Fruit Ripeness Detector for Automatic Fruit Classification Systems

Duy-Linh Nguyen[1], Xuan-Thuy Vo[2], Adri Priadana[3], Muhamad Dwisnanto Putro[4], and Kang-Hyun Jo[5]
[1,2,3,5]*Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, Korea*
[4]*Department of Electrical Engineering, Sam Ratulangi University, Manado, Indonesia*
Email: ndlinh301@mail.ulsan.ac.kr; xthuy@islab.ulsan.ac.kr; priadana@mail.ulsan.ac.kr; dwisnantoputro@unsrat.ac.id; acejo@ulsan.ac.kr

Abstract—**The development of Artificial Intelligence has led to remarkable advancement in agriculture. Many automatic tools have been developed to reduce human labor and improve accuracy. One of the most popular applications in harvesting and packaging agricultural products is the fruit classification system based on ripeness level. This paper focuses on improving the YOLOv8n architecture by replacing the original convolution operations with a new convolution module called the Receptive Field Convolution Block Attention Module for fruit ripeness detection. This module leverages the advantages of group convolution and Convolution Block Attention Module mechanisms to enhance the feature extraction ability. The experiments are trained and evaluated on the Fruit Ripening Process and Mango And Banana datasets. As a result, the proposed network achieves the best performance at 99.4% of mAP@0.5 and demonstrates superiority over other methods under the same experimental conditions.**

Index Terms—**CBAM, convolutional neural network, fruit classification systems, fruit ripeness detection, YOLOv8.**

I. Introduction

According to statistics from the Food and Agriculture Organization of the United Nations (FAO), in 2021 world fruit production reached about 910 million tons and this number may change from year to year depending on natural conditions [1]. Among them, the top-producing countries are China, USA India, Mexico, etc. With such a large annual production, it requires quick and accurate classification and packaging before delivering to the market. These tasks are essential to ensure the best quality of fruit. However, classifying fruits with different kinds and ripeness levels is not an easy task because of similarities in color, shape, and size [2]. Typically, this stage will be conducted by a group of experts or trained people with the main factors to evaluate being the color and quality of the fruit product. Manual testing often generates many errors depending on each individual's experience and judgment. Therefore, the quality of classification is not as uniform as expected. Recently, many researchers have focused on the agricultural field and provided a variety of solutions and applications for fruit detection and classification. The techniques aim to partly address complex challenges such as diversity, unevenness, and inconsistency in shape, color, and texture [3]. Prominent among them are those Computer Vision-based methods. These approaches take advantage of machine learning algorithms, especially Convolutional Neural Networks (CNNs), to distinguish the inherent characteristics of fruits. Following that trend, this paper proposes a technique to improve the YOLOv8n network architecture for fruit ripeness classification by carefully analyzing the original network architecture and completely replacing standard convolutions with Receptive Field Convolution Block Attention Module Convolution (RFCBAMConv) inside the backbone and neck modules. With the combination of lightweight architectures, CBAM attention mechanism, and computational complexity optimization, the proposed network has high accuracy and the potential to be deployed in low-computing devices for automatic fruit ripeness classification systems.

The paper provides several main contributions as follows:
1 - Proposes an efficient fruit ripeness detector based on YOLOv8n architecture for automatic fruit classification systems.
2 - The proposed method achieves better performance than other approaches on the Fruit Ripening Process and Mango And Banana datasets.

The remaining parts of the paper are arranged as follows: Section II presents related work to fruit ripeness detection and recognition. Section III introduces the proposed technique in detail. Section IV analyzes and assesses the experimental results. Section V concludes the issue and future development direction.

II. related work

This section will introduce related work to fruit ripeness detection and classification. They can be separated into traditional machine learning methodologies and CNN-based methodologies.

A. Traditional machine learning methodologies

To detect and recognize fruit ripeness, traditional machine learning methods often apply spectral and hyperspectral analysis methods. The authors in [4] proposed an automatic method to distinguish the ripeness of bananas using spectral and RGB data. It used classifiers such as random forests, multilayer perceptrons, and feedforward neural networks to classify spectral data. In [5], hyperspectral reflectance images were used to evaluate and classify three common peach diseases by analyzing spectral and image information. A study in [6] provided hyperspectral systems to combine spectral information on each pixel for fruit and vegetable quality assessment.

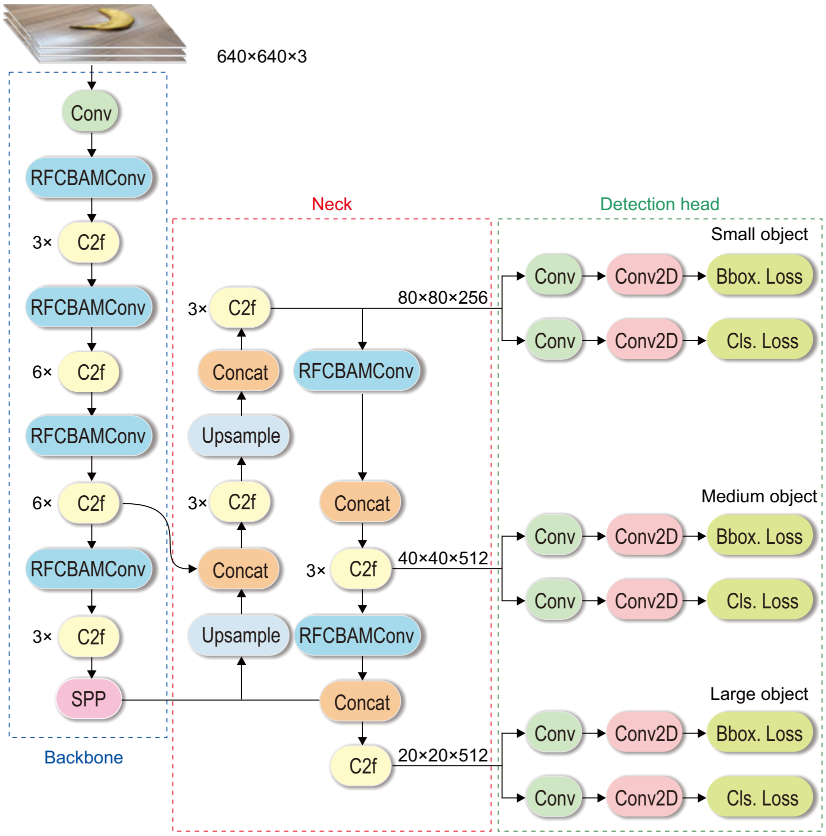


Fig. 1. The overview of the proposed fruit ripeness detection network.

These traditional machine learning methods generally achieve quite good accuracy but have high computational cost and implementation complexity. These factors hinder deployment in real-time applications.

B. *CNN-based methodologies*

An improved MobileNetV2 [7] network based on pre-trained weights from ImageNet was used to classify six fruit classes. In another study in [8], AlexNet, ResNet50, and VGG-16 were also used to classify the above six kinds of fruit. The work in [9] improved the U-Net model to detect rotten or fresh apples from peel defects. The research in [10] shows that the YOLO (You Only Look Once) network structures can automate different tasks on various fruit datasets, bringing more effective applications in the agricultural automation process. The advantage of these methodologies is their high speed and accuracy. However, the main drawback is that it requires high hardware specifications and expensive supporting devices.

III. Methodology

The fruit ripeness detection network overview is described in detail as shown in Fig. 1. This is an improved YOLOv8 architecture that consists of three parts: Backbone, Neck, and Detection head.

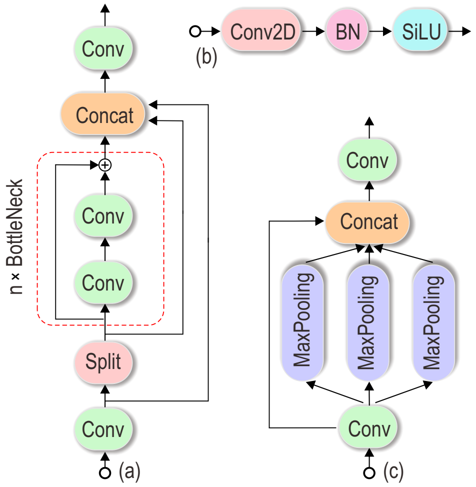


Fig. 2. The architecture of Cross Stage Partial Fast BottleNeck (C2f) (b), Conv (b), and Spatial Pyramid Pooling (SPP) (c) blocks.

A. *Proposed network architecture*

This work thoroughly analyzes and evaluates each component in the original YOLOV8 architecture [11]. From those analytic results, the research focuses on refining several blocks in the Backbone and Neck modules. Specifically, the Cross

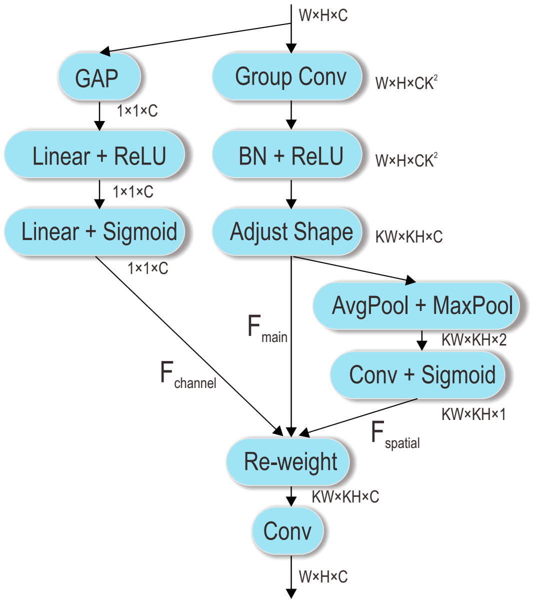


Fig. 3. The Receptive Field Convolution Block Attention Module architecture.

Stage Partial Bottleneck with two convolutions (C2f) block is reused, the Spatial Pyramid Pooling Fast (SPPF) and Conv blocks are replaced by the Spatial Pyramid Pooling (SPP) and the Receptive Field Convolution Block Attention Module Convolution (RFBAMConv) [12], respectively. The architecture details of Conv, C2f, and SPP blocks are shown in Fig. 2.

The Backbone module is redesigned based on a Conv block, followed by four identical combination blocks, and ending with an SPP block. In which, the same intermediate blocks are sequentially stacked with several C2f blocks (with repetition ratios of 3, 6, 6, and 3 times) and an RFCBAMConv block. Fig. 3 shows the architecture of RFCBAMConv which is a combination of the Receptive Field Convolution Block Attention Module (RFBAM) mechanism and a standard convolution (Conv2D). Besides, the Channel Attention Module (CAM) is replaced by the Squeeze-Excitation Attention Module (SE). The RFCBAMConv is designed to address the convolution kernel parameters sharing problem and improve the feature extraction ability for standard convolution operation. This RFCBAMConv block conducts the group convolutions which can save a large number of parameters. On the other hand, the CBAM attention mechanism guides the network to learn the important information on each feature map level. Assume that, $F \in R^{W \times H \times C}$ is the input feature map and $F' \in R^{W \times H \times C}$ is the output feature map. The operating process of RFCBAMConv can be described as follows:

$$F' = f^{3 \times 3}(F_{Channel} \times F_{Main} \times F_{Spatial}), \tag{1}$$

where $f^{3 \times 3}$ is the standard convolution operation with kernel size 3×3. $F_{Channel}, F_{Main}$, and $F_{Spatial}$ are the output feature maps of channel attention, main, and spatial attention branches, respectively. $F_{Channel}, F_{Main}$, and $F_{Spatial}$ are computed as below equations:

$$F_{Channel} = \sigma(FC(ReLU(FC(GAP(F))))), \tag{2}$$

$$F_{Main} = Reshape(BN(ReLU(g^{3 \times 3}(F))), \tag{3}$$

$$F_{Spatial} = \sigma(f^{1 \times 1}([Avg(F_M), Max(F_M)])), \tag{4}$$

in which, GAP is the Global Average Pooling layer. FC is the fully connected layer. $ReLU$ is the Rectified Linear Unit activation function. Avg and Max are Average Pooling and Max Pooling layers, respectively. The operation $[\cdot]$ describes the Concatenation layer. $f^{1 \times 1}$ denotes the standard convolution operation with kernel size 1×1. The symbol σ stands for the Sigmoid activation function. At the end of the backbone module, this work uses the SPP block from YOLOv5 [13] to replace the SPPF block. The kernel size of the Max Pooling layers varies from 3×3 to 5×5 to ensure reasonable network parameters.

The Neck module leverages the Path Aggregation Network (PAN) architecture as in the YOLOv8 network and also replaces the whole of the Conv blocks with the RFCBAMConv blocks. This module upsamples the current feature maps and aggregates them with previous low-level feature maps from the backbone module using Concatenation operations. The three scale output feature maps corresponding to the three scales of the object (small, medium, and large) are generated by the Neck module. Three feature maps have enriched the important information and go through the detection head module.

The detection head module also reuses the architecture of three detection heads from the original YOLOv8 with the decouple head and free-anchor technique. The feature maps from the output of the Neck module go to two siblings of a combination of a Conv block and standard convolution for bounding box regression (four coordinates of the box: x, y, h, w) and classification (number of classes: c) on three object scales. The Conv block is described in Fig. 2 (b). This block uses a 1×1 standard convolution layer (Conv2D), a BN, and a ReLU activation function. The Conv blocks are only used in the detection head module of the proposed network. Table 1 presents the detection head module in detail.

TABLE I
THE DETAILS OF THE DETECTION HEAD MODULE.

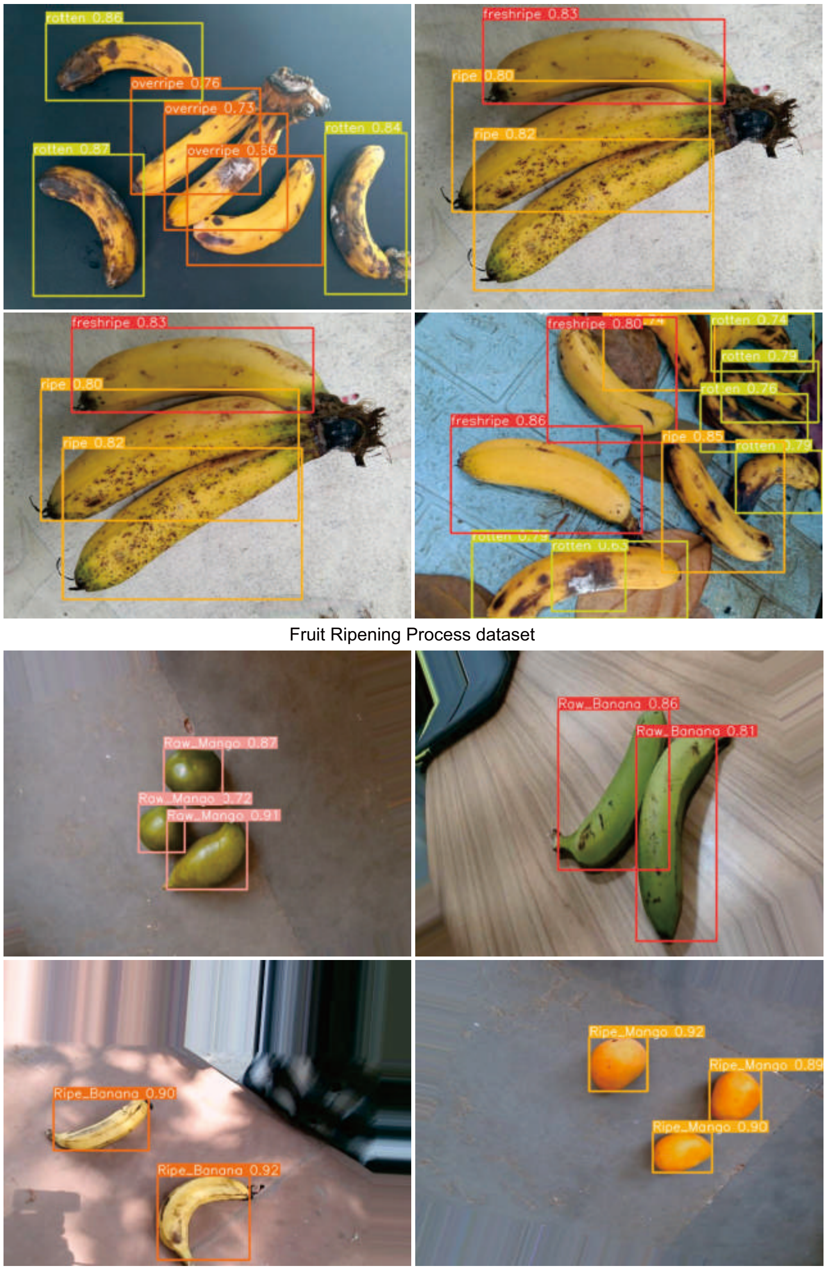
Heads	Input	Anchor	Ouput	Object
1	$80 \times 80 \times 256$	Free	$80 \times 80 \times 4/80 \times 80 \times 2$	Small
2	$40 \times 40 \times 512$	Free	$40 \times 40 \times 4/40 \times 40 \times 2$	Medium
3	$20 \times 20 \times 512$	Free	$20 \times 20 \times 4/20 \times 20 \times 2$	Large

B. *Loss function*

The proposed network using the loss function is defined as follows:

$$L = \lambda_{Box} L_{Box} + \lambda_{DFL} L_{DFL} + \lambda_{Cls} L_{Cls}, \tag{5}$$

where the bounding box regression loss combines L_{Box} and L_{DFL} and applies the CIoU loss and Distribution Focal Loss (DFL), respectively. The classification loss L_{Cls} uses the Binary Cross Entropy loss to compute. The λ_{Box}, λ_{Cls}, and λ_{dfl} are balancing parameters.



Fruit Ripening Process dataset

Mango and Banana dataset

Fig. 4. The qualitative result of the proposed network on the validation set of the Fruit Ripening Process Dataset and the Mango And Banana Dataset.

IV. Experimental Results

A. Dataset

The experiments conduct training and evaluation on the Fruit Ripening Process Dataset and the Mango And Banana Dataset. The Fruit Ripening Process Dataset [14] is collected from the Roboflow website. This dataset includes 6,789 banana images with ripeness levels (fresh ripe, fresh unripe, overripe, ripe, rotten, unripe) selected from many different sources. The dataset is divided into two subsets with 5,264 images for training and 1,525 images for evaluation. The Mango And Banana Dataset [15] is an RGB image dataset. This dataset contains 5,000 images with 640×480 resolution of bananas and mangoes focusing on classifying ripe and unripe fruits tasks. The data set is split into a training set and an evaluation set at the rate of 80% (4,000 images) and 20% (1,000 images).

B. Experimental setup

The proposed network is built using the Python programming language and the Pytorch framework. The experiments are trained and evaluated on a GeForce GTX 1080Ti 11GB GPU. The training phase applies the Stochastic Gradient

TABLE II

THE COMPARISON RESULTS OF THE PROPOSED NETWORK WITH OTHER METHODS ON THE FRUIT RIPENING PROCESS AND MANGO AND BANANA DATASETS.

Method	Parameter	GFLOPs	Weight (MB)	Fruit Ripening Process			Mango and Banana		
				mAP1	mAP2	Inf. (ms)	mAP1	mAP2	Inf. (ms)
YOLOv5n	1,769,329	4.2	3.8	91.6	64.1	1.4	99.3	83.6	1.4
YOLOv8n	3,006,428	8.1	6.2	92.8	69.9	0.6	99.2	86.6	0.6
YOLOv8s-RFCBAMConv	11,243,112	29.0	22.8	93.4	70.6	2.3	99.3	86.8	4.2
YOLOv8m-RFCBAMConv	34,109,656	86.1	68.6	93.1	70.6	4.1	99.4	86.8	6.8
YOLOv8l-RFCBAMConv	76,910,856	192.9	154.4	93.2	71.1	6.4	99.4	87.0	7.8
YOLOv8x-RFCBAMConv	120087934	301.0	240.8	93.1	70.8	9.4	93.0	70.8	9.4
Proposed mothod	**3,064,472**	**8.3**	**6.4**	**93.8**	**70.3**	**1.9**	**99.4**	**87.1**	**1.1**

- **mAP1**: mAP@0.5 (%).
- **mAP2**: mAP0.5:0.95 (%).
- **Inf.**: Inference time (ms) is evaluated on a GeForce GTX 1080Ti GPU.
- **Red color**: Best competitor.

Descent (SGD) optimization. The initial learning rate is set at 10^{-2} and ends at 10^{-4}. The momentum is set at 0.937. The training process uses 200 epochs with a batch size of 16. The balance parameters are set as follows: λ_{Box}=1.5, λ_{Cls}=0.5, and λ_{DFL}=1.5. To enhance the training dataset and avoid overfitting problems, several data augmentation methods (such as mosaic, translate, scale, and flip) are used. In the inference phase, argument configurations are set as an image size of 640×640, a batch size of 16, a confidence threshold = 0.5, and an IoU threshold = 0.5. The inference time is reported in milliseconds (ms).

C. Experimental results

To evaluate the performance of the proposed network, this work retrains from scratch the nono architectures of YOLOv5 and YOLOv8 (nano version). On the other hand, the study also conducts the same with different versions of the proposed network from small to extra-large versions (s, m, l, x). As a result, the proposed network achieves 93.8% of mAP@0.5, 70.3% of mAP@0.5:0.95 and 99.4% of mAP @0.5, 87.1% of mAP@0.5:0.95 respectively on the two datasets mentioned above. From the comparison results in Table 2, it is seen that the proposed network outperforms most of the competitors. More specifically, for the Fruit Ripening Process Dataset the performance of the proposed network is comparable to the large version YOLOv8l-RFCBAMConv (0.8%↓) and better than YOLOv8s-RFCBAMConv (0.4%↑). Meanwhile, the parameter and GFLOPs are equivalent to YOLOv8n and speed is the same as YOLOv5n (0.5 ms↓). For the Mango And Banana Datase, the performance of the proposed network is better than YOLOv8l-RFCBAMConv (0.1%↑) and better speed than YOLOv5n (0.3 ms↓). These high performances promise the proposed network to be deployed on low-computing devices for in real-time fruit ripeness classification systems. Several qualitative results on two datasets are shown in Fig. 4. The comparison results between the proposed network and YOLOv8n are clearly shown in Fig. 5. From this visualization result, it can be seen that the proposed network has better object detection ability in obscured situations, similar colors between ripeness levels context, and rotten parts inside the fruit. However, because several fruit ripeness levels are relatively close in color, detecting and differentiating them remains a huge challenge. Therefore, it requires the development of powerful object detectors and rich datasets to optimize the performance of automatic fruit ripeness classification systems.

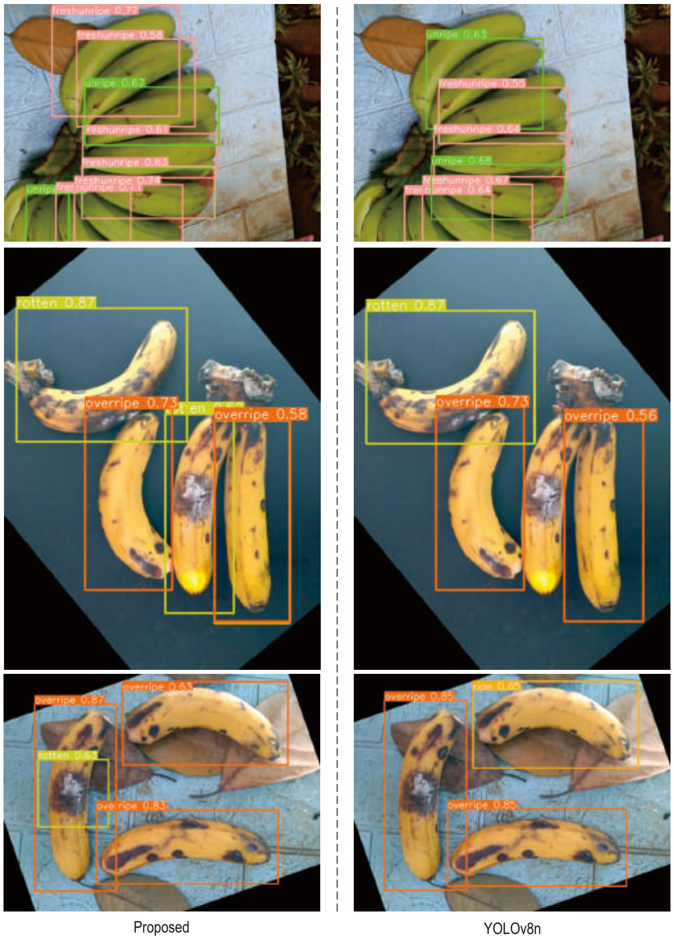


Fig. 5. The comparison result between the proposed and YOLOv8n networks on the validation set of the Fruit Ripening Process Dataset.

D. Ablation study

This research also assesses the effectiveness of each proposed module through several ablation studies. To build the different versions of the proposed network, these modules are replaced one by one, and then conducts the training and evaluation processes on the Mango And Banana dataset. The experimental results in Table III demonstrate that using the first Conv block can significantly reduce network parameters and computational complexity while the network still ensures accuracy and increases processing speed. On the other hand, replacing the SPPF block with the SPP block also increases detection accuracy and inference time. From those experiments, this study chose the combination of Conv, RFCBAMConv, and SPP blocks to improve the YOLOv8n network to achieve the best performance.

TABLE III
ABLATION STUDIES WITH DIFFERENT PROPOSED NETWORKS ON THE VALIDATION SET OF THE MANGO AND BANANA DATASET.

Blocks	**Proposed backbones**			
First Conv	✓			✓
RFCBAMConv	✓	✓	✓	✓
SPPF	✓	✓		
SPP			✓	✓
Parameter	3,064,472	3,064,915	3,064,915	3,064,472
GFLOPs	8.3	8.4	8.4	8.3
Weight (MB)	6.4	6.4	6.4	6.4
mAP@0.5	99.4	99.4	99.4	**99.4**
mAP@0.5:0.95	86.6	86.6	86.7	**87.1**
Inf. time (ms)	1.1	3.1	2.7	**1.1**

V. CONCLUSION AND FUTURE WORK

This paper proposes a technique to improve the YOLOv8n architecture for fruit ripeness detection supporting the automatic fruit ripeness classification systems. The proposed method is composed of three parts: backbone, neck, and detection head modules. The backbone and neck modules are redesigned by replacing the Conv blocks with the RFAConv blocks except the first Conv block. Besides, the SPPF block is also replaced by the SPP block with the small kernel sizes of the Max pooling layer. The detection head leverages the idea from the original architecture in YOLOv8n. The proposed network achieves the best mAP at 99.4% and is comparable to existing methods. The optimization of the number of parameters, computational complexity, inferent time, and detection precision provides the promise to deploy on real-time systems. In future work, this research will try to implement the experiment on a larger fruit dataset and compare the performance to YOLOv9.

ACKNOWLEDGEMENT

This result was supported by the "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE)(2021RIS-003).

REFERENCES

[1] FAO, "Agricultural production statistics 2000–2021," in *FAOSTAT analytical briefs*, no. 60, 2022.

[2] J. Steinbrener, K. Posch, and R. Leitner, "Hyperspectral fruit and vegetable classification using convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 162, pp. 364–372, 2019.

[3] K. Hameed, D. Chai, and A. Rassau, "A comprehensive review of fruit and vegetable classification techniques," *Image and Vision Computing*, vol. 80, pp. 24–44, 2018.

[4] B. Mithun, S. Shinde, K. Bhavsar, A. Chowdhury, S. Mukhopadhyay, K. Gupta, B. Bhowmick, and S. Kimbahune, "Non-destructive method to detect artificially ripened banana using hyperspectral sensing and rgb imaging," in *Sensing for agriculture and food quality and safety X*, vol. 10665, pp. 122–130, SPIE, 2018.

[5] Y. Sun, K. Wei, Q. Liu, L. Pan, and K. Tu, "Classification and discrimination of different fungal diseases of three infection levels on peaches using hyperspectral reflectance imaging analysis," *Sensors*, vol. 18, no. 4, p. 1295, 2018.

[6] D. Lorente, N. Aleixos, J. Gómez-Sanchís, S. Cubero, O. García-Navarrete, and J. Blasco, "Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment," *Food and Bioprocess Technology*, vol. 5, pp. 1121–1142, 05 2011.

[7] T. Ananthanarayana, R. W. Ptucha, and S. C. Kelly, "Deep learning based fruit freshness classification and detection with cmos image sensors and edge processors," in *Food and Agricultural Imaging Systems*, 2020.

[8] M.-C. Chen, Y.-T. Cheng, and C.-Y. Liu, "Implementation of a fruit quality classification application using an artificial intelligence algorithm," *Sens. Mater.*, vol. 34, no. 1, pp. 151–162, 2022.

[9] K. Roy, S. Chaudhuri, and S. Pramanik, "Deep learning based real-time industrial framework for rotten and fresh fruit detection using semantic segmentation," *Microsyst Technol*, vol. 27, pp. 3365–3375, 2021.

[10] A. Zargham, I. U. Haq, T. Alshloul, S. Riaz, G. Husnain, M. Assam, Y. Y. Ghadi, and H. G. Mohamed, "Revolutionizing small-scale retail: Introducing an intelligent iot-based scale for efficient fruits and vegetables shops," *Applied Sciences*, vol. 13, no. 14, 2023.

[11] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.

[12] X. Zhang, C. Liu, D. Yang, T. Song, Y. Ye, K. Li, and Y. Song, "Rfaconv: Innovating spatial attention and standard convolutional operation," 2023.

[13] G. Jocher and et al., "ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements," Oct. 2020.

[14] F. Ripening, "Fruit ripening process dataset." https://universe.roboflow.com/fruit-ripening/fruit-ripening-process , oct 2022. visited on 2024-02-26.

[15] A. Sutar, A. Naikare, P. Jadhav, and R. Kute, "Mango and Banana Dataset (Ripe Unripe) : Indian RGB image datasets for YOLO object detection." https://data.mendeley.com/datasets/y3649cmgg6/3, 2023. Mendeley Data, Version 3.