Smart Agricultural Technology 3 (2023) 100130

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

# Fruit maturity grading framework for small dataset using single image multi-object sampling and Mask R-CNN

Punnarai Siricharoen [a,*], Warisa Yomsatieankul [b,c], Thidarat Bunsri [c]

[a] The Perceptual Intelligent Computing Lab, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
[b] Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[c] Biosmart Material and Technology Research Group, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

ARTICLE INFO

*Keywords:*
Multi-object sampling
Instance segmentation models
Small dataset
Maturity grading
Object localization
Pineapple

ABSTRACT

Fruit maturity grading is the key factor for fruit export where maturity consistence and standard are required. This paper proposes a non-destructive approach for pineapple maturity grading and pineapple localization based on object segmentation framework which has enhanced for training a robust model with small dataset. We introduced a multi-object sampling technique in augmentation process to generate images from a small dataset taken under controlled conditions to generalize the model for a more practical dataset. We identify the robustness of object segmentation models, Mask R-CNN over other models, e.g., Faster R-CNN, RetinaNet and CenterMask through mean average precision (mAP), detection ratio to explore the false positives detected, precision-recall curve and computational time. The optimal threshold selection which is crucial especially for sensitive small dataset to achieve high detection performance is proposed in this work using mAP and detection ratio. Our proposed framework enhances the model generalization and achieves mAP of 86.7%, AP50 and AP75 at 97.98% and $AP_{unripe}$, $AP_{partially\_ripe}$ and $AP_{fully\_ripe}$ of 99.20%, 96.58% and 98.63%, respectively. Additional insights of the models developed with our small dataset are explained along with the experimental results which are suggested for future.

## 1. Introduction

Pineapple (Ananas Comosus) is an important tropical fruit which is ranked in the top 10 of world fruit production. From 2018, pineapples are produced an average of 28 million tons per year [1]. Top exporters of fresh fruits on the markets include Costa Rica, Philippines and Thailand. Pineapple production involves the process from cultivation, maintenance, harvest, and packaging [2]. The growth of pineapple from flowering to fruit usually takes about three months. Quality standards for the fresh market must meet including maturity requirement, sizing, uniformity, packaging and many more [3]. Maturity of pineapple fruits is important key for commercial use. Ripened pineapple fruits are usually stored for 4-5 days in normal conditions after harvesting and the ripened fruits are consumed fresh or used for making pineapple juice. For optimal pineapple taste and flavor, the fruit should be harvested when one-third or two-thirds of the peel color has changed from green to yellow.

Maturity grading is currently performed manually based on fruit color appearance which could subjectively result in various grading standards. Many previous work has develop techniques for automatically grading a pineapple fruit using physical and chemical characteristics [4], e.g., stiffness, soluble solid content and non-destructive image processing techniques [5,6], e.g., color, texture, etc. Advanced techniques, such as deep learning is also used for maturity grading classification for each pineapple fruit [7,8]. However, it requires sufficient dataset for training deep learning model to be accurate. In many practical cases, large dataset can be difficult to be acquired [5–8].

In this paper, non-destructive automated maturity grading and localization framework of pineapple fruits are developed especially for practical use. The contributions of this study include (1) single image multi-object sampling in augmentation process to promote a single-object training model to be able to apply with robust multi-object and cluttered background scenarios; (2) due to small dataset leading to highly sensitive models and also different pineapple maturity stage has similar textural and color patterns, the robustness of the segmentation model in our framework is identified; the model is evaluated and compared with state-of-the-arts in terms of mAP, a ratio between ground-truths and detections, precision-recall patterns, computational

* Corresponding author.
*E-mail address:* punarai.s@chula.ac.th (P. Siricharoen).

https://doi.org/10.1016/j.atech.2022.100130
Received 3 July 2022; Received in revised form 6 October 2022; Accepted 9 October 2022
Available online 13 October 2022

time, etc.; (3) the selection of optimal threshold is experimentally proposed using mAP and detection ratio. Finally, our findings provide insights and limitations for learning small dataset, similar pattern but desired for a robust model.

The remainder of this paper is organized as follows. Section 2 presents related work of quality grading of agricultural products using image processing and machine learning. Section 3 provides the methodology including the dataset acquisition and the proposed approach. Experimentation and results are reported in Section 4. Findings are detailed in the discussion in Section 5 followed by conclusions in Section 6.

## 2. Related work

Maturity grading is currently performed manually by harvester which results in various maturity grading standards. To achieve quality standards for fresh fruit export, much previous research have developed an automated system for pineapple maturity grading as summarized in Table 1. Fuzzy logic was applied in [9,10] for differentiating the maturity of the pineapple fruits. Three-side images of the "queen" pineapple fruits have been used to extract color features, particularly using yellow pixel percentage and fuzzy-logic classifier to classify the maturity of a pineapple fruits into unripe, underripe, ripe, and overripe. One pineapple was captured three sides (three images per fruit); an image was obtained under controlled conditions to obtain 80 testing samples and 100 training samples and the overall classification accuracy is approximately at 95%. Color features are commonly employed features for differentiating ripe or unripe pineapple fruits by considering the amount of yellowish regions [5,10]. RGB is converted to HSV color map before used to identify the percentage of yellowish colors in a pineapple fruit [5], and then the maturity of the Smooth Cayenne pineapple is classified using this HSV color features and support vector machine. Similarly, red and green from RGB color model and saturation from HSI color models are used as key features in [10].

Deep neural networks have been widely used in various domains including agricultural applications [11] due to the ability of the networks for self-learning of feature extraction and classification with high precision and performance. Azman and Ismail (2017) [8] used convolutional neural networks for classifying an image of a pineapple fruit into unripe, partially ripe, and fully ripe for optimal harvest. The network architecture is based on two convolution layers and two pooling layers followed by two fully connected layers. The classification accuracy is approximately at 94% on 24 test image samples. Similarly, Chaikaew et al. (2019) [7] proposed pineapple ripeness classification network using transfer learning with MobileNet architecture to classify an image of pineapple with 8 different maturity indices into 3 different types, including unripe, partially ripe, and fully ripe. The model was trained on 100 images per indices and the overall classification accuracy is at 90.77% on test images. In this research an image is acquired under constrained and factorial conditions with only one pineapple fruit on a plain white background.

Convolutional neural networks using transfer learning approaches including ResNet-50, ResNet-101, VGG16, VGG19, GoogleNet and many more were explored in this research. Behera, Rath, and Sethy (2021) [12] determined the maturity of the Lycopersicon fruits using combined techniques; convolutional autoencoder and backpropagation networks are used for background removal and five concentric color feature areas are calculated for individual fruits for determining maturity levels. The maturity grading accuracy reached 100%, but each image must contain only one cropped Lycopersicon fruit with background. Similar image acquisition, Cao et al. (2021) [13] obtained image datasets of zizania which is wild rice and belongs to the glass family; the stem and grain are edible. The images were captured individually in transmission process. Lightweight deep convolutional networks, LightNet is selected due to its compressed box with less parameters and efficient computational time for the classification process to classify zizania quality into high and defective quality. In multi-class categorization in [14] which classify oil palm fruits into 7 different ripeness levels, the convolutional neural networks used DenseNet where each layer has received feature maps of all previous layers that yield classification accuracy at 86%.

From previous work, the accuracy results are promising for an individual object obtained under constrained conditions. Multiple object detection and recognition are selected in some work to handle more practical scenarios in the field. The state-of-the-art techniques have been applied for fruit quantification, machine picking and fruit grading. Gonzalez, Arellano, and Tapia (2019) [15] obtained close-view images of blueberries and employed an instance object segmentation technique, Mask R-CNN comparing various backbones of ResNet-50, ResNet-101. This work handle fruit detection with occlusions, different in sizes and various illumination conditions and the detection accuracy using average precision score (AP50) of 0.909 which considered the IoU score at more than 0.5. Pérez-Borrero et al. (2020) [16] also employed Mask R-CNN with modification for strawberry picking. This work reduced the size of the network architectures, replacing object classifier and bounding box regressor by filtering and grouping the candidate region using non-maximum suppression. mAP score of the modified version is competitive to the original Mask R-CNN but twice computational speed. Fruit instance segmentation is also employed for fruit growth monitoring in [17] by extracting features using ResNet-50 and then fused by feature pyramid network (FPN). Then apply full convolution, ROI layer to refine the feature regions of proposals. The embedding mask branch is applied for pixel-level classification. Position attention module (PAM) is used for fusing importance information pixels which improve the robustness of the segmentation model. The model achieved the mAP score of 73.6% and 61.9% for persimmon and green apple segmentation, respectively. It is a challenging task to localize individual object and classify its maturity stage. Bazame et al. (2021) [18] applied object instance segmentation to classify and detect coffee fruits and mapping the fruits to the maturity stage during harvest time using YOLOv3-tiny because of its high processing speed with single step for predicting object bounding box and probabilities of belonging to the class. In this work, an image is practically captured in a constrained environment of

**Table 1**
Comparison with related work.

| Authors | Category | Techniques | Image samples | Average Accuracy | | Conditions |
|---|---|---|---|---|---|---|
| Aguilar et al. (2021) [5] | Unripe, ripe, overripe, unknown | Color features and Support Vector Machine | Total 100 samples | 95.3% | | Captured in a container |
| Arboleda, de Jesus, and Tia (2021) [6] | Unripe, underripe, ripe, overripe | Color features and Fuzzy-logic | 180 samples | 95% | | One pineapple with white background (manual preprocess) |
| Chaikaew et al. (2019) [7] | Unripe, partially ripe, fully ripe | CNN with MobileNet architecture | 100 images per index for training | 90.77% | | One pineapple with white background |
| Azman and Ismail (2017) [8] | Unripe, partially ripe, fully ripe | CNN with LeNet architecture | 260 image samples | 94% | | One pineapple classification (cluttered background) |
| This paper | Unripe, partially ripe, fully ripe | Mask R-CNN, Faster R-CNN, RetinaNet | 240 image samples | mAP = 86.70%, 87.81%, 91.13% | AP50 = 97.98%, 93.62%, 94.21% | Multiple pineapple fruits detection with maturity grading (cluttered background) |

the coffee harvester. With the advancement of deep learning and robustness of detection models, our research focuses on multiple fruit detection with maturity grading. To robustly train a model, we propose a single image multi-object sampling techniques for image augmentation which increase training size along with an optimal threshold selection.

## 3. Material and methods

Image dataset of pineapple fruits is initially acquired under the controlled conditions where one pineapple fruit is captured on a white background as can be seen in Fig. 1, then each image is then labelled in pixel level; we refer this original image dataset and labels as dataset #1. With limited number of acquired dataset, we apply image augmentation using rotation, scale and translation techniques, and also generate additional image dataset from the pixel-wise label to have image dataset #2 using multiple object sampling on random background which has generated multiple pineapple fruits cluttered background. These datasets are separately created for training and testing object detection and segmentation models. Then these datasets will be experimented with Mask R-CNN and finally optimal threshold selection is experimentally introduced in the last part of the section.

### 3.1. Datasets and augmentation

In some pineapple types, the maturity stage can be identified using its shell color. The color chart in Fig. 2 regarding the details in Table 2 [19] is a grading standard used to identify Maturity Index of a pineapple; the maturity indices have 7 different levels. In practical, the maturity indices have been classified into three levels: unripe, partially ripe, and fully ripe as shown in Table 2 in order to identify period of time that pineapples still stay fresh before reach the customers. For unripe pineapples, the farmer can decide whether to cut them later or if cut, pineapples can be kept for 6-7 days before they become ripe, whereas partially ripe pineapples can be kept fresh for another 3-4 days. Fully ripe pineapples should be consumed or processed as soon as possible. Thus, unripe and partially ripe pineapples practically will be packed for export, whereas fully ripe pineapples will be distributed locally.

#### 3.1.1. Data acquisition

We collect 240 images from total 60 MD2 pineapple fruits (4 images for each pineapple) using a standard smartphone (Samsung Galaxy, 12MP camera) and white paper background as shown in Fig. 3. Due to the pineapple image contains background pixels and this research focuses on the study of pineapple shell for identify different maturity levels, the pixel-level ground-truths are manually labelled in white pixels as shown in Fig. 3 for original images of different maturity levels of unripe, partially ripe, and fully ripe and their corresponding masks.

#### 3.1.2. Multi-object sampling

We generate more data based on the original images with labelled ground-truths to enhance the efficiency of the training data to be more realistic and practical. We use geometric transformation to differ various pineapple scales, orientations, and positions in an image. In addition, the background elements are included in an image. With pixel-wise labelling, we have all pineapple objects stored in the database. Multiple objects are sampled and added to the original dataset. Additional 25% of original images are added in various background images of natural pineapple field, supermarkets, factory, etc., from our private datasets and publicly available datasets [20,21]. Another 25% of original images are added in background images with multiple pineapples (2-5 pineapple fruits in an image) varying sizes, orientations, and positions to imitate the more practical and realistic images containing multiple objects. The multiple-object scenes contain overlapping objects and the examples are displayed in Fig. 4. As can be seen, each pineapple is labelled in pixel-wise and instance level. The pineapple images used for training including the original images and augmented image are completely separated from the tested dataset.

#### 3.1.3. Mask R-CNN

The main task is to detect pineapple fruits in an image and classify its maturity level using object detection or instance segmentation models. The method, Region with CNN features or R-CNN [22] which has been proposed to combine features from different convolutional neural network levels to bottom-up region proposals shows the major improvement in object localization and segmentation. The accuracy relies on the region proposal techniques which are then developed in many R-CNN based methods including Faster R-CNN, and Mask R-CNN.

Mask R-CNN architecture [23] shown in Fig. 5 extends Faster R-CNN [24] which comprises two main components, region proposals using convolutional neural networks (backbone) and a network head for object regression and classification. In addition to Faster R-CNN, Mask R-CNN outputs mask prediction for each region of interest (ROI). The convolutional backbone used in this work is based on the Residual Networks (ResNet) [25] with 50 and 101 layers and ResNeXt [26] networks. Residual networks have skip connections which solves the gradient vanishing problems and the features extracted provides promising classification accuracy in many previous work. The ROI features are extracted from different levels in feature pyramid network. The

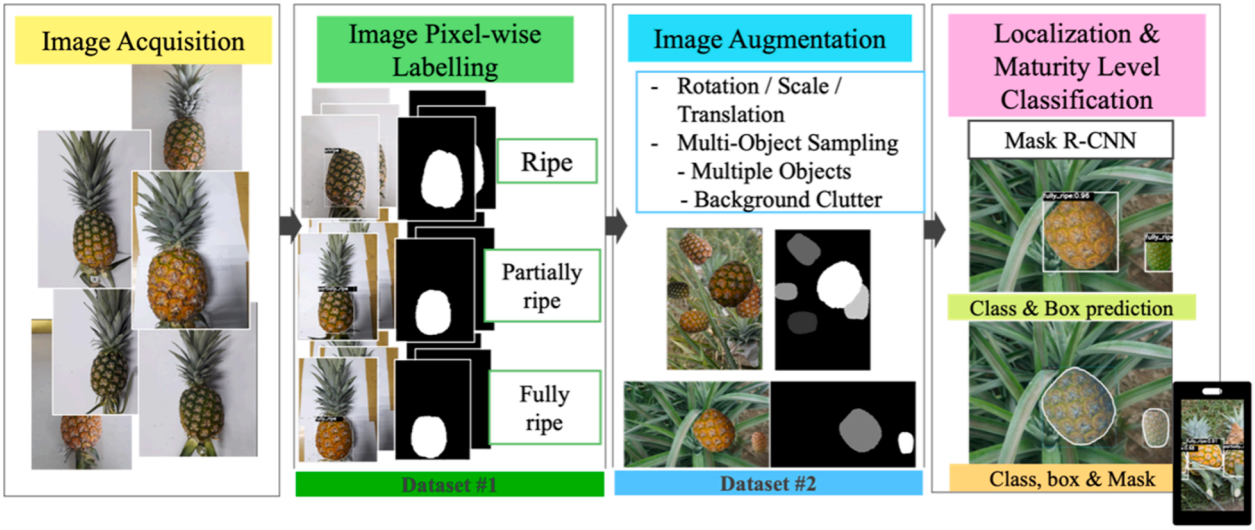


**Fig. 1.** Overview of maturity grading and localization of pineapple fruits.

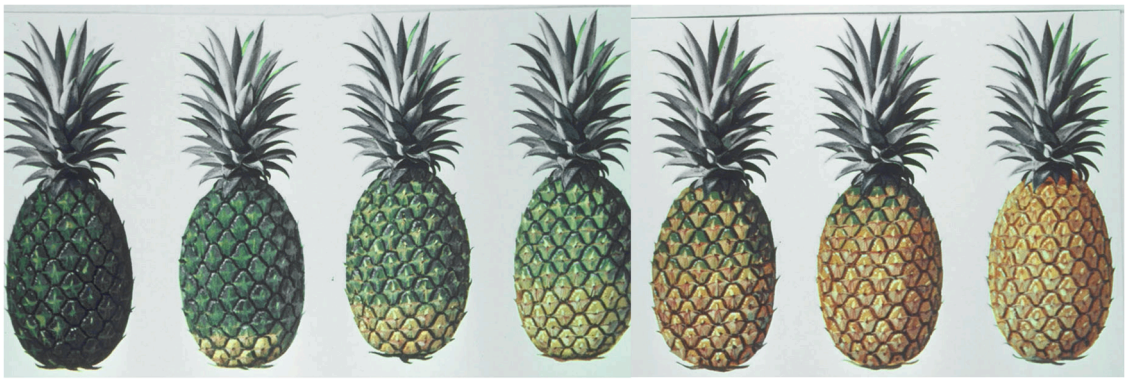

**Fig. 2.** Pineapple color chart for 0-6 maturity indices [19]

**Table 2**
Maturity levels grouping maturity indices and description [19].

| Maturity level | Maturity Index | Description |
|---|---|---|
| Unripe | 0 | Full green |
| | 1 | Slight color break |
| Partially ripe | 2 | Less than ¼ gold |
| | 3 | ¼ to less than ½ gold |
| Fully ripe | 4 | ½ to less than ¾ gold |
| | 5 | ¾ to less than full gold |
| | 6 | Full gold |

feature pyramid network (FPN) has top-down architecture with lateral connections. We employed the effective ResNet-FPN backbone for feature extraction which leverages the high performance in accuracy and computational time. The region proposal network is used for proposing object candidate using attention mechanisms to output a set of candidate objects in a bounding box. The network head of Mask R-CNN using ROI Align which uses bilinear interpolation to compute the exact values of the input features at the sampled locations, then use maximum or average for the combined result. The output of the ROI Align is used to perform the object regression and classification for each candidate object in terms of box prediction and class score. The head network for fully convolutional network (FCN) has included a convolutional mask prediction branch which has some filters for predicting mask using sigmoid and binary loss. The overall training loss includes classification loss ($L_{cls}$), bounding box loss ($L_{box}$) and average binary cross entropy loss ($L_{mask}$).

When dataset is small, we have found the model to be highly sensitive during inferencing. The optimal confidence score threshold in this paper is experimentally proposed to be selected using the relationship between the detection precision (mAP) and the ratio of the number of ground-truths and all detections (false positive and ground-truths) and will be detailed in Experimentation and Results section.

## 4. Experimentation and results

We identify the robustness of the most frequently used tehcniques, Mask R-CNN with many other state-of-the-art models evidenced in [27] where we select models with high performance in terms of average precision (AP). Initially, we identify the improved performance of using instance segmentation model, such as Mask R-CNN over bounding box detection models, such as Faster R-CNN, and RetinaNet. RetinaNet [28] is a single-stage detector applied on densely sampling candidate objects. RetinaNet comprises a backbone network and two-sub networks for box classification and box regression. Due to the class imbalance between background and foreground is the main problem to achieve promising accuracy, RetinaNet has the focal loss by training focus on negative examples. The focal loss is computed from the sum of the focal loss over all candidate objects. Training loss comprises classification loss ($L_{cls}$) which used the proposed focal loss and bounding box loss ($L_{box}$). Additionally, we experimented with another recent segmentation model, CenterMask, which is also a single-stage and anchor-free instance segmentation which has added a spatial attention-guided mask network (SAG-mask) for predicting the object mask. SAG-mask network contains spatial attention map (SAM) which plays a vital role in focusing on informative pixels and reduce noise. We employ a backbone network using VoVNetV2 architecture recommended in CenterMask architecture which outperform ResNet backbone networks in terms of accuracy and inference time. Loss function consists of four losses: center point loss ($L_p$), offset loss ($L_{off}$), size loss ($L_{size}$) and mask loss ($L_{mask}$).

The experiment is conducted using detectron2 framework [29] and PyTorch 1.8 equipped with Google Colab machine with Intel(R) 2-core Xeon(R) CPU @ 2.20 GHz, NVIDIA Tesla V100-SXM2 with 16 GB HBM2 memory. Parameter configurations are empirically selected and based on the total loss for each model; models are configured with the same set of parameters, e.g., max iteration =1000, base learning rate = 0.00025, 2 images per batch, for Mask R-CNN and Faster R-CNN also has a subset size of RPN proposals = 128. An original image has a size of 4128 × 3096 pixels, and an augmented image has a size ranged from 855 × 1280 up to 6016 × 4000 pixels. Each model is initialized with its

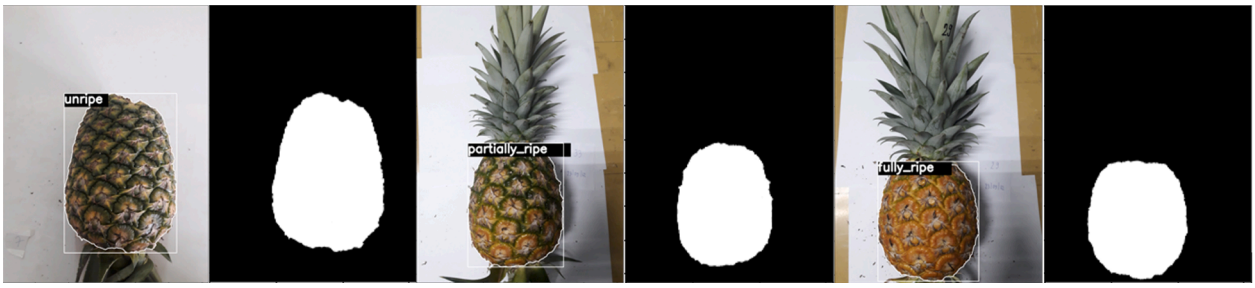


**Fig. 3.** Image samples and its manually labelled ground-truth (in white pixels) for three maturity levels: unripe, partially ripe, and fully ripe.

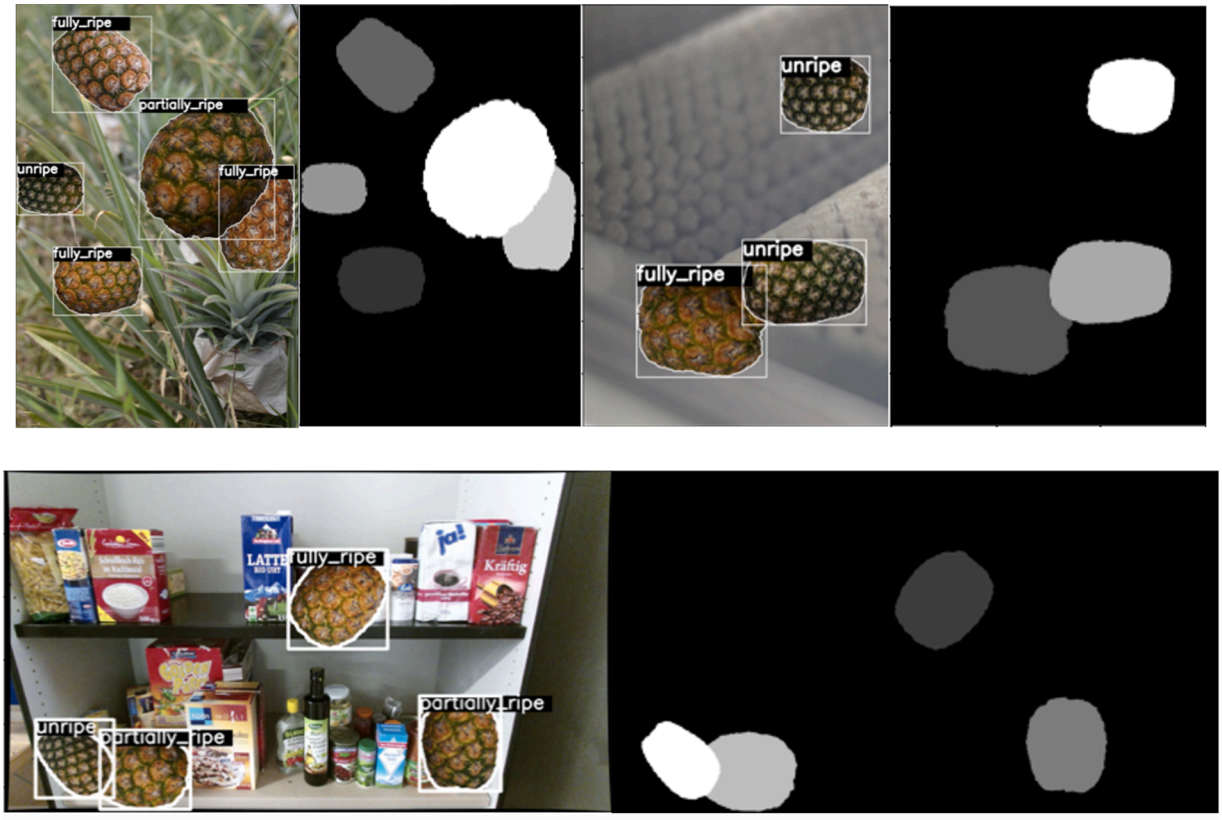


**Fig. 4.** Augmented image samples using single image multi-object sampling and corresponding instance level labelled ground-truths containing multiple and overlapping pineapple fruits varying orientations, scales and positions, and additionally cluttered background.

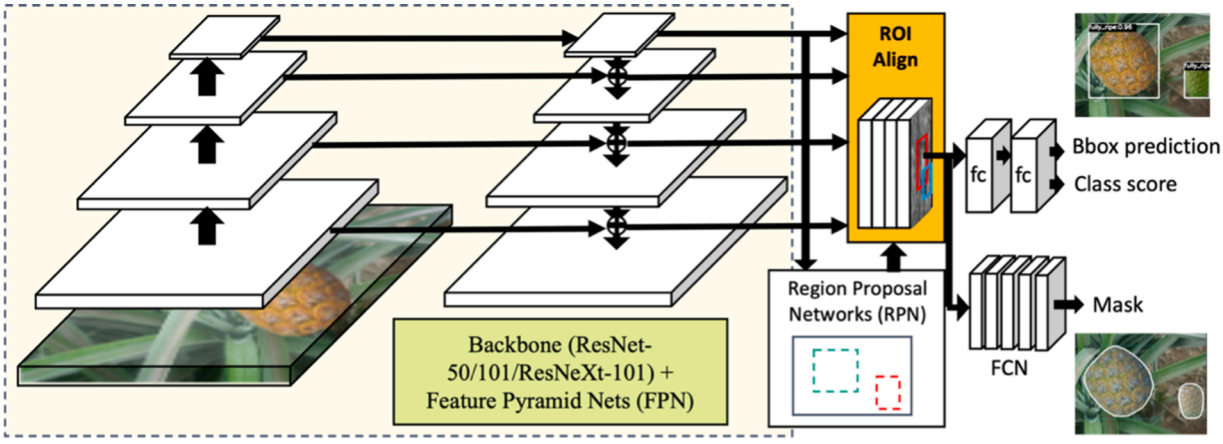


**Fig. 5.** Mask R-CNN used in our framework

original pre-trained model with COCO dataset which is then fine-tuned with our two different datasets: original pineapple dataset (dataset #1) and pineapple dataset with multi-object sampling (dataset #2) presented in Table 3.

### 4.1. Evaluation metrics

To evaluate the performance of the maturity level identification of the pineapple fruit image, we consider the similarity between the detected objects and the ground-truth bounding boxes in terms of the

**Table 3**
Datasets: original and with augmentation datasets for each maturity level.

| Dataset | Training set | | | Testing set | | | Total instances / images |
|---|---|---|---|---|---|---|---|
| | unripe | partially ripe | fully ripe | unripe | partially ripe | fully ripe | |
| Original (#1) | 56 | 56 | 56 | 24 | 24 | 24 | 240/240 |
| Original + Multi-Obj. Sampling (#2) | 105 | 118 | 117 | 41 | 48 | 42 | 471/360 |

overlapping areas and the detected class, e.g., the detected maturity level in our case. Intersection over union (IoU) measures the ratio of the overlapping areas between the bounding boxes of the predicted object ($A_p$) and the ground-truths ($A_{gt}$) [30] as presented in (1)

$$IoU = \frac{A_p \cap A_{gt}}{A_p \cup A_{gt}} \tag{1}$$

The perfect IoU score is when IoU = 1 and IoU threshold is usually set to identify how restrictive the detected object is correct. In addition, precision can be used to identify the correct positive predictions considering all detected objects as shown in (2) and recall is used to identify all relevant objects considering all ground-truth objects in the image as in (3).

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{all\ detections} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{all\ ground - truths} \tag{3}$$

Where TP is the number of the correctly detected objects of the ground-truths; FP is the number of the incorrectly detected objects or non-existing objects; FN is the number of the undetected ground-truth objects.

Also, the confidence score of the object detector can be considered. The larger confidence can be considered as positive detections. To rewrite the precision and recall as a function of confidence threshold ($\tau$) as

$$Precision(\tau) = \frac{TP(\tau)}{all\ detections(\tau)} \tag{4}$$

$$Recall(\tau) = \frac{TP(\tau)}{all\ ground - truths(\tau)} \tag{5}$$

Ideal object detector should have precision and recall equal to one. Precision-recall curve (PR curve) presents the precision-recall relationship for the different thresholds from detection; the PR curve shows in zig-zag pattern as presented in Fig. 6. Average precision (AP) is a commonly used metric which calculates area under the PR curve using all point interpolation techniques in this paper. mAP is an average over classes and threshold of the AP.

### 4.2. Maturity stage classification and localization performance

#### 4.2.1. Classification and localization performance

The detection performance of the pineapple fruit original datasets presents in Table 4; mAP is the mean average precision of the IoU threshold from 0.5 to 0.95 (0.05 step), and AP50 and AP75 are the average precision of the IoU threshold 0.5 and 0.75, respectively. RetinaNet (ResNet-50) shows the highest mAP score and Mask R-CNN and Faster R-CNN show the higher detection precision when the considering specific overlapping threshold at AP50 and AP75. CenterMask shows promising detection performance at 0.5 IoU threshold.

To explore the robustness of this model with unseen and unconstrained dataset, the model trained with the constrained dataset (dataset #1) is then tested with the unconstrained dataset (dataset #2) and the detection performance shows in Table 5. Predictively, mAP drops significantly due to the unseen background clutter, overlapping objects, various object size, and other object variations. Highest mAP is approximately at 75.22% by RetinaNet (ResNet-50) detector. Multi-object sampling technique enhanced the model to be more generalized with unconstrained dataset (dataset #2), the mAP increases for RetinaNet and it shows the highest AP50 and AP75 score using Mask R-CNN as shown in Table 6; the score between AP50 and AP75 is stable because of the high confidence score and accurate localization of the Mask R-CNN model. mAP is calculated by initially ranking confidence score followed by precision-recall curve plot and then calculate the area under the curve. Hence, the detected objects with lowest scores can be plotted and does not change the calculated area under the curve as seen in Fig. 6; many false positives are the plots at recall = 1 (when all ground-truths has been detected and precision is continuously dropped by the number of false positives. The area under the curve in Fig. 6. corresponding mAP in Table 6 for Mask R-CNN and RetinaNet shows large area under the curve using interpolation technique, so the zig-zag pattern which show the false positives found in each model is slightly neglected when calculating mAP. Hence, we analyze the detection models by calculating the ratio of all the false positives compared with the ground-truths and all detections as seen in Table 7. The number of all detections (ignoring the IoU score) is also expanded to show in five different ranges of the confidence score.

In details, R-CNN models, such as Mask R-CNN and Faster R-CNN show acceptable ratio between detections and ground-truths, the ratio between false positives (close to one) and ground-truths (close to zero), and the true positives and ground-truths (close to one); whereas the detections show a lot of false positive objects from the CenterMask and RetinaNet detectors as shown in many detected objects with low confidence score, e.g., more than 60% of detected objects have confidence score between 0.2 – 0.4 by CenterMask detector and more than 80% between 0-0.2 by RetinaNet detector. However, the percentage of detected objects considers all IoU scores, so the detected numbers of all models are large. Nevertheless, the confidence score threshold is crucial choice for accurate detection and acceptable detections of the false positives.

#### 4.2.2. Optimal threshold selection

The optimal confidence score threshold in this paper is experimentally proposed to be selected using the relationship between the

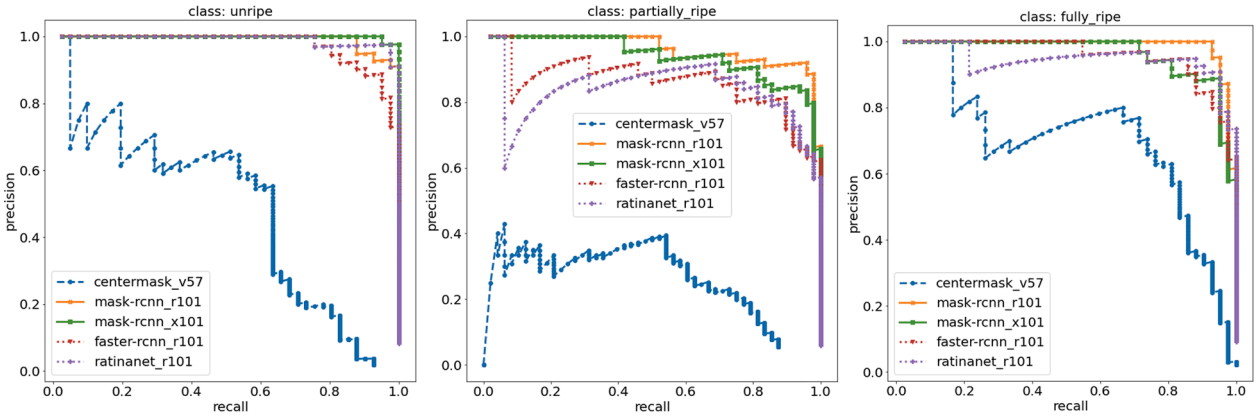


**Fig. 6.** PR curve of each detection model for unripe, partially ripe, and fully ripe class (left to right).

**Table 4**
Detection Performance of The Models Trained and Tested with the Original Datasets (Dataset #1).

| | mAP (%) | AP50 (%) | AP75 (%) | $AP_{unipe}$ (%) | $AP_{partially\_ripe}$ (%) | $AP_{fully\_ripe}$ (%) |
|---|---|---|---|---|---|---|
| CenterMask (VoVNet-99) | 53.63 | 83.36 | 63.58 | 90.33 | 66.20 | 93.82 |
| CenterMask (VoVNet-57) | 37.68 | 80.61 | 9.46 | 96.79 | 70.42 | 74.54 |
| Faster R-CNN (ResNet-101) | 91.01 | 98.94 | **98.94** | 99.18 | **98.80** | 99.07 |
| Faster R-CNN (ResNet-50) | 91.96 | 94.42 | 94.42 | 97.54 | 87.78 | 98.50 |
| RetinaNet (ResNet-101) | 91.96 | 94.42 | 94.42 | 97.54 | 87.78 | 98.50 |
| RetinaNet (ResNet-50) | **94.19** | 95.98 | 95.98 | **99.54** | 89.73 | 98.88 |
| Mask R-CNN (ResNet-101) | 87.47 | 97.47 | 97.47 | 98.92 | 94.65 | 99.17 |
| Mask R-CNN (ResNet-50) | 70.89 | 91.04 | 86.80 | 99.04 | 74.48 | **99.67** |
| Mask R-CNN (ResNeXt-101) | 88.43 | **98.71** | 98.71 | **99.54** | 98.00 | 98.86 |

**Table 5**
Detection Performance of the models trained with original datasets (training dataset #1) and tested with datasets with multi-object sampling technique (testing dataset #2).

| Models | mAP (%) | AP50 (%) | AP75 (%) | $AP_{unipe}$ (%) | $AP_{partially\_ripe}$ (%) | $AP_{fully\_ripe}$ (%) |
|---|---|---|---|---|---|---|
| CenterMask (VoVNet-99) | 32.26 | 54.69 | 34.97 | 54.28 | 36.97 | 73.16 |
| CenterMask (VoVNet-57) | 24.02 | 54.29 | 5.56 | 64.75 | 38.47 | 59.53 |
| Faster R-CNN (ResNet-101) | 69.27 | 78.11 | 77.76 | 87.53 | 56.26 | 90.97 |
| Faster R-CNN (ResNet-50) | 71.12 | 73.58 | 73.58 | 79.12 | 57.78 | 84.23 |
| RetinaNet (ResNet-101) | 69.16 | **78.24** | **78.24** | **83.02** | 59.28 | **93.29** |
| RetinaNet (ResNet-50) | **75.22** | 77.07 | 77.07 | 81.65 | **63.97** | 86.40 |
| Mask R-CNN (ResNet-101) | 62.93 | 72.37 | 72.11 | 73.02 | 59.93 | 84.75 |
| Mask R-CNN (ResNet-50) | 56.29 | 73.62 | 72.26 | 76.06 | 53.46 | 92.00 |
| Mask R-CNN (ResNeXt-101) | 62.76 | 73.00 | 72.90 | 79.47 | 58.25 | 81.71 |

**Table 6**
Detection Performance of the models trained and tested with original and generated datasets (dataset #2).

| Models | mAP (%) | AP50 (%) | AP75 (%) | $AP_{unipe}$ (%) | $AP_{partially\_ripe}$ (%) | $AP_{fully\_ripe}$ (%) |
|---|---|---|---|---|---|---|
| CenterMask (VoVNet-99) | 14.11 | 26.62 | 13.10 | 52.59 | 27.36 | 0.00 |
| CenterMask (VoVNet-57) | 30.19 | 50.86 | 37.19 | 50.16 | 29.72 | 73.00 |
| Faster R-CNN (ResNet-101) | 87.81 | 93.62 | 93.62 | 97.46 | 87.83 | 95.89 |
| Faster R-CNN (ResNet-50) | 77.72 | 96.67 | 93.79 | 98.52 | 94.31 | 97.65 |
| RetinaNet (ResNet-101) | **91.13** | 94.21 | 94.21 | 99.19 | 88.02 | 95.65 |
| RetinaNet (ResNet-50) | 90.11 | 93.09 | 93.09 | 99.73 | 89.96 | 89.90 |
| Mask R-CNN (ResNet-101) | 86.70 | **97.98** | **97.98** | 99.20 | **96.58** | **98.63** |
| Mask R-CNN (ResNet-50) | 82.81 | 94.76 | 94.76 | 97.18 | 89.36 | 98.00 |
| Mask R-CNN (ResNeXt-101) | 82.98 | 96.68 | 96.68 | **99.88** | 94.11 | 96.37 |

**Table 7**
Detection Ratios Of The Models Trained And Tested with Original and Generated Datasets (Dataset #2).

| | Detections:GT (~1) | FP:GT Ratio (~0) | TP:GT Ratio (~1) | % Detected Objects within five ranges of the confidence score | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | [0,0.2) | [0.2,0.4) | [0.4,0.6) | [0.6,0.8) | [0.8,1.0] |
| CenterMask (VoVNet-99) | 37.97 | 34.41 | 0.64 | 36.6 | 61.9 | 1.5 | 0 | 0 |
| CenterMask (VoVNet-57) | 39.03 | 34.98 | 0.91 | 29.6 | 65.1 | 4.8 | 0 | 0 |
| Faster R-CNN (ResNet-101) | 2.80 | 1.80 | 0.98 | 35 | 13 | 11 | 11.3 | 29.1 |
| Faster R-CNN (ResNet-50) | 2.93 | 2.07 | 0.87 | 36 | 17 | 8 | 9.4 | 29.9 |
| RetinaNet (ResNet-101) | 13.46 | 12.43 | **0.99** | 84.4 | 4.3 | 2.8 | 3.6 | 4.9 |
| RetinaNet (ResNet-50) | 15.32 | 14.59 | **0.99** | 86.2 | 4.9 | 2.6 | 2.4 | 3.9 |
| Mask R-CNN (ResNet-101) | 2.98 | 1.98 | 0.98 | 41 | 12 | 8 | 11.1 | 28.3 |
| Mask R-CNN (ResNet-50) | 2.69 | 1.76 | 0.97 | 36 | 14 | 7 | 11.2 | 32.2 |
| Mask R-CNN (ResNeXt-101) | **2.35** | **1.40** | 0.98 | 32 | 11 | 11 | 11.9 | 34.4 |

detection precision (mAP) and the ratio of the number of ground-truths and all detections (false positive and ground-truths) as presented in Fig. 7 by varying the confidence score. In a perfect case, both values should be equal to one. Hence, we select the threshold value from each model which lies in the line drawn from the origin (0,0) to (1,1). The score confidence thresholds 0.3, 0.6, 0.45, 0.5, and 0.6 are selected for CenterMask, Mask R-CNN (ResNet), Mask R-CNN (ResNeXt), Faster R-CNN and RetinaNet models, respectively.

The computational time relies on the complexity of the model to localize and identify the maturity level for each pineapple fruit. Fig. 8 (left) shows the inference time per image of each detection model vs. mAP. The larger size of layers, e.g., R101 (ResNet101), V99 (VoVNet99) compared with ResNet50, VoVNet59 always require more computational time. Mask R-CNN extends Faster R-CNN with the fully connected network for mask prediction which requires slightly more time. RetinaNet and Faster R-CNN have the highest computational speed, but a lot of false positives shown in RatinaNet as presented in Fig. 8 (right) in terms of the ratio between the ground-truths and all detected objects (false positives and true positives) which is ideally equal to one, so the ratio is low compared with the R-CNN models.

Figs. 9 and 10 compares the models before and after applied optimal score confidence thresholding. In Fig. 9, CenterMask is able to roughly localize the pineapple fruits but show many false positive bounding boxes similar to RetinaNet which is more accurate in localization. R-

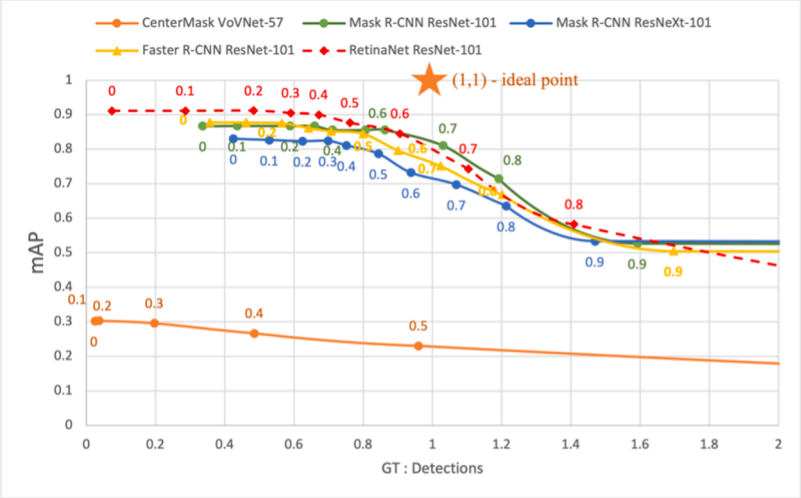


**Fig. 7.** mAP vs. GT:Detections varying score confidence threshold.

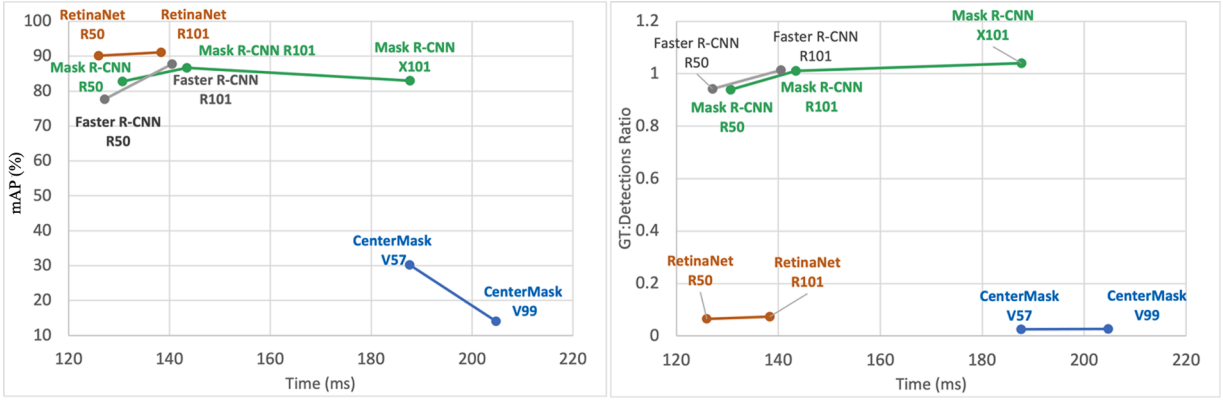


**Fig. 8.** Performance comparisons: mAP vs. Time (left), and a ratio between number of ground-truths and all detected objects vs. Time (right).

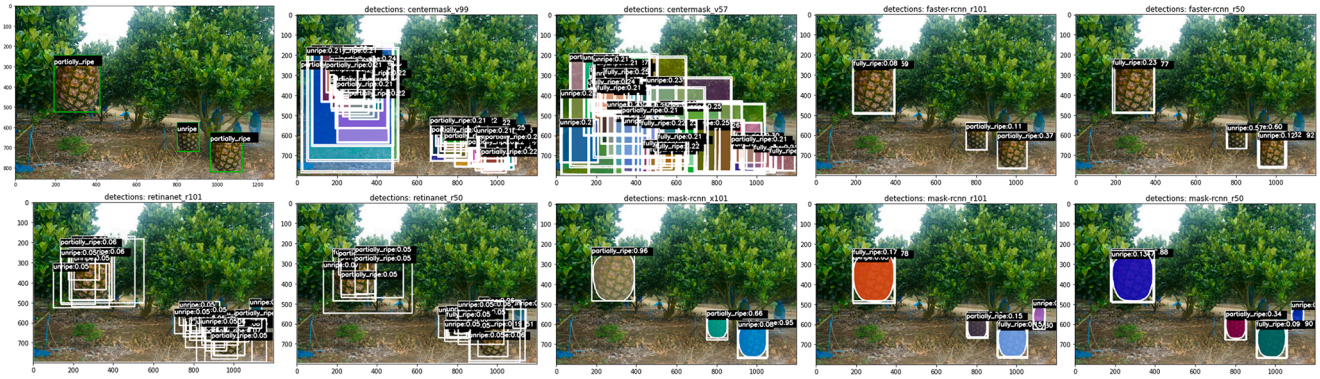


**Fig. 9.** Maturity level detection of the pineapple fruits (original image on the top left) without confidence score thresholding using CenterMask-V99, CenterMask-V59, Faster R-CNN-R101, Faster R-CNN-R50 from left to right, top row, and RetinaNet-R101, RetinaNet-R50, Mask R-CNN-R101, Mask R-CNN-R50, Mask R-CNN-X101 from left to right, bottom row.

CNN models show accurate detection and mask segmentation with few repeated bounding boxes with low score confidence. For Fig. 10 after applied empirically score confidence thresholding, all the models show more accurate results in terms of detection and maturity level classification.

Fig. 11 show many pineapples just harvested from the field. Most models focus on detecting clear and large objects, whereas Mask R-CNN-X101 and Faster R-CNN are able to detect some objects in the back of the scene and can cope with slightly overlapping objects. Figs. 12 and 13 compared the performance of the model trained with original images (one pineapple fruit in white background) and original images added with sampled multiple objects in plain and cluttered background), respectively. In this case, Mask R-CNN-R101 show perfect localization and classification from both models but with image augmentation, it shows higher score confidence. RetinaNet-R101 and CenterMask-V99 are unable to detect object within selected threshold due to lower

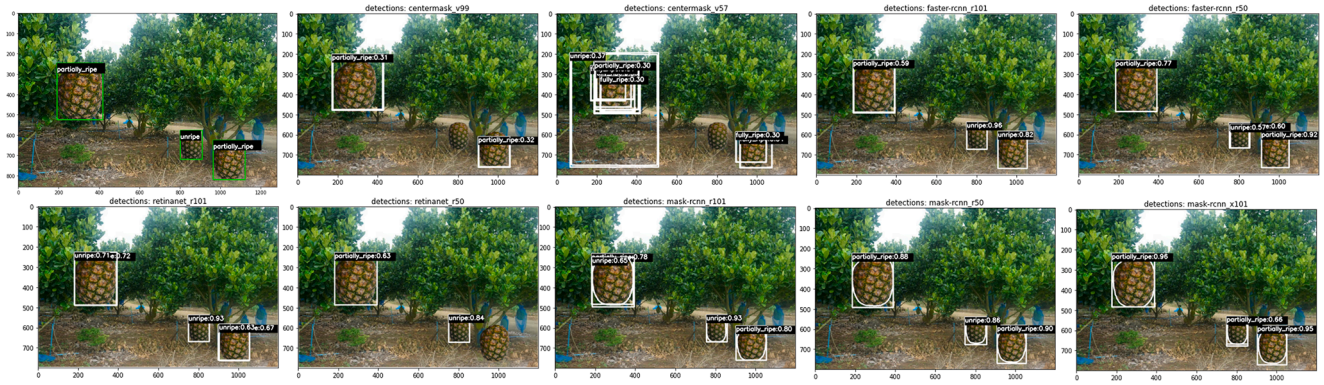


**Fig. 10.** Maturity level detection of the pineapple fruits (original image on the top left) with confidence score thresholding using CenterMask-V99, CenterMask-V59, Faster R-CNN-R101, Faster R-CNN-R50 from left to right, top row, and RetinaNet-R101, RetinaNet-R50, Mask R-CNN-R101, Mask R-CNN-R50, Mask R-CNN-X101 from left to right, bottom row.

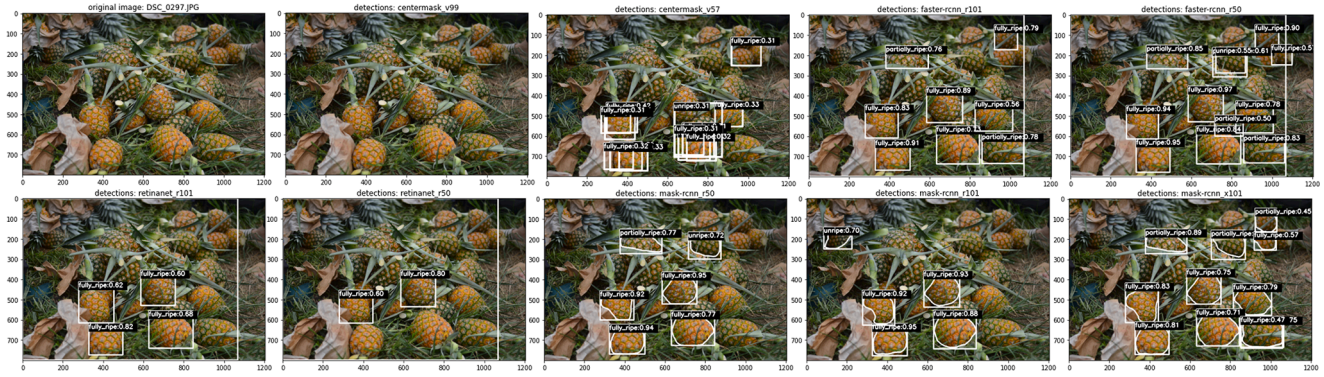


**Fig. 11.** Maturity level detection of the pineapple fruits (original image on the top left) with confidence score thresholding using CenterMask-V99, CenterMask-V59, Faster R-CNN-R101, Faster R-CNN-R50 from left to right, top row, and RetinaNet-R101, RetinaNet-R50, Mask R-CNN-R101, Mask R-CNN-R50, Mask R-CNN-X101 from left to right, bottom row.

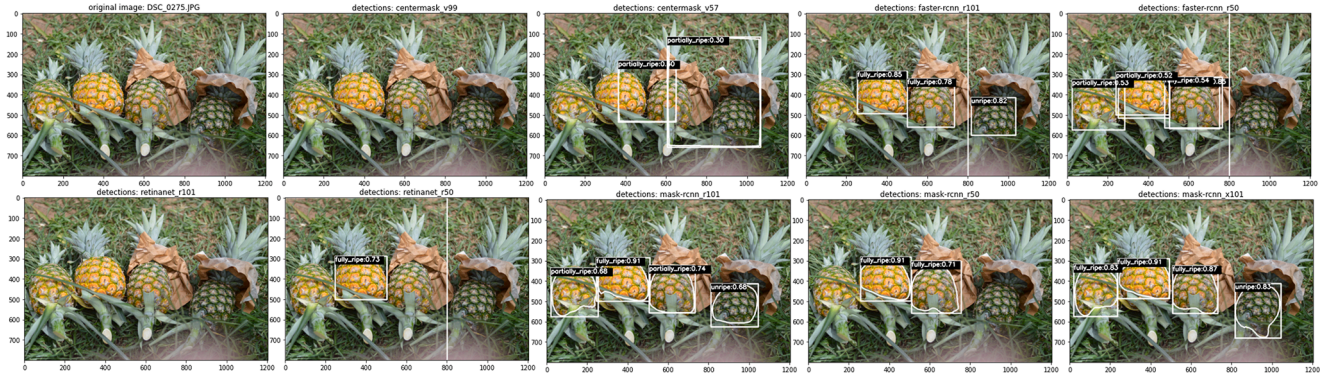


**Fig. 12.** Maturity level detection of the pineapple fruits (original image on the top left) with optimal confidence score threshold using CenterMask-V99, CenterMask-V59, Faster R-CNN-R101, Faster R-CNN-R50 from left to right, top row, and RetinaNet-R101, RetinaNet-R50, Mask R-CNN-R101, Mask R-CNN-R50, Mask R-CNN-X101 from left to right, bottom row. The model trained with constrained dataset (dataset #1).

detected score confidence. Faster R-CNN shows competitive performance in localization and classification similar to Mask R-CNN, only the object mask is not considered, so the confidence is slightly lower than Mask R-CNN.

## 5. Discussions

Previous work has developed ripeness identificaion of a pineapple fruits which are acquired under the constranined conditions, e.g., a fruit captured on a white background. With the sample size ~100-300 images, the applied techniques are limited to the consideration of the color distribution [5,6,10] and much research work [7,8] applied convolutional neural networks for extracting key freatures from the fruit and that mainly focused on image classification tasks as compared in Table 1. Our proposed framework has demonstrated to the ability to localize the pineapple fruit and identify the maturity level accurately, and the ability to identify multiply objects presented in the image with slight overlap and overcome background clutter problem with small dataset.

Our proposed augmentation technique by adding random multi-

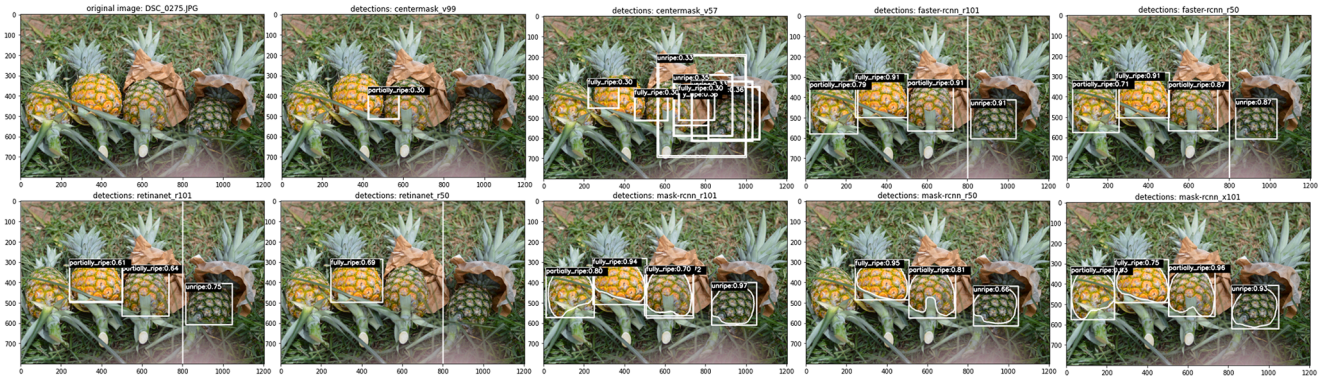


**Fig. 13.** Maturity level detection for each pineapple fruits (original image in the top left) with optimal confidence score threshold using CenterMask-V99, CenterMask-V59, Faster R-CNN-R101, Faster R-CNN-R50 from left to right, top row, and RetinaNet-R101, RetinaNet-R50, Mask R-CNN-R101, Mask R-CNN-R50, Mask R-CNN-X101 from left to right, bottom row. The model trained with original dataset and dataset with multi-object sampling.

object sampling, varying size and orientation along with the cluttered background are generated from the pixel-wise labelled dataset. This technique improves the detection precision and also increases the confidence scores of the detected objects.

The robustness of Mask R-CNN which is included in our framework is identified and compared with other state-of-the-art object detection and instance segmentation models, such as RetinaNet, Faster R-CNN, and CenterMask. RetinaNet has demonstrated the highest precision in detection, but there also exists so many false positives. The detection ratio of the ground-truths and the false positives is also evaluated. The confidence score ranges show a large number of low confidence score detected in CenterMask and RetinaNet. While R-CNN models show less false positives or low confidence score. Hence, the performance relies on the selected score confidence threshold for each model. We propose score optimal confidence selection techniques using mAP and detection ratio for select best threshold for each detection model. The results show great improvement for RetinaNet to remove false positives and repeated bounding boxes with low confidence score. Similarly, a lot of false positives are detected from CenterMask model which may require larger dataset.

In general, R-CNN models show reliable results, specifically Mask R-CNN has high confidence scores for detected objects. This concludes that object mask is required for reliable and accurate detection model. However, it shows the tradeoff between detection precision and computational time which needed to be considered for real-time practical applications where the transmission time or/and computed time in the limited computational resource might be additionally included.

## 6. Conclusion

This paper presents a detection and classification framework for identifying the pineapple ripeness stage using Mask R-CNN. The multi-object sampling techniques is proposed to generate a variety of sampling images from the original single object image with white background to cope with small dataset problem and also pixel-wise labelling is tedious work. In our framework, we employ Mask R-CNN, and also identify the robustness of the model compared with others. With small dataset, model can be sensitive to unseen data, we introduce the optimal thresholding technique based on mAP and detections ratio for Mask R-CNN. We demonstrate that the model developed with small dataset is robust and archives mean average precision of 86.7% with the ratio of ground truth and number of detections is nearly 0.9 (very close to one) and can be applied for multi-fruit grading with slightly overlapping under cluttered background in the field.

## Author contributions

Punnarai Siricharoen conducted the research; Warisa Yomsatieankul and Thidarat Bunsri provide conceptualization and ideas; all authors analyzed the data; Punnarai Siricharoen wrote the paper; all authors had approved the final version.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

The authors wish to thank Mr. Anon Rodthanong, Thailand's pineapple farmer association, for providing facilitation and supporting datasets for this research.

## References

[1] FAOSTAT, Production quantities of Pineapples by country, The Food and Agriculture Organization (2022). https://www.fao.org/faostat/en/#data/QCL/visualize.
[2] H. MF, F. Hossain, World pineapple production: an overview, Afr. J. Food, Agric. Nutr. Dev. 16 (4) (2016) 11443–11456, https://doi.org/10.18697/ajfand.76.15620.
[3] United Nations, Pineapple: United Nation Conference On Trade and Development, An INFOCOMM Commod. Profile New York (1) (2016) 1–22.
[4] S. Pathaveerat, A. Terdwongworakul, A. Phaungsombut, Multivariate data analysis for classification of pineapple maturity, J. Food Eng. 89 (2) (2008) 112–118, https://doi.org/10.1016/j.jfoodeng.2008.04.012.
[5] E.J.L. Aguilar, J.F. Villaverde, Determination of pineapple ripeness using support vector machine for Philippine standards, in: 7th International Conference on Control Science and Systems Engineering, 2021, pp. 283–287.
[6] E.R. Arboleda, C.L.T. de Jesus, L.M.S. Tia, Pineapple maturity classifier using image processing and fuzzy logic, IAES Int. J. Artif. Intell. 10 (4) (2021) 830–838, https://doi.org/10.11591/ijai.v10.i4.pp830-838.
[7] A. Chaikaew, T. Thanavanich, P. Duangtang, K. Sriwanna, W. Jaikhang, Convolutional neural network for pineapple ripeness classification machine, in: Proc. 16th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2019, 2019, pp. 373–376, https://doi.org/10.1109/ECTI-CON47248.2019.8955408.
[8] A.A. Azman, F.S. Ismail, Convolutional neural network for optimal Pineapple harvesting, J. Electr. Eng. 16 (2) (2017) 1–4, https://doi.org/10.11113/elektrika.v16n2.54.
[9] I.H. Kao, Y.W. Hsu, Y.Z. Yang, Y.L. Chen, Y.H. Lai, J.W. Perng, Determination of Lycopersicon maturity using convolutional autoencoders, Sci. Hortic.

(Amsterdam). 256 (May) (2019), 108538, https://doi.org/10.1016/j.scienta.2019.05.065.
[10] A.B. Badrul Adrul Hisham, I. Asnor Juraiza, S. Rosnah, W.H. Wan Zuha, Ripeness level classification for Pineapple, J. Theor. Appl. Inf. Technol. 57 (3) (2013) 587–593.
[11] A. Kamilaris, F.X. Prenafeta-Boldú, Deep learning in agriculture: a survey, Comput. Electron. Agric. 147 (July 2017) (2018) 70–90, https://doi.org/10.1016/j.compag.2018.02.016.
[12] S.K. Behera, A.K. Rath, P.K. Sethy, Maturity status classification of papaya fruits based on machine learning and transfer learning approach, Inf. Process. Agric. 8 (2) (2021) 244–250, https://doi.org/10.1016/j.inpa.2020.05.003.
[13] J. Cao, et al., An automated zizania quality grading method based on deep classification model, Comput. Electron. Agric. 183 (January) (2021) 1–8, https://doi.org/10.1016/j.compag.2021.106004.
[14] Herman, T. W. Cenggoro, D. Science, A. Susanto, B. Pardamean, and D. Science, Deep Learning for Oil Palm Fruit Ripeness Classification with DenseNet, (2021).
[15] S. Gonzalez, C. Arellano, J.E. Tapia, Deepblueberry: quantification of Blueberries in the wild using instance segmentation, IEEE Access 7 (2019) 105776–105788, https://doi.org/10.1109/ACCESS.2019.2933062.
[16] I. Pérez-Borrero, D. Marín-Santos, M.E. Gegúndez-Arias, E. Cortés-Ancos, A fast and accurate deep learning method for strawberry instance segmentation, Comput. Electron. Agric. 178 (February) (2020), 105736, https://doi.org/10.1016/j.compag.2020.105736.
[17] W. Jia, et al., FoveaMask: A fast and accurate deep learning model for green fruit instance segmentation, Comput. Electron. Agric. 191 (106488) (Dec. 2021), https://doi.org/10.1016/j.compag.2021.106488.
[18] H.C. Bazame, J.P. Molin, D. Althoff, M. Martello, Detection, classification, and mapping of coffee fruits during harvest with computer vision, Comput. Electron. Agric. 183 (March) (2021), https://doi.org/10.1016/j.compag.2021.106066.
[19] A. A. Kader, Pineapple: Maturity & Quality, (1996), https://postharvest.ucdavis.edu/Commodity_Resources/Fact_Sheets/Datastores/Fruit_English/?uid=50&ds=798 (Accessed June 18, 2022).
[20] P. Jund, N. Abdo, A. Eitel, and W. Burgard, The Freiburg Groceries Dataset, (2016), [Online]. Available: http://arxiv.org/abs/1611.05799.
[21] M. George, C. Floerkemeier, Recognizing products: A per-exemplar multi-label image classification approach, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 8690 LNCS (PART 2) (2014) 440–455, https://doi.org/10.1007/978-3-319-10605-2_29.
[22] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2014) 580–587, https://doi.org/10.1109/CVPR.2014.81.
[23] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 386–397, https://doi.org/10.1109/TPAMI.2018.2844175.
[24] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.
[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem (2016) 770–778, https://doi.org/10.1109/CVPR.2016.90.
[26] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017-Janua, 2017, pp. 5987–5995, https://doi.org/10.1109/CVPR.2017.634.
[27] S. Minaee, Y.Y. Boykov, F. Porikli, A.J. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: a survey, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1–22, https://doi.org/10.1109/TPAMI.2021.3059968.
[28] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 318–327, https://doi.org/10.1109/TPAMI.2018.2858826.
[29] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2. https://github.com/facebookresearch/detectron2.
[30] R. Padilla, W.L. Passos, T.L.B. Dias, S.L. Netto, E.A.B. Da Silva, A comparative analysis of object detection metrics with a companion open-source toolkit, Electron 10 (3) (2021) 1–28, https://doi.org/10.3390/electronics10030279.