모델의 클래스별 Confusion Matrix를 나타낸다. Red Pepper 클래스는 689개의 객체를 거의 완벽하게 분류하여(정답 707건 중 689건, 97%) 가장 높은 분류 성능을 나타내었으며, 이는 3Class 결과와 마찬가지로 배경 및 다른 클래스와의 색상 대비가 뚜렷하기 때문으로 판단된다.

Stem 클래스는 922개의 객체를 정확하게 검출하였으나, 233개의 객체가 Background로 분류되어 일부 미검출(False Negative, 20%)이 발생하였다. 이는 줄기가 가늘고 길며 잎이나 지지대와 형태적 특성이 유사하여 특징 추출이 어려웠기 때문으로 분석된다.

Leaf 클래스는 1772건의 정답 중 1106건(62%)을 정확히 검출하였으나, 648건(37%)이 Background로 오분류되어 5개 클래스 중 Background와의 혼동 비율이 가장 높게 나타났다. 이는 잎이 배경 식생 및 그림자와 시각적으로 유사하여 독립 클래스로서의 경계 학습이 여전히 도전적임을 보여준다.

Green Pepper 클래스는 326개의 객체를 정확하게 검출하였으나(정답 448건 중 73%), 106개(24%)가 Background로, 16개(4%)가 Leaf로 오분류되었다. 3Class 결과(Green의 Background 오분류 비율 약 40%, 179/448)와 비교하면 Leaf 클래스 도입으로 Green의 Background 혼동이 다소 완화된 경향을 보였다.

또한 Background 영역이 Leaf(662건), Green(197건), Stem(233건), Red(27건)로 오검출되는 현상이 관찰되었으며, 특히 Leaf에 대한 오검출 비율이 가장 높게 나타났다(정답 background 표본 1119건 중 59%). 이는 Leaf 클래스 자체가 배경 식생과 매우

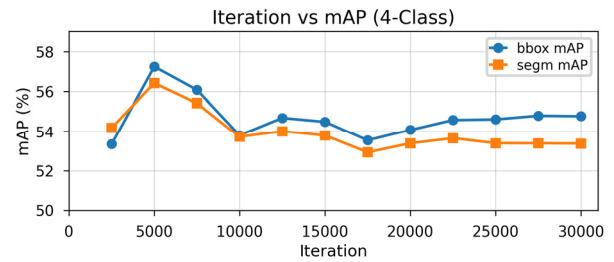


Fig. 12 mAP changes according to iteration

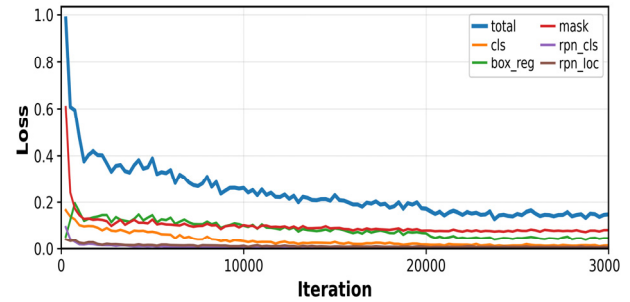


Fig. 13 Loss change trend according to iteration

유사하여 Leaf↔Background 간 양방향 혼동이 여전히 주요 오류 원인임을 시사한다.

Fig. 10의 결과는 Table 7의 클래스별 AP와 Fig. 14의 Precision-Recall 곡선에서 나타난 경향과 일치하였다. Red Pepper 클래스는 높은 AP와 함께 거의 완전한 분류 성능을 보인 반면, Green Pepper와 Stem 클래스는 대부분의 오류가 다른 객체로의 오분류가 아니라 Background로 분류되는 미검출 형태로 발생하였다. 이는 제안한 모델이 객체 간 클래스 구분 능력은 우수하지만, 녹색 계열 객체의 경계 검출에는 여전히 한계가 있음을 의미한다. 따라서 Green Pepper와 Leaf의 특징 공간을 명확히 분리하기 위한 추가 클래스 정의 및 데이터셋 확장이 향후 성능 향상의 핵심 요소가 될 것으로 판단된다.

3.2 학습 수렴 특성

Fig. 12는 Iteration에 따른 bbox 및 segm mAP 변화를 나타낸 것으로 학습 초반 5,000 iteration 부근에서 일시적으로 높은 mAP가 관찰되었으나, 이후 10,000 iteration까지 하락과 조정을 거쳐 20,000 iteration 이후 bbox mAP 약 54.7~54.8%, segm mAP 약 53.3~53.4% 수준에서 안정적으로 수렴하는 경향을 보였다.

Fig. 13은 Iteration에 따른 정확도 및 손실 변화를 나타내었다. total_loss는 초기 약 0.98에서 30,000 iteration 시점에는 약 0.15 수준까지 꾸준히 감소하고 25,000 iteration 이후로는 total_loss가 0.13~0.16 사이에서 안정적으로 수렴하였다. loss_mask는 초반(iter 250 부근 약 0.61)에 가장 큰 비중을 차지했다가 이후 급격

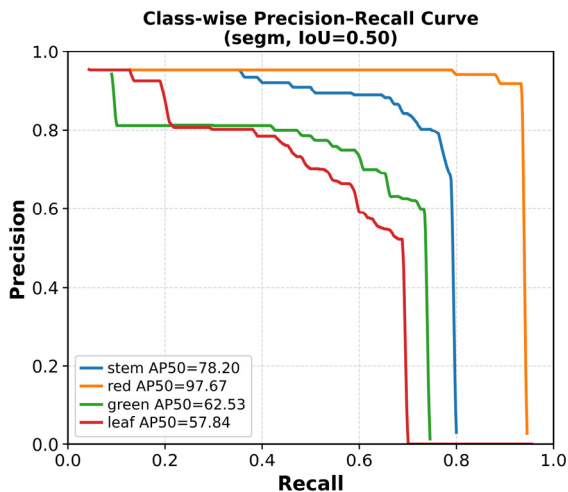


Fig. 14 Class-wise precision-recall curve

Table 8 Class-wise IoU statistics (4-Class model, score ≥ 0.5 , IoU ≥ 0.5 matched detection instances)

Class	N	Mean bbox IoU	Std bbox IoU	Mean mask IoU	Std mask IoU
Red	689	0.936	0.036	0.929	0.028
Stem	922	0.815	0.102	0.774	0.103
Green	326	0.860	0.063	0.862	0.064
Leaf	1106	0.843	0.083	0.856	0.086

히 감소하여 20,000 iteration 부근에서 learning rate가 감쇠 ($2.5e-05 \rightarrow 2.5e-06$)되며 loss 곡선의 변동폭이 눈에 띄게 줄어드는 것이 확인되었다.

Fig. 14는 4-Class 모델에서 IoU=0.50 기준으로 각 클래스 (red, green, stem, leaf)에 대한 Precision-Recall 곡선을 나타낸다. Red 클래스는 대부분의 Recall 영역에서 Precision이 거의 저하되지 않아 가장 우수한 검출 성능을 보였으며, Stem 클래스 역시 Red 클래스에 다소 못 미치나 Green·Leaf 클래스 대비 높은 Precision을 유지하여 두 번째로 우수한 성능을 나타냈다. 반면 Green 클래스는 Recall이 증가할수록 Precision이 상대적으로 빠르게 감소하는 경향을 보였는데, 이는 녹색 고추와 잎 객체 간에 남아있는 색상·형태 유사성 및 가림(occlusion) 현상에 기인한 것으로 판단된다. Leaf 클래스 역시 Green 클래스와 유사한 수준의 곡선 형태를 보였으며, 특히 Recall 중간 구간 이후 Precision 저하가 두드러지게 나타났는데, 이는 Leaf가 형태적으로 배경 색상과 유사하여 Background와 혼동되는 사례가 상대적으로 많았던 것과 일치하는 결과이다. 이러한 결과는 Table 7의 클래스별 AP 비교 결과와 일치하며, 다만 IoU 기준이 높아지는 AP75에서는 Stem의 성능이 Green보다 낮아지는 경향을 보여, 이는 Table 8에서 확인된 Stem 클래스의 큰 마스크 IoU 편차($\sigma=0.103$)와도 부합한다.



Fig. 15 Segmentation results in an object-dense environment



Fig. 16 Recognition results in complex backgrounds

Table 8은 각 클래스별로 IoU ≥ 0.5 기준으로 정답지(true positive)된 인스턴스만을 대상으로, bbox IoU와 mask IoU의 평균 및 표준편차를 산출한 결과이다. Red 클래스는 평균 mask IoU 0.929(표준편차 0.028)로 네 클래스 중 가장 높은 위치 정밀도와 가장 낮은 분산을 동시에 보여, 검출 위치가 매우 안정적이고 일관됨을 확인하였다. 반면 Stem 클래스는 평균 mask IoU 0.774로 상대적으로 낮고 표준편차(0.103) 또한 가장 커서, 줄기 특유의 가늘고 긴 형상적 특성상 마스크 경계의 정밀도가 개체별로 편차가 크다는 것을 시사한다. Green과 Leaf 클래스는 평균 IoU(0.862, 0.856)는 서로 비슷한 수준이나, Leaf 클래스의 표준편차(0.086)가 Green(0.064)보다 다소 높게 나타났다. 이는 4.2절에서 확인된 Leaf - Background 간 분류 혼동과 마찬가지로, Leaf가 배경 색상과 형태적으로 유사하다는 근본 원인에서 비롯된 것으로 추정되며, 다만 이번에는 검출 여부가 아닌 정답지된 인스턴스의 마스크 경계 정밀도 측면에서 그 영향이 나타난 것으로 해석된다.

3.3 정성적 사례 분석

Fig. 15 밀집된 고추 환경에서도 인스턴스 간 분리가 효과적으로 수행되었음을 보여 주었으며, 복잡한 배경에서는 Fig. 16과 같이 복잡한 배경에서도 줄기 객체를 안정적으로 인식하였음을 보여 주었다.

3.4 실시간성 및 현장 적용 결과

전체 파이프라인의 평균 처리 속도는 약 5FPS로 측정되었으며, 구성 모듈별 평균 및 최대 지연 시간은 Table 9에 정리하였으며, 전체 파이프라인의 end-to-end 지연 특성과 처리 속도는 Table 10과 Fig. 17과 Fig. 18에 나타내었다.

Table 9 Average latency per configuration module

Module	Average latency (ms)	Max delay time (ms)	Description
Video reception	20	30	RGB-D video reception and frame buffering
Inference	120	200	Object Detection and Instance Segmentation
Post-processing	30	50	Mask post-processing and center point calculation
Coordinate transformation	15	30	Camera-robot coordinate transformation
Control commands	15	25	Generating and transmitting control commands
Total	200	335	End-to-end latency (worst-case)

Table 10 End-to-end pipeline latency summary

Metric	Value	Description
Average end-to-end latency	200 ms	Average delay from image input to control command output
Maximum end-to-end latency	335 ms	Total pipeline delay under worst-case conditions
Average throughput	5 FPS	Steady-state processing speed in streaming environments
Real-time control criterion	≤ 500 ms	Robot control application criteria
Criterion satisfied	Yes	Meet real-time control requirements

4. 고 찰

4.1 Red Pepper 클래스의 높은 성능 원인 분석

본 연구에서 제안한 Mask R-CNN 기반 영상인식 시스템은 Table 11과 같이 붉은 고추(red pepper) 클래스에서 3Class 학습 시 AP-bbox 88.0%, AP-segm 86.6%의 우수한 성능을 나타내었다. 4Class 학습 이후에도 Red 클래스는 AP-bbox 87.7%, AP-segm 86.8%로 거의 동일하거나 오히려 소폭 향상된 최상위 성능을 유지하였다. 이는 잎 클래스의 추가가 색상·형상 대비가 명확한 적색 고추의 특징 추출에는 영향을 미치지 않았음을 의미하며, 뒤에서 논의할 Green Pepper의 개선 효과와 대조적으로 Red Pepper의 인식 성능은 클래스 구성 변화와 무관하게 안정적임을 확인하였다.

4.2 Leaf 클래스 도입에 따른 Green Pepper 및 Stem 검출 성능 개선 효과

3.1절의 결과에서 나타난 바와 같이, 기존 3-Class 구성에서는

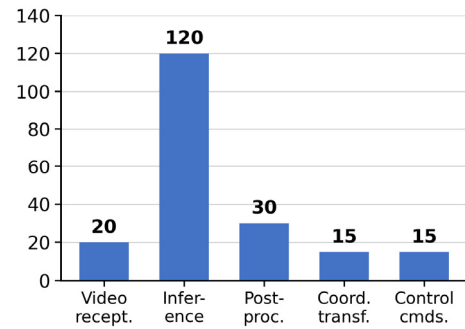


Fig. 17 Average latency of the configuration module

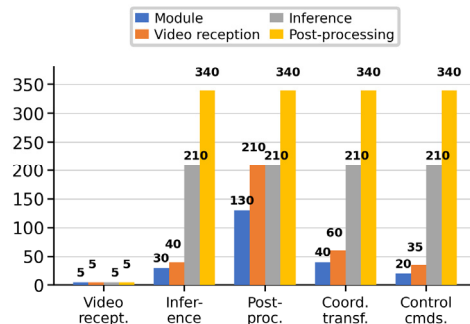


Fig. 18 Latency distribution by component module

Table 11 3-Class vs 4-Class performance comparison (based on identical experimental conditions)

Class	bbox AP			segm AP		
	3-Class	4-Class	Δ	3-Class	4-Class	Δ
Stem	48.36	48.37	+0.01	40.20	42.15	+1.95
Red	88.05	87.69	-0.36	86.61	86.82	+0.21
Green	44.10	44.16	+0.06	40.61	44.24	+3.63
Leaf	—	38.86	New	—	40.27	New

잎을 배경으로 처리함에 따라 녹색 고추와 배경 잎 사이의 특징 공간이 충분히 분리되지 못해 Green Pepper 클래스의 검출 성능이 상대적으로 저하되는 경향이 확인되었다. 2.1절에서 설계한 대로, 동일한 학습 조건 하에서 3-Class 모델과 4-Class 모델을 비교한 결과는 Table 11과 같다.

Table 11에서 확인할 수 있듯이, Green Pepper 클래스의 segmentation AP는 40.61% $\rightarrow$ 44.24%로 3.63%p 향상되었으며, 이는 잎 클래스를 독립적으로 학습시킴으로써 녹색 고추와 배경 잎 사이의 특징 공간이 보다 명확히 분리된 결과로 해석된다. 또한 Stem 클래스의 segmentation AP 역시 40.20% $\rightarrow$ 42.15%로 1.95%p 향상되었는데, 이는 줄기와 인접한 잎 영역의 경계가 명확히 구분됨으로써 줄기 마스크의 경계 정밀도가 개선되었기 때문으로 판단된다.

반면 Red Pepper 클래스는 두 조건에서 거의 동일한 성능을 유지하였으며(86.61% $\rightarrow$ 86.82%), 이는 4.1절에서 논의한 바와 같

이 색상 대비가 뚜렷한 객체는 클래스 구성 변화에 민감하지 않음을 재확인하는 결과이다.

신규 도입된 Leaf 클래스 자체의 성능은 segm AP 40.27% (AP50 57.8%)로, Stem·Green과 유사한 수준을 보였다. 예비 분석(score threshold 0.1~0.3 구간 Confusion Matrix 기준)에서는 Leaf가 Background와 상호 혼동되는 비율이 상대적으로 높게 나타났는데, 이는 잎 자체가 형태적으로 배경 식생과 유사하여 독립 클래스로서의 경계 학습이 여전히 도전적임을 시사한다.

4.3 Stem 인식의 의미 및 절단점 산출 관점의 해석

줄기 클래스는 4Class 학습에서도 여전히 Red Pepper 대비 낮은 AP를 보였으나(seg AP 42.15%), 절단점 산출이라는 실제 응용 목적 관점에서는 3Class 대비 오히려 향상된 성능(+1.95%p)을 확보하였다. 이는 앞서 4.2절에서 논의한 바와 같이 잎 클래스 분리에 따른 부수적 개선 효과로 볼 수 있으며, 줄기가 배경 잎·잔가지·지시대와 형태적으로 유사하다는 기존의 한계가 부분적으로 완화되었음을 시사한다. 그럼에도 줄기 자체의 가늘고 긴 형상적 특성으로 인한 annotation 민감도, 조명·그림자에 의한 경계 손실 문제는 4Class 조건에서도 여전히 존재하는 것으로 판단된다.

그럼에도 불구하고 본 연구에서는 줄기를 단순한 보조 객체가 아닌 독립된 인식 대상으로 정의하고, 인스턴스 세그멘테이션을 적용함으로써 절단점 산출에 필요한 최소 수준의 공간 정보를 안정적으로 확보하였다. 따라서 본 연구에서의 stem 인식은 단순 객체 탐지를 넘어, “수확 로봇 제어를 위한 핵심 기능 객체”라는 의미를 가진다.

4.4 실시간 처리 성능의 의미와 로봇 제어 적합성

실시간 처리 속도 분석 결과, 전체 영상 인식-제어 파이프라인의 평균 응답 시간은 약 200 ms로 측정되었다. 이는 로봇 제어 분야에서 일반적으로 요구되는 500 ms 이내 기준을 만족한다. 전체 지연 시간 중 약 60%가 모델 추론 단계에서 발생하였으며, 이는 Mask R-CNN 구조의 연산 복잡도에 기인한 것으로 분석되었다. 평균 응답 시간 200 ms는 고속 이동체에는 다소 긴 지연 시간일 수 있으나, 본 연구의 대상인 고추 수확 로봇의 작업 환경을 고려할 때 실시간 제어를 위한 충분한 성능을 확보한 것으로 평가된다. 일반적으로 수확 로봇의 말단 장치는 과실의 손상을 방지하기 위해 저속(0.05~0.1 m/s)으로 구동된다. 따라서 5Hz 주기의 비전 피드백은 ROS 기반의 하위 제어가 20 ms~100 ms(10~50Hz) 주기로 수행하는 국부 경로 계획 및 보간 제어를 위한 목표 좌표 업데이트 주기로서 충분히 수용 가능하다. 결과적으로, 본 시스템은 비전 인식의 지연 시간 내에 로봇의 물리적 이동 거리가 오차 범위

Table 12 Model performance evaluation comparison

Item	Zu (2021)	This study
Target crops	Tomato (greenhouse)	Chili peppers (open field)
Input	RGB-D	RGB + Depth synchronization
bbox mAP	84.26	54.7
segm mAP	83.67	53.3
AP50 (bbox / segm)	-	74.3 / 74.1
red Class AP (bbox / segm)	-	87.67 / 86.82
Inference speed	Varies by literature conditions	About 200 ms/frame

이내로 유지되므로, 노지 수확 플랫폼의 실시간 객체 추종 및 수확 작업에 적용하기에 적합함을 확인하였다.

4.5 기존 연구와의 비교 및 본 연구의 차별성

기존 영상처리 연구 중 비교 대상으로 가장 적절한 사례는 Table 12와 같이 Zu 등(2021)은 Mask R-CNN 기반 토마토 인식 방법을 제안한 연구이다^[8]. 이 연구는 RGB-D 융합과 Soft-NMS를 적용하여 bbox mAP 약 54.7~54.8%, segm mAP 약 53.3~53.4% 달성하였다. 그러나 Zu 등의 연구는 비교적 배경이 일정한 온실 환경을 대상으로 수행되었으며, 줄기 객체를 독립적으로 활용하지 않았다는 차이가 있다.

반면 본 연구는 다음과 같은 차별성을 가진다.

첫째, 노지(open-field) 환경을 대상으로 하였다. 노지 환경은 조명 변화, 그림자, 배경 복잡성 및 객체 가림 현상이 심하여 영상인식 난도가 훨씬 높다.

둘째, 줄기(stem)를 독립 클래스화하였다. 이는 단순 객체 검출이 아닌 절단점 산출 및 로봇 제어 연동을 위한 설계라는 점에서 기존 연구와 차별화된다.

셋째, RGB-D 기반 3차원 위치 추정을 실제 로봇 제어와 연동하였다. 대부분의 기존 연구가 객체 검출 단계에 집중한 반면, 본 연구는 “영상 인식 과 절단 위치 계산에 이어 로봇 제어” 까지 포함하는 통합 구조를 구현하였다.

넷째, 실제 농가 환경에서 200회의 수확 실험을 수행하여 약 90% 수준의 성공률을 검증하였다. 이는 실험실 기반 benchmark 수준을 넘어 실제 현장 적용 가능성을 확인하였다는 점에서 중요한 의미를 가진다.

다섯째, 클래스 구성의 한계(배경 잎 처리)를 실험적으로 검증하고 개선하였다. 잎을 독립 클래스로 재정의한 4Class 모델을 통해 Green Pepper(+3.63%p) 및 Stem(+1.95%p) 클래스의 검출 성능이 개선됨을 정량적으로 확인하였으며, 이는 단순 알고리즘 성

능 비교를 넘어 실제 심사 피드백을 반영한 반복적 모델 개선 과정을 보여주는 사례로서 의미가 있다.

종합적으로 본 연구는 단순 알고리즘 성능 비교를 넘어, 실제 수확 로봇 응용을 위한 시스템 통합 관점에서 의미 있는 기여를 가진다고 판단된다.

Table 12는 기존 영상처리 연구와 본 연구의 모델 성능평가를 비교한 것으로 COCO 기준 AP는 bbox 54.7, segm 53.3, AP50은 bbox 74.3, segm 74.1로 나타났으며, red 클래스는 AP-bbox 87.67, AP-segm 86.82로 높은 성능을 보였다. 이 결과는 온실 토마토와 달리 노지 고추의 복잡한 배경, 가림, 줄기 혼재 조건을 고려하면 충분히 실용적인 수준으로 평가할 수 있다. 추론 단계에서는 프레임당 평균 200 ms 수준을 유지하였고, 전체 파이프라인 응답 시간은 약 500 ms 이내였다.

종합적으로, 본 연구에서 제안한 Mask R-CNN 기반 고추 객체 탐지 시스템의 성능은 단일 알고리즘의 우수성보다는, 데이터셋 설계, 인스턴스 분할 전략, 마스크 기반 위치 추정, 실시간 연동 구조가 유기적으로 결합된 결과로 도출되었다고 판단되었다. 특히 “영상 인식 과 절단 위치 계산에 이어 로봇 제어”로 이어지는 일관된 파이프라인은 기존 농업 영상인식 연구와 차별화되는 핵심 기여 요소이다.

이러한 고찰 결과는 본 연구의 접근 방식이 고추뿐만 아니라 토마토, 파프리카 등 유사한 생육 구조를 갖는 작물에도 확장 가능성을 시사하며, 향후 농업용 수확 로봇의 시각 인식 모듈 설계에 중요한 참고 자료가 될 수 있다.

5. 결 론

농업 현장의 노동력 부족 문제를 해결하기 위한 수확용 로봇 플랫폼 개발의 핵심 요소로서, Mask R-CNN 기반 영상인식 시스템을 제안하고, 이를 실제 고추 수확 환경에 적용하여 성능을 실험적으로 검증하였다. 특히 심사 과정에서 제기된 클래스 구성의 한계를 실험적으로 검증·개선하여, 3Class (Red/Green/Stem) 및 4Class(Red/Green/Stem/Leaf) 두 조건에서의 성능을 비교 분석하였다. 본 연구의 주요 성과는 다음과 같이 정리할 수 있다.

첫째, Mask R-CNN 기반 인스턴스 세그멘테이션을 농업 수확 로봇에 적용하여, 고추와 줄기를 픽셀 단위로 분리 인식할 수 있는 영상인식 시스템을 구현하였다. 실험 결과, 붉은 고추 클래스에서는 3Class 및 4Class 조건 모두에서 AP-segm 86% 이상의 높은 정확도를 안정적으로 달성하였으며, 클래스 구성 변화와 무관하게 우수한 인식 성능을 유지함을 확인하였다.

둘째, 줄기를 독립된 인식 대상으로 정의함으로써, 기존의 바운딩 박스 기반 탐지 방식에서 한계로 지적되었던 절단점 산출 문제

를 효과적으로 해결하였다. 마스크 기반 중심 좌표 추출과 깊이 정보 융합을 통해, 절단 위치 계산에 필요한 3차원 좌표 정보를 안정적으로 도출할 수 있음을 확인하였다.

셋째, 배경으로 처리되던 잎(leaf)을 독립된 4번째 클래스로 재정의하고, 동일 조건에서 3Class 대비 4Class 모델의 성능을 정량적으로 비교하였다. 그 결과 Green Pepper 클래스의 segmentation AP가 40.6% → 44.2%로 약 3.6%p 향상되었고, Stem 클래스 역시 40.2% → 42.1%로 약 2.0%p 향상되어, 잎과 배경 간 색상·형태 유사성으로 인한 기존의 검출 성능 저하 문제가 실험적으로 개선됨을 확인하였다. 이는 본 연구가 단순 알고리즘 성능 검증에 그치지 않고, 심사 피드백을 반영한 반복적 모델 개선 과정을 실증하였다는 점에서 의미가 있다.

넷째, 실시간 처리 및 로봇 제어 연동 성능을 검증하였다.

전체 영상 처리-제어 파이프라인의 평균 응답 시간은 약 200 ms로 측정되어, 실시간 로봇 제어 기준(500 ms 이내)을 만족하였다. 또한 프레임 누적 및 안정화 기법을 적용하여,

단일 프레임 오탐이 제어 오류로 직결되지 않도록 시스템 안정성을 확보하였다.

다섯째, 실제 농지 환경에서의 수확 실증 실험을 통해 시스템의 실용성을 검증하였다. 총 200회의 고추 수확 실험에서 90%의 성공률을 기록하였으며, 이는 제안된 영상인식 시스템이 실험실 수준을 넘어 실제 농업 현장에 적용 가능함을 보여주었다.

제안된 인스턴스 세그멘테이션 기반 구조는 고추뿐 아니라 줄기 기반 절단이 필요한 토마토, 파프리카, 가지 등의 과채류 작물에도 적용 가능하다. 특히 RGB-D 기반 3차원 위치 추정 구조는 작물 종류와 무관하게 활용 가능하므로, 향후 다양한 수확 로봇 플랫폼으로 확장될 수 있을 것으로 판단된다. 다만 본 연구의 4Class 실험에서도 Leaf 클래스 자체의 배경과의 경계 모호성은 완전히 해소되지 않았으므로, 향후 연구에서는 RGB-D 기반 깊이 특징 강화, 조명 변화 기반 augmentation, multispectral camera 적용 등을 통해 Leaf 및 Green Pepper 클래스의 검출 성능을 추가로 향상시킬 필요가 있다.

References

- [1] Yang, C. H., Xiong, L. Y., Wang, Z., Wang, Y., Shi, G., Kuremoto, T., Zhao, W. H., Yang, Y., 2020, Integrated Detection of Citrus Fruits and Branches Using a Convolutional Neural Network, Comput. Electron. Agric., 174 105469, <https://doi.org/10.1016/j.compag.2020.105469>.
- [2] Husin, Z., Shakaff, A. Y. M., Aziz, A. H. A., Farook, R. S. M., Jaafar, M. N., Hashim, U., Harun, A., 2012, Embedded Portable

Device for Herb Leaves Recognition Using Image Processing Techniques and Neural Network Algorithm, Comput. Electron. Agric., 89 18-29, <https://doi.org/10.1016/j.compag.2012.07.009>.

- [3] Park, S., Moon, Y. S., 2014, Wild Image Object Detection using a Pretrained Convolutional Neural Network, IEIE Trans. Smart Process. Comput., 3:6 366-371, <https://doi.org/10.5573/IEIESPC.2014.3.6.366>.
- [4] Stein, M., Bargoti, S., Underwood, J., 2016, Image Based Mango Fruit Detection, Localisation and Yield Estimation Using Multiple View Geometry, Sensors, 16:11 1915, <https://doi.org/10.3390/s16111915>.
- [5] Kang, H., Chen, C., 2020, Fruit Detection, Segmentation and 3D Visualisation of Environments in Apple Orchards, Comput. Electron. Agric., 171 105302, <https://doi.org/10.1016/j.compag.2020.105302>.
- [6] He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017, Mask R-CNN, Proceedings of the IEEE International Conference on Computer Vision, 2980-2988, <https://doi.org/10.1109/ICCV.2017.322>.
- [7] Ren, S., He, K., Girshick, R., Sun, J., 2015, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Proceedings of the 29th International Conference on Neural Information Processing Systems, 1 91-99, <<https://dl.acm.org/doi/10.5555/2969239.2969250>>.
- [8] Zu, L., Zhao, Y., Liu, J., Su, F., Zhang, Y., Liu, P., 2021, Detection and Segmentation of Mature Green Tomatoes Based on Mask R-CNN with Automatic Image Acquisition Approach, Sensors, 21:23 7842, <https://doi.org/10.3390/s21237842>.



Kyoung Ill Jeon

Ph. D. Candidate in the Department of Artificial Intelligence IT Convergence, Soongsil University. His research interests include harvesting robot platform and dual arm robot.
E-mail: kijeon@senestechonology.com



Dong-Sung Kim

Professor, School of Artificial Intelligence IT, Soongsil University. His research interests include medical imaging, computer vision, mobile robots.
E-mail: dongsung@ssu.ac.kr