특집논문 (Special Paper)
방송공학회논문지 제27권 제1호, 2022년 1월 (JBE Vol.27, No.1, January 2022)
https://doi.org/10.5909/JBE.2022.27.1.44
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

토마토 위치 및 자세 추정을 위한 데이터 증대기법

장 민 호[a], 황 영 배[b)‡]

Data Augmentation for Tomato Detection and Pose Estimation

Minho Jang[a] and Youngbae Hwang[b)‡]

요 약

농업 관련 방송 콘텐츠에서 과일에 대한 자동적인 정보 제공을 위해서 대상 과일의 인스턴스 영상 분할이 요구된다. 또한, 해당 과일에 대한 3차원 자세에 대한 정보 제공도 의미있게 사용될 수 있다. 본 논문에서는 영상 콘텐츠에서 토마토에 대한 정보를 제공하는 연구를 다룬다. 인스턴스 영상 분할 기법을 학습하기 위해서는 다량의 데이터가 필요하지만 충분한 토마토 학습데이터를 얻기는 힘들다. 따라서 적은 양의 실사 영상을 바탕으로 데이터 증대기법을 통해 학습 데이터를 생성하였다. 실사 영상만을 통한 학습 결과 정확도에 비해서, 전경과 배경을 분리해서 만들어진 합성 영상을 통해 학습한 결과, 기존 대비 성능이 향상되는 것을 확인하였다. 영상 전처리 기법들을 활용해서 만들어진 영상을 사용한 데이터 증대 영상의 학습 결과, 전경과 배경을 분리한 합성 영상보다 높은 성능을 얻는 것을 확인하였다. 객체 검출 후 자세 추정을 하기 위해 RGB-D 카메라를 이용하여 포인트 클라우드를 획득하였고 최소제곱법을 이용한 실린더 피팅을 진행하였고, 실린더의 축 방향을 통해 토마토 자세를 추정하였다. 우리는 다양한 실험을 통해서 대상 객체에 대한 검출, 인스턴스 영상 분할, 실린더 피팅의 결과가 의미있게 나타난다는 것을 보였다.

Abstract

In order to automatically provide information on fruits in agricultural related broadcasting contents, instance image segmentation of target fruits is required. In addition, the information on the 3D pose of the corresponding fruit may be meaningfully used. This paper represents research that provides information about tomatoes in video content. A large amount of data is required to learn the instance segmentation, but it is difficult to obtain sufficient training data. Therefore, the training data is generated through a data augmentation technique based on a small amount of real images. Compared to the result using only the real images, it is shown that the detection performance is improved as a result of learning through the synthesized image created by separating the foreground and background. As a result of learning augmented images using images created using conventional image pre-processing techniques, it was shown that higher performance was obtained than synthetic images in which foreground and background were separated. To estimate the pose from the result of object detection, a point cloud was obtained using an RGB-D camera. Then, cylinder fitting based on least square minimization is performed, and the tomato pose is estimated through the axial direction of the cylinder. We show that the results of detection, instance image segmentation, and cylinder fitting of a target object effectively through various experiments.

Keyword : Data augmentation, Tomato detection, Tomato pose estimation, Fruit detection, Instance segmentation

Ⅰ. 서 론

최근 디지털 영상처리 기술이 급속도로 발전하면서 산업, 의료, 농업 등 다양한 분야에서 컴퓨터 비전 기술이 적용되고 있다[1][2]. 또한 최근 인공지능을 활용한 창작 기술, 방송 및 문화예술 콘텐츠 등의 기술이 제안되고 있다. 특히 다른 분야에 비해 농업 분야는 컴퓨터 비전 기술의 활용될 수 있는 부분이 많이 있지만, 빅데이터 수집과 자동화 시스템 등의 부족으로 앞으로도 많은 연구와 데이터 획득 및 실적용이 필요할 것으로 보여진다. 농업 분야 인구의 고령화 및 노동력 부족으로 인해 컴퓨터 비전 기술을 통한 자동화 연구는 필수적인 현실이다.

최근 농업 분야에서 과일 성숙도 결정, 질병 인식, 해충 제거작업 등에서 심층 컨볼루션 신경망의 사용이 증가하고 있다[3][4][5]. 그리고, 많은 노동력을 요구하는 수확 작업의 자동화를 위해 수확 로봇에 대한 연구가 활발히 진행되고 있다. 과일을 수확할 때 대상 물체를 정확하게 인식하는 것뿐만 아니라 인스턴스 영상 분할을 통한 픽셀 수준의 위치를 인식하는 것도 중요한 작업이다.

최근 Mask-RCNN[6]은 인스턴스 수준의 검출과 영상 분할을 동시에 수행할 수 있는 결과로 인해 농업 분야에서 널리 사용되고 있다. Y. Yu 등은 딸기 검출에 Mask-RCNN을 활용하였고[7] Gonzalez 등은 블루베리 검출에 활용하였다[8]. 과일의 수확을 위해서는 위치 인식뿐만 아니라 자세 추정이 필수적이다. W. Yin 등은 Mask-RCNN을 통해 포도를 검출하고 자세 추정 알고리즘을 제안했다[9]. 또한 N. Wagner 등은 VGG16을 통한 자세 추정 네트워크를 제안하였다[10].

인스턴스 영상 분할을 위해서는 많은 영상과 어노테이션(annotation) 정보가 필요하다. 하지만 농업을 위한 공개된 딥러닝 데이터셋은 다른 분야에 비해서 아직 많지 않고 특히, 본 연구에서 다루는 과일 수확을 위한 인스턴스 영상 분할을 위한 데이터베이스 구축을 위해서는 레이블링 시간이 크게 소요되기 때문에 현실적인 어려움이 있다. 따라서 본 논문에서는 토마토를 대상 객체로 설정하여 데이터 증대를 통해 인스턴스 영상 분할을 수행하였다. 적은 양의 실사 영상 기반으로 하여 전경과 배경을 분리해서 합성하는 기법과 영상 전처리를 통한 데이터 증대 기술을 적용해서 학습한 결과를 각각 양적, 질적으로 비교한다. 또한 인식된 결과를 기반으로 RGB-D 카메라를 통해 포인트 클라우드를 생성시켜 토마토의 자세를 추정하는 연구도 진행하였다.

본 논문의 기여점은 인스턴스 영상분할 관점에서 데이터베이스가 부족한 토마토 인식 문제에 대해서 합성 기반의 방법과 영상 전처리 기반의 방법을 적용하였을 때 어떤 경우가 더 향상이 되는지를 보였으며, 특히 두 가지 방법 모두 데이터 증대를 사용하지 않았을 경우보다 상당한 성능 향상을 보이는 것을 확인하였다. 또한, 토마토에 대한 수확 등 농업 관점에서 다양한 정보를 제공하기 위해서 실린더 피팅 기반의 토마토의 자세 추정 방법을 적용해서 실제 테스트 영상에 대해서 토마토의 자세를 정확하게 추정하는 것을 보였다.

Ⅱ. 관련 연구

1. 데이터 증대

데이터 증대는 학습 시 데이터셋의 크기 및 품질, 다양성 등을 향상하여 더 나은 딥러닝 모델을 구축할 수 있는 기술이다. N. Srivastava 등은 딥러닝 모델 학습 시 과적합 현상을 방지하기 위해 드롭아웃 방법을 제안한다[11]. 드롭아웃은 과적합을 줄이고 다른 정규화 방법보다 크게 개선된다는 결과를 도출한다. T. DeVries 등은 인풋 영상에서 사각 영역을 무작위로 마스킹하는 간단한 정규화 기법 Cutout을 제안한다[12]. 이는 신경망의 견고성과 전반적인 성능을 향상시킬 수 있음을 보여준다. 드롭 아웃과는 달리 네트워크

a) 충북대학교 바이오시스템공학과(Chungbuk National University Dept. of Biosystems Engineering)
b) 충북대학교 지능로봇공학과(Chungbuk National University Dept. of Intelligent Systems &Robotics)
‡ Corresponding Author : 황영배(Youngbae Hwang)
E-mail: ybhwang@cbnu.ac.kr
Tel: +82-43-261-3641
ORCID:https://orcid.org/0000-0002-3400-0493
※ 이 논문은 충북대학교 국립대학육성사업(2020)지원을 받아 작성되었음.
(This research was supported by Chungbuk National University Korea National University Development Project (2020)).
· Manuscript received November 25, 2021; Revised December 30, 2021; Accepted January 14, 2022.

입력 단계에서 삭제되고 그 부분을 0으로 채운다는 점이 다르다. 또한 삭제되는 박스 크기에 따라 성능이 바뀌는 점이 특징이다.

Z. Zhong 등은 입력 영상의 작은 영역에 대해 지나치게 집중하는 것을 방지하기 위해 무작위 영역을 지우는 Random Erasing을 제안한다[13]. 이 과정에서 폐색 (occlusion)을 가진 학습 영상이 생성되어 폐색에 강한 모델이 생성된다. 랜덤 크롭은 배경의 기여를 줄이고 객체에 초점을 맞춰 객체 일부를 기반한 학습 모델을 만들 수 있다. 이와 비교하여 논문에서 제시하는 Random Erasing은 객체의 전체 구조를 유지하며 일부만을 가릴 수 있다. 또한 지워진 영역을 임의 값으로 지정할 수 있기 때문에 영상에 노이즈를 추가한다고 볼 수 있다. 실험을 통해 랜덤 크롭 (crop) 및 플립 (flip)등 일반적으로 사용되는 데이터 증대 기술을 보완한다는 결과가 도출되었다. 윤상두 등은 단순히 픽셀을 제거하는 대신 제거된 영역을 다른 영상의 패치로 대체하는 Cutmix를 제안하였다[14]. 레이블의 경우 결합 영상 픽셀 수에 비례하여 혼합된다. 이는 학습 중 정보를 제공하지 않은 픽셀이 없다는 특성과 동시에 지역 드롭아웃 (dropout)의 장점을 유지한다.

2. 과일 인식 및 자세 추정

Y. Ge 등은 Mask R-CNN을 통해 딸기를 검출하고 수확 작업의 자동화를 위한 주변 환경을 인식 방법을 제안한다[15]. 영상에서 주변 환경과 딸기를 검출한 뒤 RGB-D 카메라 깊이 정보를 통해 포인트 클라우드를 생성한다. 주변 환경과 딸기와의 관계를 3D 공간상에서 계산한 뒤 수확 안전 영역과 위험 영역을 판단한다. W. Yin 등은 Mask-RCNN을 통해 포도 검출과 RANSAC 알고리즘을 사용한 실린더 모델을 통해 자세 추정을 제안한다[9]. 실사 영상 150장을 통해 데이터 증대 영상 1050장을 학습하였고 테스트는 실사 영상 30장을 증대하여 210장을 통해 진행되었다. 자세는 포인트 클라우드 데이터에 그리드를 나누어서 실린더 모델을 성장시키고 완성된 실린더 모델을 통해 추정한다. N. Guo 등은 3D 재구성, 유클리드 분할, ICP 접근법을 사용하여 포인트 클라우드 기반 자세 추정을 제안한다[16]. 3D 재구성을 통해 얻은 포인트 클라우드를 레이저 스캐너 기반 템플릿과 매칭한 뒤 ICP 알고리즘으로 자세를 추정한다. 과일의 기하학적 이해는 RANSAC 알고리즘을 통해 수행되고 과일은 구, 원통, 원뿔로 분류되었다.

G. Lin 등은 저렴한 RGB-D 센서를 통해 구아바의 위치와 자세 추정을 제안한다[17]. FCN 네트워크를 사용하여 과일 및 가지의 영상 분할을 수행한다. 그 이후, 그룹화를 위해 RGB-D 깊이 영상을 기반으로 유클리드 클러스터링을 적용한다. 분할된 가지의 재구성을 위해 3D 라인 검출 방법이 제안된다. 과일의 자세는 중심 위치와 가장 가까운 가지의 정보를 사용한다. H. Li 등은 피망의 대칭성을 이용한 자세 추정을 제안한다[18]. 먼저 포인트 클라우드의 로컬 평면에 대한 정규분포를 계산한 뒤 후보 평면으로 구분한다. 그 이후 각 후보 평면의 점수를 대칭 평면을 구하기 위해 계산하여 가장 낮은 평면이 피망 포인트 클라우드가 선택된다. 대칭축은 대칭면을 통해 계산될 수 있다.

김정인 등은 토마토 수확 로봇을 위해 YOLOv3 모델로 토마토 객체 검출을 하고 토마토 3차원 위치추정을 통해 매니플레이터 제어 시스템을 개발하였다[19]. Kaggle 토마토 데이터셋 500장을 사용하였으며 3차원 위치추정은 토마토의 중심점을 통해서 이루어졌다. 이우영 등은 숙도별 토마토 추적이 가능한 인스턴스 영상분할 기법을 제안하였다[20]. 본 논문에서는 Mask R-CNN[15]을 이용해서 객체를 분할하였으며, Kalman 필터를 사용해서 추적을 하였다. 두 논문 모두 토마토를 인식한다는 점에서 유사한 부분이 있지만, 본 논문은 인식률을 향상시키기 위한 데이터 증대 기법과 실린더 피팅 기반의 자세 추정을 한다는 점에서 차별성이 있다.

Ⅲ. 데이터 증대

1. 실사 영상 데이터

토마토 실사 영상 데이터는 국립농업과학원에서 제공하였다. 온실 내에서 줄기에 있는 토마토를 촬영하였고 익은 토마토와 안 익은 토마토가 혼합되어있다. 영상은 총 63장이고 각 영상에 3개에서 10개의 토마토가 있다. 실사 영상은 그림 1과 같다.

그림 1. 토마토 실사 영상
Fig. 1. Real tomato images

2. 전경-배경 분리기반 합성 영상 데이터

합성 영상은 cocosynth[21]를 사용하여 생성하였다. 합성 영상을 생성하기 위해서는 전경 영상과 배경 영상이 필요하다. 전경 영상과 배경 영상 각 예는 그림 2와 3에 나타나 있다. 토마토는 주로 토마토 잎과 줄기 주변에 위치하기 때문에 AI hub의 농촌 지식베이스[22]에서 토마토 잎과 줄기 영상 74장을 가져왔다. 전경 영상은 국립농업과학원 실사 영상과 Kaggle 토마토 데이터셋[23]에서 GIMP 2.10.28 영상 편집기를 통해 생성하였다. 생성된 전경 영상은 익은 토마토 134, 안 익은 토마토 14장으로 총 148장이다. 합성 영상을 생성시킬 때, 전경의 최대 개수, 각도 범위, 밝기, 스케일 변경 등의 파라미터가 포함된다. 전경의 최대 개수는 사용자가 지정한 최대값 안에서 랜덤으로 생성되는 것이다. 합성 영상은 1개의 배경 영상과 전경 최대 개수 안에서 무작위로 전경 영상을 선택한다. 또한, 전경 영상의 각도, 밝기, 스케일이 범위 안에서 랜덤으로 변경된다. 토마토는 주로 다양한 각도로 위치하기 때문에 360 ° 안에서 랜덤으로 설정하였다. 최대 전경 개수는 3과 5로 설정하였고 그림 4와 같이 합성 영상 각 10,000장을 생성하였다. 전경-배경 분리기반 합성 영상 데이터는 전경 영상이 배경 영상 내 무작위로 위치하기 때문에 과일 등 전경과 배경을 특정지을 수 있는 분야에만 적용할 수 있다.

그림 2. 전경 영상, 영상에서 하나의 객체를 잘라서 생성
Fig. 2. Foreground image, one object in image is cut

그림 3. 배경 영상, 토마토 잎과 줄기
Fig. 3. Background image, tomato stems and leaves

(a) (b)

그림 4. 최대 전경 개수를 (a)3과 (b)5로 설정한 합성 영상
Fig. 4. Synthetic images from the maximum number of foreground as (a)3 (b)5

3. 영상 전처리를 통한 데이터 증대 영상 데이터

영상 전처리를 통한 데이터 증대는 Albumentation 라이브러리[24]를 통해 수행되었다. Albumentation은 픽셀 단위로 변환시키는 Pixel-Level Transform과 영상 자체에서 변환이 이루어지는 Spatial-Level Transform으로 이루어져 있다. 본 논문에서는 영상의 다양성을 위해 Blur, RandomShadow 등 6개의 Pixel-Level 변환과 Resize, CoarseDropout, RandomGridShuffle 등 7개의 Spatial-Level 변환을 사용하였다. 각 함수는 적용될 확률을 설정할 수 있다.

데이터 증대가 이루어질 때 설정된 확률로 함수의 적용 유무가 결정된다. 토마토는 잎 또는 줄기에 의해 토마토 객체가 일부 가려지거나 그림자 지는 경우가 다분하다. 따라서 영상의 일정 부분을 랜덤으로 삭제하는 CoarseDropout과 영상에 랜덤으로 그림자를 입혀주는 RandomShadow의 확률을 다른 함수들보다 높게 설정하였다.

데이터 증대는 국립농업과학원 실사 영상에 적용하였다. 실사 영상을 통한 데이터 증대 영상의 예는 그림 5와 같다. 실사 영상의 어노테이션을 마스크 영상으로 불러온 뒤에 영상에 적용되는 데이터 증대를 동일하게 적용하였다. 데이터 증대가 적용된 영상과 마스크 영상을 통해 COCO format[25]의 어노테이션 파일을 생성할 수 있다. 어노테이션 파일을 생성할 때 각 객체는 색상을 통해 구분한다. 하지만 실사 영상의 어노테이션을 통해 불러온 마스크 영상은 이진 영상이다. 따라서, 각 객체의 해당 위치에 서로 다른 RGB 값을 입력해주어 각 객체를 구분하는 COCO format 어노테이션 파일을 생성하였다. 영상과 같은 데이터 증대가 적용되고 서로 다른 색상을 가진 마스크 영상의 예는 그림 6과 같다. 영상 전처리를 통한 데이터 증대는 전경이나 배경의 조합에 대한 제한이 없기 때문에 다양한 어플리케이션에 적용이 가능할 것으로 예상된다.

Ⅳ. 모델 학습 및 자세 추정

본 논문은 데이터 증대를 통한 모델 학습과 자세 추정으로 구분되며 전체적인 흐름도(flowchart)는 그림 7과 같다.

1. Mask-RCNN 프레임워크

Mask-RCNN 모델은 컨볼루션 네트워크, 영역제안 네트워크 (Region Proposal Network), 관심영역 정렬 (Region

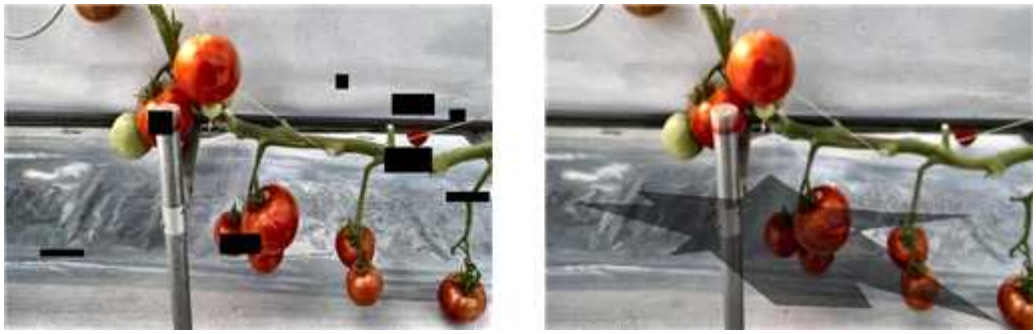
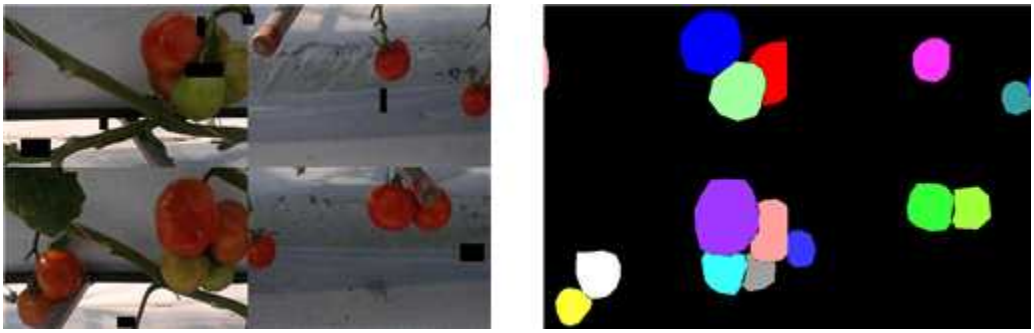


그림 5. 실사 영상을 통한 데이터 증대 영상
Fig. 5. Data augmentation image from real image data



(a) (b)

그림 6. 데이터 증대가 적용된 영상 (a)와 각 객체의 색상이 다른 마스크 영상 (b)
Fig. 6. Image with data augmentation (a) and mask image with different colors each object (b)

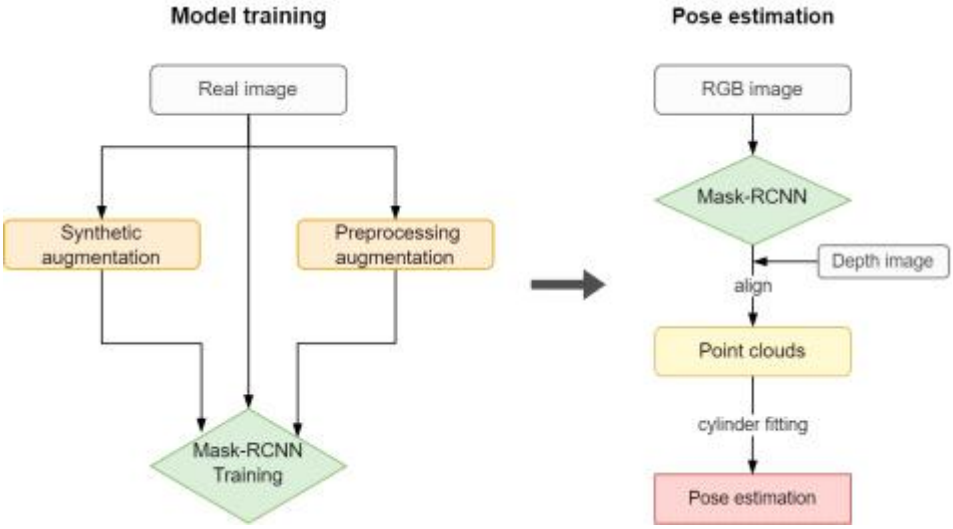


그림 7. 본 논문의 전체적인 흐름도
Fig. 7. The overall flowchart of this paper

of Interest alignment), 분류, 마스크 예측, 바운딩 박스 회귀 분석으로 구성되어있다. Mask-RCNN의 프레임워크는 그림 7에 나타나 있다. 백본 네트워크로 ResNet-101[26]을 사용하였다. 101개의 레이어를 통해 영상의 정보가 담긴 특징(feature)맵이 형성된다. 영역제안 네트워크에서는 특징맵을 입력 (input)하고 후보 바운딩 박스 생성한다. 그 후에 관심 영역 정렬에서 좌표 오류를 없애기 위해 이중 선형 보간법을 사용하여 후보 영역에 해당하는 특징맵을 풀링(pooling)한다. 마지막으로 분류와 바운딩 박스 회귀 파트에서 특징맵의 분류 및 바운딩 박스 회귀를 담당한다. 마스크 예측 파트는 객체의 영상 분할을 담당한다.

2. 실사 영상, 실사 영상을 통한 데이터 증대 영상, 합성 영상을 통한 모델 학습

Mask-RCNN 모델은 사전 학습된 COCO 파라미터에서 전이 학습 (transfer learning)을 통해 재학습 된다. 실사 영상 데이터 63장, 2개의 합성 영상 데이터 각 10,000장, 실사 영상을 통한 데이터 증대 영상은 10,017장으로 구성되어 있다. 각 데이터는 동일한 영상으로 테스트 되었다. 테스트 데이터는 Laboro[27]에서 제공하는 토마토 데이터셋 643장에서 카테고리 6개를 학습데이터와 맞게 1개로 변경 후 사용하였다. 테스트 영상은 그림 8과 같다.

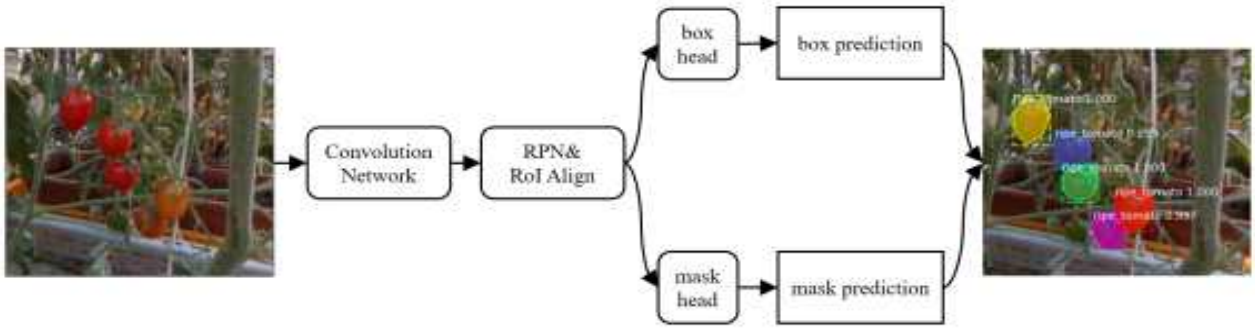


그림 8. 토마토 인식과 객체(instance) 영상 분할테이션을 위한 Mask-RCNN
Fig. 8. Mask-RCNN for tomato detection and instance segmentation

학습에 사용된 배치 사이즈는 1, 학습률은 0.01이고 전체 에포크(epoch) 절반 이상부터 0.001이다. 모든 데이터의 클래스는 토마토와 배경으로 2개이고 총 에포크는 100, 에포크 당 스텝 수는 1,000으로 설정 하였다. 실사 영상과 실사 영상을 통한 데이터 증대 영상 사이즈는 600×800이고 합성 영상은 710×710, 테스트 영상은 3024×4032, 4160×3120로 구성되어 있다. 실사 영상, 영상 전처리를 통한 데이터 증대 영상, 최대 전경 개수를 3으로 설정한 합성 영상, 최대 전경 개수를 5개로 설정한 합성 영상, 총 4번의 학습이 진행되었다. 본 논문에서 사용한 GPU는 Nvidia Geforce RTX 2080 이다.

3. 자세 추정

Mask-RCNN으로 인식된 토마토의 정확한 자세를 추정하기 위해선 깊이 정보가 필요하다. 따라서 토마토의 수확 환경(조명, 일조량, 거리 등)을 고려해 스테레오 액티브 적외선 방식으로 RGB와 깊이 정보를 같이 알 수 있는 Intel 사의 Real Sense D435 센서를 채택하였다.

토마토의 자세를 정확하게 추정하기 위해서는 3D 공간상의 위치를 알아야 한다. 따라서 3D 공간상의 데이터인 포인트 클라우드가 필요하다. RGB-D 카메라를 통해 3채널 RGB 영상과 1채널 깊이 영상을 얻을 수 있다. 이때 깊이 영상은 16비트로 정보의 손실이 적다. RGB 영상과 깊이 영상은 그림 9와 같다. RGB 영상과 그에 해당하는 깊이 영상, 그리고 일련의 카메라 파라미터가 주어지면 Open3D 라이브러리[28]를 통해 두 영상을 정렬(align)시켜 포인트 클라우드를 생성시킬 수 있다. 카메라 파라미터에는 넓이, 높이, fx, fy, cx, cy가 포함되어야 한다. fx, fy는 카메라 초점거리, cx, cy는 주점 (Principal Point)이다. 초점거리는 카메라 렌즈 중심과 센서 사이의 거리를 픽셀로 표현한 것을 의미한다. 주점은 영상 센서에서 수선의 발을 내린 카메라의 실질적인 중심 좌표이다. 카메라 행렬을 바탕으로 식 $x = (u - c_x) * z / f_x,\ y = (v - c_y) * z / fy,\ z = d / depth$

그림 9. Laboro에서 제공한 토마토 데이터셋
Fig. 9. Tomato dataset provided by Laboro

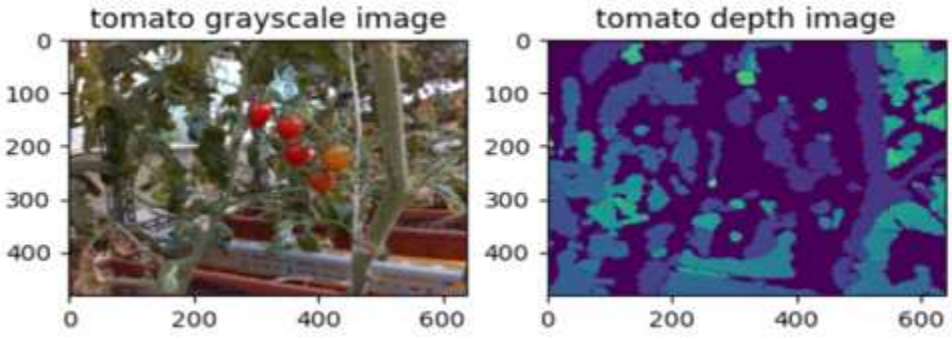


그림 10. RGB 영상과 깊이 영상
Fig. 10. RGB image and depth image

를 도출해 낼 수 있고 이를 통해 포인트 클라우드 데이터가 생성된다. RGB영상과 깊이 영상을 통해 생성시킨 포인트 클라우드는 그림 10에서 볼 수 있다.

그림 11. RGB와 깊이 정렬을 통해 생성된 포인트 클라우드 데이터
Fig. 11. Pointcloud data created of RGB depth align

각 토마토 객체에 대한 자세를 추정하기 위해서 포인트 클라우드에서 토마토 위치를 알아야 한다. 먼저 토마토를 Mask-RCNN을 통해 2D 영상에서 인식한다. 인식된 토마토만 존재하는 영상을 통해 깊이 영상에서 해당 토마토 정보만 제외하고 삭제한다. 인식된 토마토 영상과 토마토 정보만 남은 깊이 영상으로 포인트 클라우드를 생성시켜 3D 공간에서 토마토 위치를 인식한다. 인식된 2D 토마토 영상과 3D 포인트 클라우드에서 토마토 위치 검출은 그림 11과 같다. 본 연구에서는 일반 토마토가 아닌 방울토마토에 대한 자세 추정 연구를 진행하였다. 방울토마토는 일반적으로 양옆보다 위, 아래의 길이가 긴 모양을 지니고 있고 장축을 기준으로 대칭이다. 따라서 방울토마토의 포인트 클라우드 데이터에 실린더를 피팅하여 실린더의 축 방향을 통해 자세를 추정하였다. 실린더 피팅은 최소 제곱법을 이용한 cylinder_fitting[29]을 사용하였다. 최소 제곱법은 실제 데이터와의 오차 제곱의 합 또는 평균이 최소가 되는 해를 구하는 방법이다. 먼저 데이터의 질량 중심을 원점으로 변환한다. 최소 제곱법을 통해 토마토 포인트 클라우드 데이터에서의 중심축을 찾은 뒤 실린더를 피팅한다.

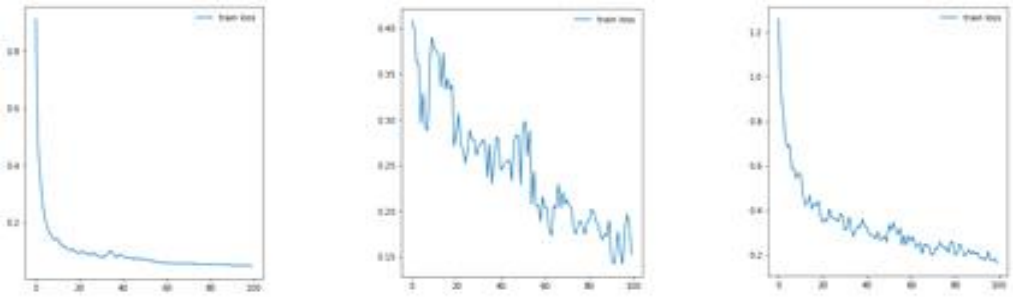
V. 결 과

1. 실사 영상 데이터 학습 결과

실사 영상 데이터의 학습 소요 시간은 약 200분이다. 최종 로스 (loss)는 0.047이고 그래프는 그림 12(a)와 같다. 본 논문의 성능을 평가하기 위한 지표로 평균 정밀도(AP, Average Precision)을 사용하였다. AP는 모든 검출 결과 중 정답을 검출한 비율의 평균이다. 정답은 실제 영역(ground truth)과 예측 영역의 IoU(Intersection of Union)을 통해 판단된다. AP50, AP75에서 뒤의 숫자는 정답 여부를 판단할 때의 신뢰(confidence) 값이고 mAP는 신뢰값 0.5부터 0.95까지의 평균을 의미한다. 테스트 영상 데이터에 의해 계산된 mAP는 표 1과 같이 약 0.4198이다. 실사 영상 데이터는 학습에 사용되는 영상이 적기 때문에 로스가 상대적으로 낮다고 판단된다. 테스트 영상에 대한 영상 분할 결과는 그림 13과 같다.

(a) (b)

그림 12. (a) 2D 방울토마토 인스턴스 영상 분할 영상, (b) 3D 포인트 클라우드의 토마토 위치
Fig. 12. (a) 2D cherry tomato instance segmentation image, (b) Tomato location of 3D point cloud



(a) (b) (c)

그림 13. 각 데이터의 학습 로스 (a) 실사 영상, (b) 합성 영상, (c) 영상 전처리 기반 데이터 증대 영상
Fig. 13. Train loss of each data (a) real image, (b) synthetic image, (c) Pre-processing based data augmentation images

2. 전경-배경 분리 기반 합성 데이터 학습 결과

데이터 증대 전, 합성 기반의 데이터 증대에서 최대 전경 3개와 5개 적용 결과, 영상 전처리 기반의 데이터 증대에 대한 정확도 결과가 표 1에 나타나 있다. 정확도는 테스트 영상에 대해 계산되었다. 합성 기반 데이터 증대 방법 중 더 좋은 결과는 최대 전경 5개로 것이고 mAP는 약 43.765 이다. 이는 데이터 증대 없이 학습한 결과 대비 약 1.78 향상된 수치를 보여준다. 학습 결과 소요 시간은 약 211분이다. 최종 로스는 약 0.1523이고 그래프는 그림 12(b)와 같다. 테스트 영상에 대한 실험 결과는 그림 14와 같다. 데이터 증대 없이 학습한 그림 13과 비교했을 때, 크기가 작은 객체에 대한 인식이 더 좋게 나타났다. 이는 합성 영상을 생성할 때 전경 영상의 스케일 변화를 주기 때문이라고 판단된다. 하지만 안 익은 토마토 전경 영상 14개 만을 가지고 합성 영상을 생성시켰기 때문에 안 익은 토마토에 대한 인식률이 낮다고 판단된다.

표 1. 각 학습 데이터의 평균정확도(AP) 결과
Table 1. AP result of each training data

Types of images	mAP	AP50	AP75
Without data augmentation	41.978	59.580	48.067
Syn. Aug. with Max. foregrounds 3	41.908	60.336	47.845
Syn. Aug. with Max. foregrounds 5	43.764	62.803	50.246
Pre-processing based Data augmentation	47.896	64.833	54.866

3. 영상 전처리 기반 증대 데이터 학습 결과

영상 전처리 기반 증대된 데이터의 학습 결과 소요 시간은 약 207분이다. 최종 로스는 약 0.1619이고 그래프는 그림 13(c)와 같다. 테스트 영상에 의해 계산된 AP는 표 1과 같이 약 47.896이다. 이는 데이터 증대 없이 학습한 결과 대비 약 5.92 향상된 수치를 보여준다. 테스트 영상에 대한

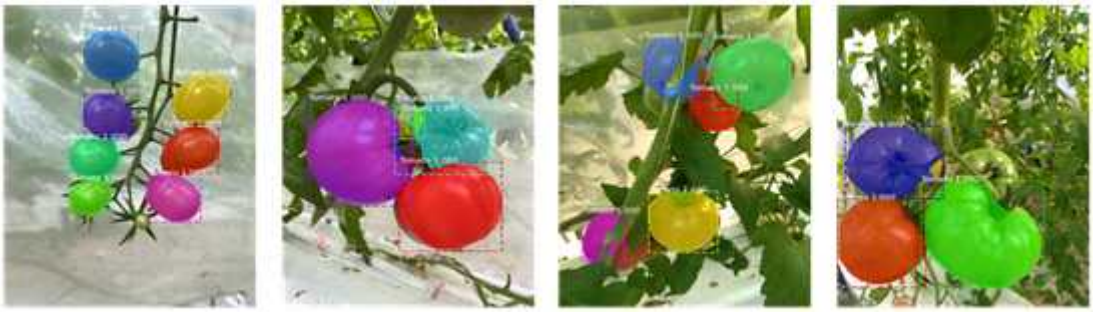


그림 14. 실사 영상 기반 학습 모델의 탐지 및 인스턴스 영상 분할 테스트 결과
Fig. 14. Results of the detection and instance segmentation test of the model learned with real image data

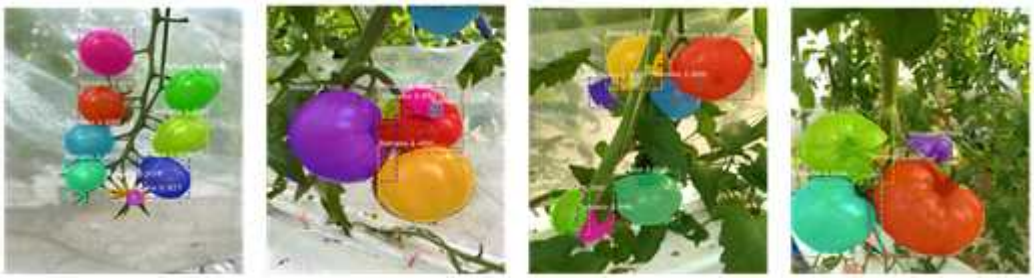


그림 15. 합성 영상 기반 학습 모델의 탐지 및 인스턴스 영상 분할 테스트 결과
Fig. 15. Results of the detection and instance segmentation of the model learned with synthetic image data

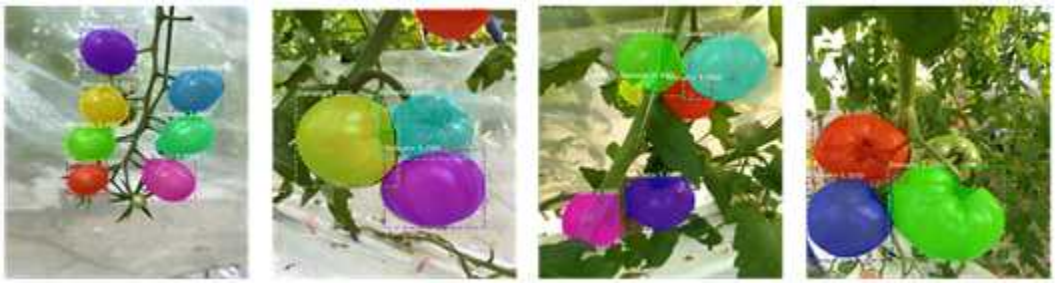


그림 16. 실사 영상을 사용한 데이터 증대 영상 학습 모델의 탐지 및 인스턴스 영상 분할 테스트 결과
Fig. 16. Results of the detection and instance segmentation of the model learned with data augmentation image of real image

실험 결과는 그림 15와 같다. 데이터 증대 없이 학습한 그림 13과 비교하였을 때, 증대된 영상이 폐색에 강하고 정밀한 결과를 보여준다. 하지만 데이터 증대 없이 학습한 모델에서와 같이 작은 객체에 대한 인식 오류를 나타낸다. 실사 영상 내 토마토 객체의 크기가 크기 때문에 실사 영상과 영상 전처리를 기반으로 한 데이터 증대 데이터의 학습 결과 작은 객체에 취약하다고 판단된다. 작는 객체에 좋은 성능을 얻기 위해서는 영상 합성 기반의 데이터 증대가 더 적당하다고 생각된다.

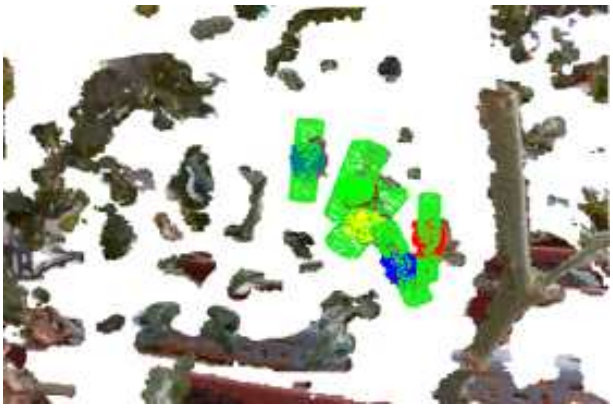


그림 17. 실린더 피팅 결과
Fig. 17. Result of cylinder fitting

4. 실린더 피팅 결과

최소 제곱법을 이용한 실린더 피팅 결과는 그림 16와 같다. 인스턴스 영상 분할된 그림 12(a)와 비교하였을 때, 방울토마토의 자세가 대략 일치하는 것을 알 수 있다.

Ⅵ. 결 론

본 논문에서, Mask-RCNN을 통해 데이터 증대가 적용되지 않은 실사 영상과 2개의 합성 영상 기반의 데이터 증대 데이터, 영상 전처리기반 증대 데이터에 대한 학습을 진행하였다. 그 후에 학습에 사용하지 않은 실사 데이터를 통해

공정한 테스트를 진행하였다. 적은 수의 실사 영상만 이용해서 얻은 결과는 정확도 41.978%이다. 최대 전경 개수 3, 5개로 생성된 합성 영상 중 더 좋은 결과를 나타낸 것은 최대 전경 5개이다. 정확도는 43.765%를 획득하였다. 증대된 실사 영상의 학습 테스트 결과는 47.896%이다. 본 논문에서 수행한 학습데이터 중 가장 좋은 정확도 값을 나타낸 데이터는 영상 전처리 기반의 데이터 증대를 사용한 학습 데이터이다. 실제 영상에 적용한 결과에서 실사 영상을 사용한 모델은 작은 객체에 대한 인식 성능이 좋지 않았다. 합성 영상을 사용한 모델은 작은 객체에 대한 인식은 되었지만 안 익은 토마토에 대한 데이터가 적어 인식 오류가 발생한다고 판단된다. 영상 전처리를 통해 증대된 영상을 사용한 모델은 정교한 인식과 폐색에 대해 강한 결과를 나타냈다. 하지만 실사 영상과 마찬가지로 작은 객체에 대해 인식이 좋지 않은 한계점이 있었다. RGB-D 카메라를 통해 포인트 클라우드를 생성하고 실린더 피팅을 사용해 방울토마토의 자세를 추정하였다. 방울토마토의 자세는 실제와 어느 정도 일치한 것으로 나타났지만 실측 데이터를 구하기 힘들어 정확한 수치적 검증은 할 수 없었기 때문에, 추후에 실측 데이터를 통한 정확도를 검증할 예정이다. 본 방법은 토마토 뿐만이 아니라 다양한 과일에 적용이 가능할 것으로 보이기 때문에 향후 연구로써 토마토 이외의 파프리카, 사과, 딸기 등 유사한 형태의 과일에 대해서 인식과 자세추정을 적용하여 성능을 확인할 예정이다. 본 연구를 통해 적은 수의 농업용 데이터베이스가 주어졌을 때 다양한 데이터 증대 기법을 통해서 인식률을 향상시키고 3차원 자세 추정이 가능하다는 것을 보였다.

참 고 문 헌 (References)

[1] Jongseo Lee, Mangyu Kim, and Hakil Kim, “Camera and LiDAR Sensor Fusion for Improving Object Detection”, JBE Vol. 24, No. 4, July 2019. https://doi.org/10.5909/JBE.2019.24.4.580

[2] Jinbae Park, Teerath Kumar, and Sung-Ho Bae, “Search for Optimal Data Augmentation Policy for Environmental Sound Classification with Deep Neural Networks”, JBE Vol. 25, No. 6, November 2020. https://doi.org/10.5909/JBE.2020.25.6.854

[3] Y. Onishi, T. Yoshida, H. Kurita, T. Fukao, H. Arihara, and A. Iwai, “An automated fruit harvesting robot by using deep learning.” Robomech Journal, Vol. 6, No.13, November 2019.

[4] K. I-His, H. Ya-Wen, Y. Ya-Zhu, C. Ya-Li, L. Yi-Horng, and P. Jau-Woei, “Determination of Lycopersicon maturity using convolustional autoencoders”, Scientia Horticulturae, Vol. 256, No.108538, October 2019.

[5] A. Elhassouny, and F. Smarandache, “Smart mobile application to recognize tomato leaf diseases using convolutional Neural Networks,” International Conference of Computer Science and Renewable Energies, pp. 1-4, 2019.

[6] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN.” In ICCV, pp.2961-2969, 2017.

[7] Y. Yu, K. Zhang, L. Yang, and D. Zhang, “Fruit detection for stawberry harvesting robot in non-structural environment based on mask-RCNN.” Comput Electron Agricult, Vol. 163, No.104846, June 2019.

[8] S. Gonzalez, C. Arellano, and J. E.Tapia, “Deepblueberry: Quantification of blueberries in the wild using instance segmentation.” IEEE Access, Vol. 7, pp. 105776-105788, August 2019.

[9] W. Yin, H. Wen, Z. Ning, J. Ye, Z. Dong, and L. Luo, “Fruit Detection and Pose Estimation for Grape Cluster–Harvesting Robot Using Binocular Imagery Based on Deep Neural Networks.” Frontiers in Robotics and AI, Vol. 8, No.626989, June 2021.

[10] N. Wagner, R. Kirk, M. Hanheide, and G. Cielniak, “Efficient and Robust Orientation Estimation of Strawberries for Fruit Picking Applications”, IEEE International Conference on Robotics and Automation, pp. 13857-13863, May, 2021.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting” The journal of machine learning research, Vol. 15, No.1, pp. 1929-1958, June 2014.

[12] T. DeVries, and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout” arXiv, Vol. 1708, No.04552, 2017.

[13] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random Erasing Data Augmentation” Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No.07, pp. 13001-13008. 2020.

[14] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features” In Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 6023-6032, 2019.

[15] Y. Ge, Y. Xiong, G. L. Tenorio, and P. J. From, “Fruit localization and environment perception for strawberry harvesting robots” IEEE Access, Vol. 7, pp. 147642-147652, October 2019.

[16] N. Guo, B. Zhang, J. Zhou, K. Zhan, and S. Lai, “Pose estimation and adaptable grasp configuration with point cloud registration and geometry understanding for fruit grasp planning” Computers and Electronics in Agriculture, Vol. 179, pp. 105818, December 2020.

[17] G. Lin, Y. Tang, X. Zou, J. Xiong, and J. Li, “Guava detection and pose estimation using a low-cost RGB-D sensor in the field” Sensors, Vol. 19, No.2, pp. 428. January 2019.

[18] H. Li, Q. Zhu, M. Huang, Y. Guo, and J. Qin, “Pose estimation of sweet pepper through symmetry axis detection” Sensors, Vol. 18 No.9, pp. 3083, September 2018.

[19] J. Kim, J. Kim, H. Son, “Development of Deep Learning-based Tomato

Detection and Manipulator Control System for Tomato Harvesting Robot", Institute of Control, Robotics and Systems, pp. 525-526, 2020.
[20] W. Lee, K. Ko, J. Kang, H. Park, I. Jang, "Instance Segmentation based Recognition System Tracking Tomatoes by Ripeness in Natural Light Conditions", Journal of Institute of Control, Robotics and Systems, vol. 26, no. 11, pp. 940-948, 2020.
[21] A. Kelly, cocosynth, 2019, https://github.com/akTwelve/cocosynth.git.
[22] AI Hub, Agricultural knowledge base, 2018, https://aihub.or.kr/aidata/129.
[23] Kaggle, Tomato Detection, 2020, https://www.kaggle.com/andrewmvd/tomato-detection
[24] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations" Information, Vol. 11, No.2, pp. 125, February 2020.
[25] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context" In European conference on computer vision, pp. 740-755, 2014.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.
[27] Laboro, Laboro Tomato, 2020, https://github.com/laboroai/Laboro Tomato.git
[28] Q. Zhou, Y. Park, and V. Koltun, "Open3D: A modern library for 3D data processing" arXiv, Vol. 1801, No.09847 2018.
[29] X. Pan, cylinder_fitting, 2017, https://github.com/xingjiepan/cylinder_fitting.git

저 자 소 개

장 민 호

- 2016년 ~ 현재 : 충북대학교 바이오시스템공학과 학사과정
- ORCID : https://orcid.org/0000-0002-5286-843X
- 주관심분야 : 컴퓨터비전, 데이터 증대

황 영 배

- 2009년 : KAIST 전기 및 전자공학 공학박사
- 2011년 : 삼성테크윈 로봇사업그룹 책임연구원
- 2019년 : 전자부품연구원 지능형영상처리연구센터 책임연구원
- 2019년 ~ 현재 : 충북대학교 지능로봇공학과 조교수
- ORCID : https://orcid.org/0000-0002-3400-0493
- 주관심분야 : 컴퓨터비전, 딥러닝, 의료영상처리