Detecting and Localizing Strawberry Centers for Robotic Harvesting in Field Environment

Zixuan He*, Manoj Karkee, Qin Zhang*****

** Center for Precision and Automated Agricultural Systems, Department of Biological Systems Engineering, Washington State University, Prosser, WA 99350, USA (Tel: +1 509-339-4605; email: zixuan.he2@wsu.edu)*

*** Center for Precision and Automated Agricultural Systems, Department of Biological Systems Engineering, Washington State University, Prosser, WA 99350, USA (email: manoj.karkee@wsu.edu)*

**** Center for Precision and Automated Agricultural Systems, Department of Biological Systems Engineering, Washington State University, Prosser, WA 99350, USA (email: qinzhang@wsu.edu)*

Abstract: Automated or robotic harvesting methods are being investigated worldwide and have shown promising alternatives to manual harvesting in strawberry production. In robotic strawberry harvesting, the critical task of its machine vision system is to detect the presence and maturity of strawberries and estimate their precise location in the canopies. This study focused on the estimation and localization of strawberry centers in field environment to provide the 3D location of strawberry centers. It first applied a YOLOv4 approach to detect strawberries of different maturity levels (flower, immature, nearly mature, mature, and overripen) from an acquired RGB image. Matured strawberries detected by YOLOv4 were then used as inputs to a YOLOv4-tiny model to estimate berry centers in field conditions. A strawberry canopy dataset including 1300 selected RGB images was used for training the YOLOv4 model. Validation tests using 100 RGB images showed that the trained YOLOv4 model achieved an average precision (*AP*) of 91.73% in detecting mature strawberries at a reasonably high processing speed of 55.19ms. A dataset containing 750 images of single-strawberry was used in training the YOLOv4-tiny model. The trained model could detect the strawberry center in a processing time of 4.18ms per strawberry and achieved a mean average precision (*mAP*) of 86.45%. The average errors in estimating strawberry center locations were 1.65 cm on the x-axis, 1.53 cm on the y-axis, and 0.81 cm on the z-axis when the ZED camera was installed at ~100 cm. With precise detection of centers of strawberries by combining YOLOv4 and YOLOv4-tiny, the manipulator could receive accurate location information of strawberries to avoid inaccurate or failed picking during harvesting.

Keywords: machine vision, deep neural network, strawberry detection, object detection, depth image, YOLO

1. INTRODUCTION

Strawberry is one of the most widely cultivated small fruit globally due to its delicious taste and rich nutrition (Yu et al., 2019). Based on the U.S. Department of Agriculture report, the total crop value in the U.S. was over $2.2 billion in 2020 (USDA, 2021). Robotic harvesting technologies are being investigated in the strawberry industry to address labor-related challenges such as availability and cost. Thus, it is essential to develop a machine vision system that can provide precise locations of the strawberries in the field to support robotic harvesting.

Machine vision systems utilized in agricultural applications have generally improved in recent years regarding their accuracy and robustness. Classification, object detection, and segmentation methods based on deep learning techniques have been widely applied in the detection and grading of agricultural products (e.g., fruit recognition - Hussain et al., 2018; apple canopy segmentation - Xin et al., 2019). Convolutional Neural Networks (CNN) and different improvements, including Region-Based Convolutional Neural Networks (R-CNN), Fast RCNN, and Faster RCNN-based systems, have increasingly been used in detecting and localizing various types of fruit for robotic harvesting. That has helped robots accurately locate target fruits and estimate fruit maturity (Zhang et al., 2017; Gao et al., 2020).

One method for object detection studied extensively recently, You-Only-Look-Once (YOLO), could get detection results, including bounding boxes and class probability, directly with a single feed-forward network, making it computationally much more efficient than a two-stage network, such as RCNN-based networks (Redmon et al., 2016). YOLOv2 with the anchor was significantly improved on detection accuracy and the learning process from YOLO (Redmon et al., 2017). Modified YOLOv3 models were applied in fruit detection, and promising results were obtained with average precision over specific tasks (e.g., strawberry detection in Yu et al. (2020)). However, these studies have mainly focused on only RGB images and estimated only the 2D location of the detected objects in the images. The YOLOv4 achieved an *AP* of 43.5%

for the MS COCO dataset (Lin et al., 2014) and ~65 fps processing time on a Tesla V100 GPU (Bochkovsiy et al., 2020). He et al. (2021) also compared the results on field strawberry detection among YOLOv4, YOLOv3, and YOLOv2. The results showed that YOLOv4 performed better than the other two models in strawberry detection tasks. In this study, the YOLOv4 model was used to detect multiple classes of strawberries in the canopy images. As shown in Figure 1, the YOLOv4 includes a backbone (CSPDarknet 53), a neck (SPP and PAN), and a head (YOLOv3). CSPDarknet was used for feature extraction, an improved version of Darknet53 with better ability of gradient flow in the network. YOLOv4, with the additional structure of SPP and PAN, was not impacted on its speed while achieving improved performance in separating the target features.

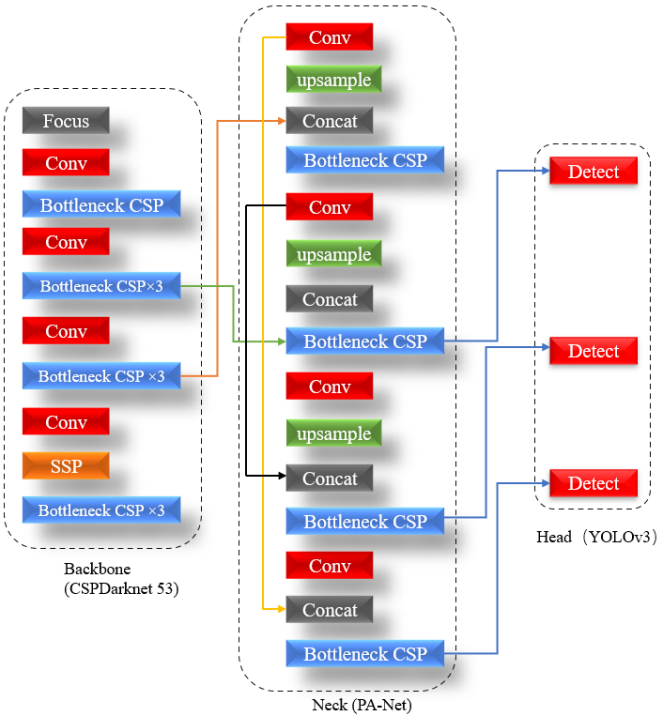

Figure 1. The generic architecture of a YOLOv4 model including backbone (CSPDarknet53), neck (SSP and PAN), and dense prediction block (YOLOv3).

There are, however, still limited studies in applying YOLOv4 for fruit detection with multiple maturities in field conditions (He et al., 2021). In addition, previous methods for finding picking points or the centers of strawberries were mainly based on regular image processing methods by following the color or shape of strawberries, which lacks robustness in the field environment. In most strawberry fields, a large proportion of fruit is in overlapped and occluded conditions. It is, therefore, essential to have a fast, robust, and reliable method to detect fully and partially visible strawberries and estimate their centers for picking. A deep-learning-based method (YOLOv4-tiny) was proposed in this study to estimate the center of mature strawberries after the YOLOv4 model detected mature strawberries.

The specific objectives of this study were to i) apply a trained YOLOv4 and YOLOv4-tiny to detect strawberries and center regions of mature strawberries; and ii) provide 3D location of target strawberries for robotic harvesting.

2. MATERIAL AND METHODS

2.1. Data Acquisition

The image data used to support this study was acquired from a commercial field located near Orlando, FL, between February 15 and 22, 2020. A ZED2 camera (*Stereolab inc., US*) was used for collecting RGB and depth images of strawberry canopies simultaneously in a natural open field environment. All images were collected from a fixed height of 100 cm above the strawberry beds. Around 200 strawberry canopy images were captured. Besides, around 2,000 RGB images (without depth information) of strawberry canopies were also collected using a ZED2 camera. An example color and depth image are shown in figure 2.

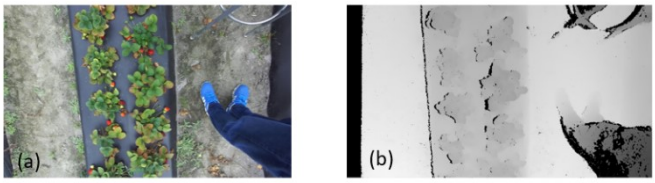

Figure 2. An example image of a strawberry canopy acquired using ZED2 camera; a) RGB image; and b) depth image.

2.2. Data processing

As RGB and depth images acquired with the ZED2 camera had the same resolution, an additional depth image could be obtained through the ZED2 camera (e.g., Figure 3) to directly get depth information after fruits were detected. A dataset was built with 1,400 selected strawberry canopy images to train the strawberry detection network. Strawberries were divided with five classes: flower, immature, nearly mature, mature, and overripen. Manual labeling of strawberries in these classes (except for the flower class) was based on the fruit grading method described by Barnes et al. (1976) as follows:

Immature group – Strawberries with green and white colors

Nearly mature group – Strawberries with red color on 1/4 to 3/4 of the surface area

Mature group – Strawberries with red color over 3/4 of the surface area

Overripe group – senescent strawberries with apparent corruption and withering

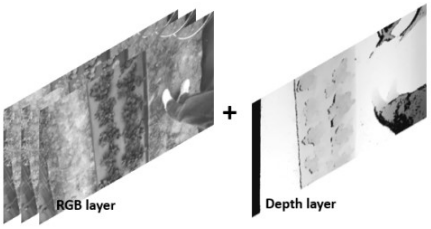

Figure 3. Example of pre-processed RGB image with corresponding depth image.

To detect strawberries and distinguish them to different maturity levels, all strawberries in the images were labeled with bounding boxes and classes, as shown in Figure 4. After the images were labeled, 850 images, each with a single strawberry of 'mature' level, were cropped from the canopy images according to the bounding boxes on the mature group. The fruit center regions were labeled manually according to the shape and location of mature strawberries in the original canopy images, as shown in Figure 5.

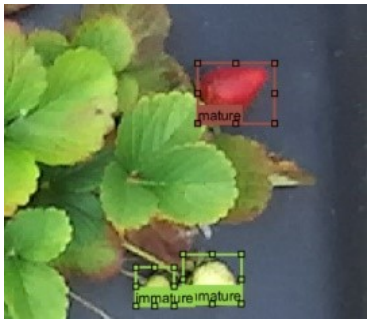

Figure 4. An example of labeled image of a strawberry canopy.

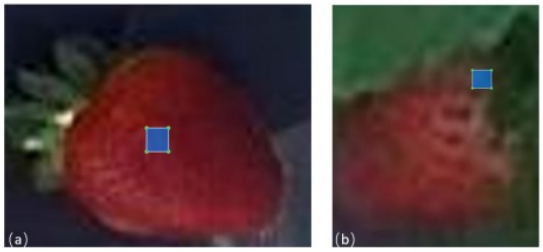

Figure 5. Examples of fruit centers labeling on detected strawberries.

2.3 YOLO-based object detection

In this study, a YOLOv4-tiny model was applied to detect the center of strawberries after being detected using a YOLOv4. Besides, we limited the output of YOLOv4-tiny to generate only one box with the highest scores/percentage on the center of strawberries. YOLOv4-tiny, a compressed version of YOLOv4, was designed for implementation in light computational platforms such as smartphones and single-board computers. Although the performance of YOLOv4-tiny was not better than YOLOv4 based on the COCO dataset, it was about 2–3 times faster than a YOLOv4 model.

Table 1 lists the training parameters of the YOLOv4 and YOLOv4-tiny models. As mentioned above, the YOLOv4 model was used for detecting strawberries of multiple classes. After a bounding box around a strawberry was generated by this model, such a bounding box was then input to the YOLOv4-tiny model for locating the fruit strawberry center. The input image sizes for YOLOv4 and YOLOv4-tiny were set as 648×726 and 416×416, respectively, to keep the main features of objects (strawberries with multiple maturities and center regions) in the images. Due to different output classes (YOLOv4 for 5 classes and YOLOv4-tiny for 1 class), the data batch size was set to 10,000 for YOLOv4 and 5,000 for YOLOv4-tiny. Besides, the filters before YOLO layers were 30 for YOLOv4 and 18 for YOLOv4-tiny. The learning rate, which decreased gradually during the training progress, was set to be 0.001 in the first 8,000 iterations, 0.0001 between 8,001 and 9,000 iterations, and 0.00001 between 9,001 and 10,000 iterations for YOLOv4. For YOLOv4-tiny, the learning rate was set to be 0.001 in the first 2,400 iterations, 0.0001 between 2,401 and 2,700 iterations, and 0.00001 between 2,701 and 3,000 iterations.

As mentioned before, the YOLOv4-tiny model was limited to generate only one output per image, as a strawberry could have only one center. The example of selection method is shown in Figure 6. In later processing, the bounding box with the highest proportion was kept while the others were deleted. After the center region of the bounding box was obtained, the center point could be generated by calculating the width and height of the bounding box and then matched with the depth image.

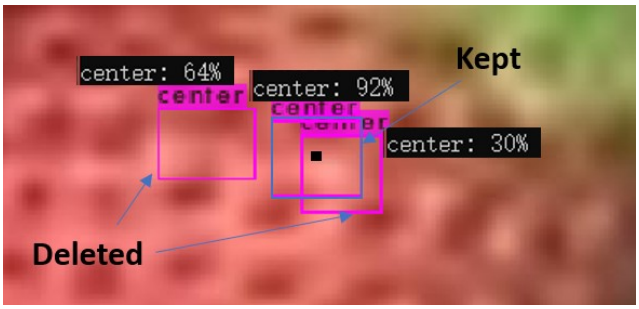

Figure 6. Example output of YOLOv4-tiny: selected bounding box (blue box) with the highest score during detection.

The YOLOv4 model was trained with 1,200 canopy images, whereas the YOLOv4-tiny was trained with 750 single-strawberry images (detected mature strawberries).

Table 1. Parameters of YOLOv4 and YOLOv4-tiny training

Parameter	YOLOv4	YOLOv4-tiny
Size of the input image	648×726	416×416
Subdivisions	64	64
Max training batch	10000	3000
Number of classes	5	1
Filter before each YOLO layers	30	18
Step	8000, 9000	2400, 2700
Number of Output	Not limited	1

2.4 Estimating the 3D location of strawberry centers

Once the deep-learning networks detected bounding boxes and the center of matured strawberries in an image, the pixel location of the fruit centers would be transformed into 3D coordinates using the depth layer appended to color images. Figure 7 presents a flowchart of the process used to estimate the 3D coordinates.

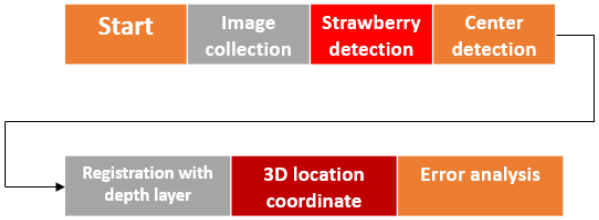

Figure 7. A flowchart for determining 3D location of strawberry centers using RGB and depth images.

For 3D calibration, three fake strawberries were placed on a strawberry canopy at a fixed distance from the ZED2 camera as calibrating objects, as shown in Figure 8, and the depth and RGB images of those calibrating objects were collected. The trained YOLOv4 model was used to detect these strawberries, with their centers obtained using the YOLOv4-tiny model. The depth information was then estimated using a co-registered depth layer. After the depth registration, the relative position in *X*-*Y* coordinates with reference to origin 'O' was estimated by using following equations:

$$X = k_1 x + b_1 \quad (1)$$

$$Y = k_2 x + b_2 \quad (2)$$

$$k_1 = (x_b - x_c)/(X_b - X_c) \quad (3)$$

$$k_2 = (y_a - y_c)/(Y_a - Y_c) \quad (4)$$

Where X and Y were coordinates of strawberry centers about the origin O; and x and y are the coordinates of strawberry centers (x, y) in RGB image. After calibration process is over, the relative location of strawberries and camera origin point O could be calculated when the pixel location was known from the YOLOv4 and the YOLOv4-tiny image processing tool.

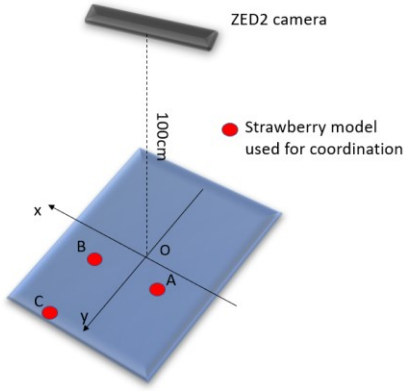

Figure 8. 3D calibration for estimating strawberry coordinates

2.5. Performance Assessment

Strawberry detection result was evaluated using average precision (*AP*) of mature strawberry, mean Average Precision (*mAP*) under an intersection-over-union (*IOU*) of 50%, processing speed, and the error (e) of 3D location on mature strawberries.

AP and *mAP* were calculated as follows:

$$AP = \sum_n (r_{n+1} - r_n) \max_{\tilde{r}: \tilde{r}^3 r_{n+1}} p(\tilde{r}) \quad (5)$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \quad (6)$$

$$p = TP/(TP + FP) \quad (7)$$

$$r = TP/(TP + FN) \quad (8)$$

$$F1\ score = 2 \times r \times p/(p + r) \quad (9)$$

Where p is precision, r is recall, TP is the number of true positive objects/strawberries detected, FP is the falsely detected objects/strawberries, and FN is the number of objects falsely not detected as strawberries.

The final location of strawberries was evaluated using *average errors (e)* by comparison between the calculated location and real location, which can be calculated as follows:

$$e_X = \frac{1}{N} \sum_{i=1}^{N} (X_i - X_{ri}) \quad (10)$$

$$e_Y = \frac{1}{N} \sum_{i=1}^{N} (Y_i - Y_{ri}) \quad (11)$$

$$e_z = \frac{1}{N} \sum_{i=1}^{N} (Z_i - Z_{ri}) \quad (12)$$

Where (X_i , Y_i , Z_i) is calculated location of mature strawberries, (X_r, Y_r, Z_r) is real location of strawberries. The data in (X_r , Y_r , Z_r) were acquired manually through measurement on the center point of center of strawberries to the origin point O (e.g., X_r cm to the origin point O).

3. Evaluation and Results

3.1. Performance on strawberry detection

The trained YOLOv4 model was evaluated qualitatively and quantitatively using a test dataset including 100 RGB images with a resolution of 1200×1000 pixels. An example detection result generated by YOLOv4 model was shown in Figure 9. The performance of YOLOv4 is shown in Table 2.

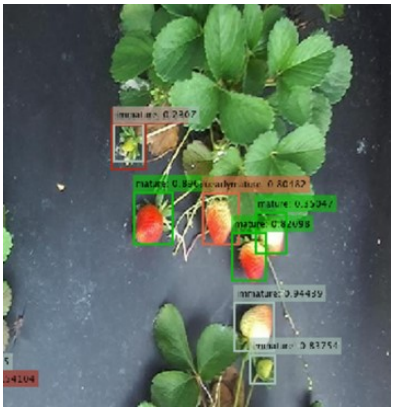

Figure 9. An example detection results generated by the YOLOv4 model trained with RGB images of strawberry canopies.

Table 2. Overall performance of YOLOv4 in strawberry detection

mAP (%)	*F1 score*	Processing time (ms)
80.68	0.80	55.19

Table 3. Model performance over individual fruit classes

AP (%)				
Flower	Immature	Nearly mature	Mature	Overripen
71.51	87.71	85.28	91.73	68.99

The results (Table 2) showed that the *mAP* on testing strawberry canopy dataset was 80.68%. The processing time for YOLOv4 per input image of 648×726 pixels was 55.19ms. Besides, additional results (Table 3) showed that AP in the flower group and overripen fruit group were 71.51% and 68.99%, respectively, which caused a decrease in *F1 score* and *mAP*. The features of flower and overripen groups were more complex than the other 3 groups, which was the main reason the trained YOLOv4 could not perform as well in recognizing flower and overripen groups. However, as the mature strawberries have outstanding appearances, detection of mature strawberries (which was the most crucial target for detection) achieved the highest *AP* of 91.73% on the test dataset. YOLOv4 also performed well in the immature and nearly mature groups, with *APs* of 87.71% and 85.28%, respectively. Overall, the trained YOLOv4 model showed a good ability to detect strawberries with different maturity levels, especially for the mature strawberries.

3.2. Performance of Strawberry Center Detection Model

The performance of the trained YOLOv4-tiny model in detecting centers of mature strawberries are shown in Table 4 and Figure 10. The average processing speed per image with a single mature strawberry (416×416 pixels) was only 4.18ms. The *mAP* achieved was 86.45%. In our dataset, one full canopy image consisted of up to 15 strawberries, meaning that the total processing time per strawberry image was less than 60 ms.

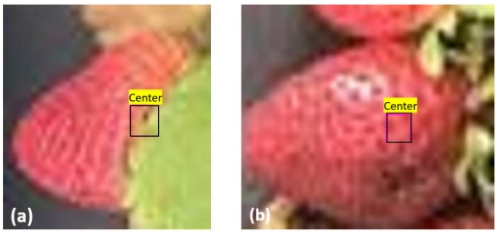

Figure 10. Detection examples on single strawberry.

Table 4. Performance of YOLOv4-tiny

Processing time(ms)	*mAP*(%)
4.18	86.45

3.3. Performance on 3D location of strawberries

The horizontal information of strawberries (X and Y) was estimated using equations (1) and (2), whereas the depth information (Z) was provided by the ZED2 camera. The estimated location (X_i, Y_i, Z_i) was compared against the real location (X_r, Y_r, Z_r). 50 calculated locations were recorded and compared with their real location. The errors on the x, y, and z axis were then estimated using equations (10) (11) and (12); the average errors on x, y, and z-axis are shown in Table 5.

Table 5. Errors in estimating X, Y, and Z coordinates of strawberry centers.

Axis	e (cm)
x	1.65
y	1.53
z	0.81

From Table 4, the average error on the z-axis was 0.81 cm while the average errors on the x-axis and y-axis were 1.65 cm and 1.53 cm, respectively with a fixed height (100 cm) of the ZED2 camera.

The error on the z-axis was mainly from the camera errors during obtaining depth images while the errors in x- and y-axis were mostly from the uneven ground in strawberry field, which resulted in ZED 2 camera not being strictly perpendicular to the strawberry bed. Finally, the real average error distance between calculated and the real locations was 1.7 cm in 3D space when the ZED2 camera put at a fixed height of 100 cm.

4. CONCLUSION

An in-field object detection method based on YOLOv4 and YOLOv4-tiny was developed for providing 3D location of strawberries in this study. The YOLOv4 model was trained using 1,300 RGB images and tested using 100 images with an input resolution of 648×726 pixels. The evaluation results showed that YOLOv4 method had a good potential to detect strawberries of different maturity levels with an *AP* of 91.73% and a short processing time of 55.19ms. The Strawberry center detection technique using YOLOv4-tiny model achieved a *mAP* of 86.45% with processing speed of 4.16ms on a single image of a 416×416-pixel resolution. The final location estimation technique achieved an average error of less than 2 cm at a fixed camera height of 100 cm. This study focused on strawberry detection on RGB and depth images under a field environment, and post-processing method for improving the accuracy and efficiency of the machine vision system. In conclusion, this deep learning-based method could be used to efficiently detect and locate strawberries and their centers under field conditions for guiding robotic arm to reach the target fruit, which is crucial for automated/robotic, targeted strawberry harvesting.

Further improvement is essential in decreasing the processing time of deep learning-based models for real-time applications (e.g., the modified structure of YOLOv4). Additionally,

increasing the total number of images in the training and testing dataset could be an excellent way to improve the performance of the model. Furthermore, other object detection models, including YOLOv5 and YOLOx (the latest two versions of YOLO) could be tested and applied for strawberry detection.

ACKNOWLEDGMENT

This research was supported in part by the National Science Foundation (NSF; award# 1924640), and Washington State University (WSU). The first author was also supported by the China Scholarship Council (CSC). Any options, findings, and conclusions expressed in this publication are those of the authors and do not reflect any view from NSF, WSU, or CSC. The authors would like to acknowledge ArcLab at the University of Central Florida for their support during the field data collection.

References

Barnes, M. F., & PATCHETT, B. J. (1976). Cell wall degrading enzymes and the softening of senescent strawberry fruit. *Journal of food science, 41*, 1392–1395.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Gao, F., Fu, L., Zhang, X., Majeed, Y., Li, R., Karkee, M., & Zhang, Q. (2020). Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. *Computers and Electronics in Agriculture, 176*, 105634.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence, 37*, 1904–1916.

He, Z., Karkee, M., & Upadhayay, P. (2021). Detection of strawberries with varying maturity levels for robotic harvesting using YOLOv4. *2021 ASABE Annual International Virtual Meeting*, (pág. 1).

Hussain, I., He, Q., & Chen, Z. (2018). Automatic fruit recognition based on dcnn for commercial source trace system. *Int. J. Comput. Sci. Appl, 8*, 01–14.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision*, (págs. 740–755).

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (págs. 8759–8768).

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (págs. 7263–7271).

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (págs. 779–788).

USDA.National agricultural statistics database. (2021). Obtenido de https://www.nass.usda.gov/

Yu, Y., Zhang, K., Yang, L., & Zhang, D. (2019). Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Computers and Electronics in Agriculture, 163*, 104846.

Zhang, J., He, L., Karkee, M., Zhang, Q., Zhang, X., & Gao, Z. (2017). Branch detection with apple trees trained in fruiting wall architecture using stereo vision and regions-convolutional neural network (R-CNN). *2017 ASABE annual international meeting*, (pág. 1).

Zhang, X., Fu, L., Karkee, M., Whiting, M. D., & Zhang, Q. (2019). Canopy segmentation using ResNet for mechanical harvesting of apples. *IFAC-PapersOnLine, 52*, 300–305.