DeeplabV3+ Model with CBAM and CSPM Attention Mechanism for Navel Orange Defects Segmentation

Guo Jinmei[1], Wan Nurshazwani Wan Zakaria[2]*, Wei Bisheng[3], Muhammad Azmi Bin Ayub[4]
College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia[1, 2, 4]
School of Mechanical and Electronic Engineering, Jiangxi College of Applied Technology, 341000 Ganzhou, China[1, 3]

***Abstract*—Accurate defect detection of navel oranges is the key to ensuring the quality of navel oranges and extending their storage life. An improved DeeplabV3+ model integrating attention mechanism is proposed to increase the current low recognition accuracy and slow detection speed of defect detection in navel oranges grading and sorting process. The improved lightweight backbone network HECA-MobileV3 is applied in the DeeplabV3+ model to reduce the amount of computational data and improve the image processing speed. In addition, the Convolutional Block Attention Module (CBAM) and Channel Space Parallel Mechanism CSPM are integrated to the DeeplabV3+ model. ASPP structure is redesigned and the low feature extraction network is optimized to enhance the capture of target edge information and improve the segmentation effect of the model. Experimental results show that the proposed model exhibits a better MIoU and MPA with 89.50% and 94.02%, respectively, while reducing parameters by 49.42M and increasing detection speed by 55.6fps, which are 7.27% and 3.51% higher than the basic model. The results are superior than U-Net, SegNet and PSP-Net semantic segmentation networks. As a results, the proposed method provides better real-time performance, which meets the requirements of industrial production for detection accuracy and speed.**

Keywords*—Navel oranges; defect detection; DeeplabV3+; HECA-MobileNetV3; CBAM attention mechanism; CSPM mechanism

I. INTRODUCTION

Jiangxi Gannan, home to the world's largest navel oranges plantation, has an annual output of up to one million tons annualy. However, despite its large-scale production, the harvesting and sorting processes still rely heavily on manual methods leading to high labour cost, prominent seasonality and high labour intensity. As the production of navel oranges increases year by year and sales channels continue to expand, the automation of grading and sorting of navel oranges are gradually emerging. After being picked from the trees, navel oranges need to go through disinfection and cleaning, sterilization and waxing, drying and weighing, colour and size grading, skin defect sorting, sugar content density quality analysis, and packaging and labeling before the fruit can be shipped to various parts of the world [1]. At present, the navel orange sorting line can quickly sort navel oranges based on their size and colour, but the recognition of local defects on the skin is not accurate at lower detection rate, which affected the overall quality, delayed the storage time and sorting process of navel oranges. Recent trends in machine vision and advancement in deep learning have led to a proliferation of studies that apply both methods in the field of navel oranges skin defect detection.

The traditional machine vision algorithm is mainly used to grade and classify navel orange defects based on the differences in data such as the RGB colour of the navel orange peel, surface brightness distribution, spectral imaging band curve and edge threshold. Abdelsalam et al. [2] detected the external defects of orange citrus fruits using multi-spectral imaging sensor. They segmented the defects based on the near-infrared (NIR) and RGB images of orange fruits and used threshold technology to detect defects in seven colour components of orange fruits. The overall accuracy of the algorithm exceeded 95%. Rong et al. [3] designed a fast edge detection algorithm for navel oranges surface defects to solve the problem of low defect detection accuracy caused by surface brightness by using the threshold edge segmentation method. Zhang et al. [4] proposed an Otsu threshold segmentation method based on image segmentation according to the different characteristics of navel orange surface defects, and the defect recognition rate is approximately 92.7%. Luo et al. [5] used a visible-near-infrared hyperspectral imaging system with a wavelength range of 325 to 1000 nm to collect citrus hyperspectral images, and used guided soft shrinkage (BOSS) and BOSS-SPA (BOSS-continuous projection algorithm) combined algorithms to optimize the spectral variables. Based on the extracted four defect wavelength images, they proposed a fast multispectral image processing algorithm combined with global threshold theory for rotten orange detection, with an overall classification accuracy of up to 98.6%.

Nowadays, deep learning techniques have been widely applied mainly on 1) object detection and 2) semantic segmentation. Object detection involves recognizing and locating target objects in an image, including algorithms such as R-CNN, YOLO and SSD. Semantic segmentation assigns semantic labels to each pixel in the image, including FCN, U-Net, Deeplab, SegNet and PSPNet. Iqbal et al. [6] determined the difference in fruit surface quality by training the RGB image combination data based on different fruit surface colour data. Asriny et al. [7] applied deep convolutional neural networks to grade the quality of navel oranges, establishing a database of over 1000 navel orange images and achieving a detection accuracy of 96% for different categories of navel oranges. Cai et al. [8] proposed a multi-resolution knowledge distillation strategy by integrating multi-scale pyramid modules and semi-resolution reconstruction branches, training the FastSegfermer

model, which effectively improved the segmentation accuracy of the network, achieving a MIoU of 88.78%.

However, it can be concluded that there are several shortcomings in the navel orange defect detection research. The research up to now has been mainly based on traditional machine vision algorithms where the key challenge is the algorithm is too complex with computational burden resulting in difficulty to achieve real-time online detection. Although there are few common types of surface defects such as anthrax, sun spots and scratches, the highest chances to detect the similarity defect are low due to the nature of the fruit. Especially in the same image, precise segmentation is not achievable, particularly for small defects, thus failing to meet the requirements for grading and classification.

With the rapid development of computing power and artificial intelligence, researchers are also developed other fruit and vegetable detection using spectral technology, ultrasonic imaging and deep learning. Da Costa et al. [9] introduced the ResNet50 model into tomato external defect detection with the accuracy rate of 94.6%. Liang et al. [10] proposed a semantic segmentation method based on BiSeNet V2 deep learning for apple defect detection, and its average pixel accuracy MPA value approximately 99.66%. Hao et al. [11] applied the DeeplabV3+ model for kiwi defect recognition, using the lightweight convolutional neural network MobileNetV2 to extract image features, reducing the training time and achieving an average classification recognition rate of 96%. Gu et al. [12] used the phantom network and coordinate attention module to construct CA-ChostNet as the backbone feature extraction network of DeeplabV3+ for tomato target recognition, which reduced the number of network parameters while improving the model's segmentation capacity for small target categories. In order to improve the real-time performance of apple defect detection, Fan et al. [13] reduced the number of channels and network depth in the YOLOV4 network, reducing the size of the network model to 8.82MB with lower detection time of only 8.36ms for each image. It can be concluded that there is growing interest in deep learning application as main method for fruit and vegetable detection.

A. Navel Oranges Defect Types

There is multiple type of navel orange defects which are spots, scars, mildew, damage, blemishes, and enlarged fruit heads. If navel oranges are simply divided into good and defective fruits can cause great waste. Therefore, it is necessary to accurately identify the type and size of defects to better achieve navel orange grading and sorting.

Semantic segmentation algorithms include classic algorithms such as FCN, U-Net, SegNet, as well as modern deep learning algorithms such as PSPNet, Deeplab, and Mask R-CNN. Among them, the Deeplab network is a model with outstanding semantic segmentation performance at present, and has been gradually optimized from DeeplabV1 to DeeplabV3+. In 2018, DeeplabV3+ introduced an encoder-decoder structure, integrated multi-scale information, and improved the accuracy of image segmentation, becoming the most outstanding model for semantic segmentation. In view of the problems existing in the current research on navel oranges defect detection, combined with the problems of DeeplabV3+ model with many parameters and weak extraction of small target boundary features, this study proposes an improved DeeplabV3+ navel oranges defect real-time detection and segmentation model. The main research contributions of this paper are:

1) MobileNetV3 is used to replace the backbone network Xception, and the improved ECA attention mechanism is used to replace the SE mechanism in the MobileNetV3 network, which greatly reduces the amount of calculation parameters and improves the real-time performance of detection.

2) Redesigned the Atrous Spatial Pyramid Pooling (ASPP) structure of the DeeplabV3+ model. The CBAM attention mechanism and the CSPM mechanism are integrated to dynamically adjust the weight share of the feature channel to increase the attention to important areas of the image and comprehensively improve the recognition capacity of different types of navel orange defects.

3) The CBAM attention mechanism is added to the extraction of low-order feature information to make the extracted low-order features more representative and discriminative, and improve the segmentation effect and stcapacity of the model for boundary features.

4) Navel orange images are collected and a database of more than 2,000 navel orange defects is created to reduce model overfitting and provide more accurate and reliable model evaluation.

II. DeeplabV3+ Network Model with Improved Attention Mechanism

A. DeeplabV3+ Model

The Deeplab model was proposed in 2015. Over the years, with the continuous iteration and optimization of algorithms and technologies, DeeplabV1 [14], DeeplabV2 [15], and DeeplabV3 [16] models have been proposed one after another, continuously improving the model structure while improving the image segmentation. To address the issues of reduced image resolution, lower accuracy, and loss of details caused by max pooling and downsampling in deep convolutional neural networks (DCNNs), DeeplabV1 introduces the Atrous convolution algorithm and fully connected CRF structure. This approach expands the receptive field and connects DCNNs with CRF, thereby improving segmentation accuracy. DeeplabV2 improves the model's backbone network from VGG to ResNet and constructs an ASPP structure. This configuration captures information at multiple scales with high accuracy and capacity through parallel sampling. DeeplabV3 enhances the capacity to capture multi-scale information in images by varying the unit dilated rate of Atrous convolutions. To solve the problem of prolonged processing time and incomplete detail information in high-resolution images with DeeplabV3, DeeplabV3+ introduces an encoder-decoder structure, enhancing network capacity while ensuring the accuracy of feature extraction.

DeeplabV3+ consists of DCNN with dilated convolution and ASPP as the main structure of the encoder. Due to pooling and strided convolutions in the feature extraction process, some image details, particularly boundary features, are lost. To address this, the model integrates high-level features from the encoder with low-level features from the DCNN, enhancing

boundary segmentation accuracy. The DCNN utilizes the Xception backbone, a complex structure with Entry, Middle, and Exit flow layers, leading to a high number of computational parameters and slower training and inference speeds. The ASPP module includes one 1x1 convolution, three 3x3 convolutions with different dilation rates, and one image pooling layer, aimed at dimensionality reduction, multi-scale context information extraction, and global context capture of the input image, respectively. The decoder connects low-level feature maps from the DCNN through a 1x1 convolution with the encoder's 4x upsampled high-level semantic feature maps. It further refines features with 3x3 convolutions and produces accurate prediction maps after another 4x upsampling. The DeeplabV3+ model architecture is illustrated in Fig. 1.

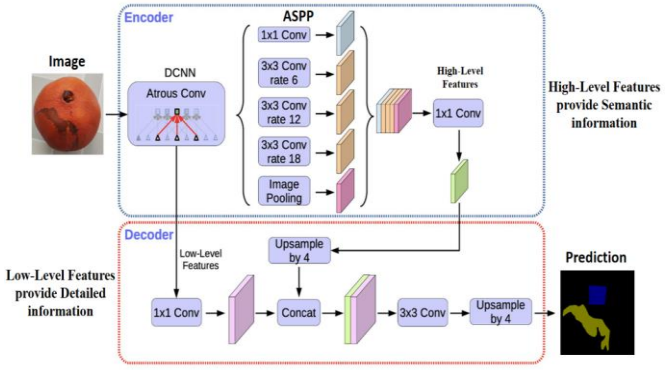


Fig. 1. DeeplabV3+ model structure diagram.

Although the DeeplabV3+ model has demonstrated excellent performance, it still faces challenges such as complex network and large amount of computation, limited capacity to capture details, and strong data dependency. Currently, many scholars have replaced lightweight backbone networks (such as MobileNet and Thin-xception) [17] to reduce the amount of computation and improve real-time performance, or introduced attention mechanisms (such as SE and CBAM) [18] in the model to improve the accurate segmentation of details, or added more multi-scale feature fusion modules (such as FPN [19] and ASPP [20]) to improve the recognition capacity for large-scale targets.

B. Improved MobileNetV3 Backbone Network

The DeeplabV3+ model employs the Xception network as the backbone for feature extraction, which inadequate for navel orange grading and sorting since the network required a large number of parameters and not suitable for the real-time application. Therefore, an improved lightweight MobileNetV3 backbone network is proposed to replace Xception.

The MobileNet network was proposed by the Google team in which MobileNetV3 is a lightweight network model that is continuously improved and optimized based on V1 and V2, which achieved excellent performance in tasks such as image classification and semantic segmentation [21]. Since traditional convolutional neural networks have large computational complexity and consume a lot of memory, depthwise convolution and pointwise convolution are combined in MobileNetV1 to construct a deep separable convolution structure, reducing its parameter volume and computational complexity to one square of the convolution kernel. After continuous verification, it was found that most of the computational parameters of V1's depthwise convolution were zero, limiting its effectiveness. Therefore, MobileNetV2 incorporated the inverted residual block and optimized activation functions, resulting in improvements in segmentation accuracy and processing time compared to the V1 structure. MobileNetV3 retains the depthwise separable convolution and inverted residual block from V2, adds the SE attention mechanism, updates activation functions, and redesigns the structure of time-consuming layers [22]. The network structure of MobileNetV3 is shown in Fig. 2.

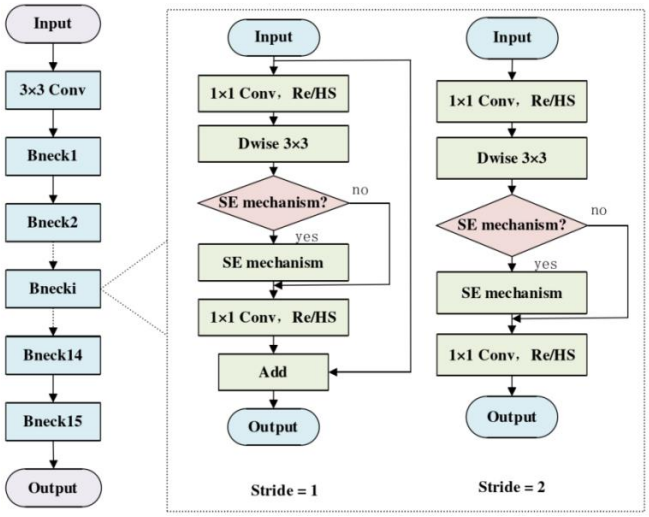
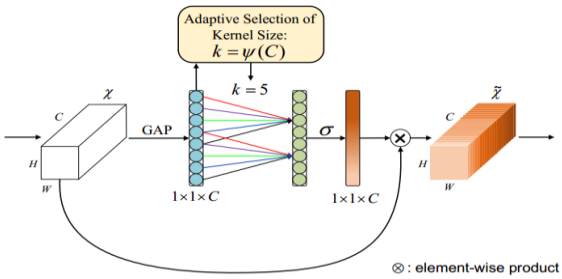


Fig. 2. MobileNetV3 model structure diagram.

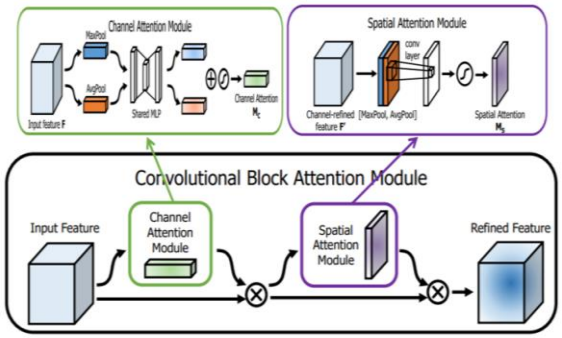
The attention mechanism allows the model to improve its representation ability by focusing on different parts or features of the data during processing data. The main attention mechanisms currently include SE (Squeeze-and-Excitation Networks), ECA (Efficient Channel Attention Module), CA (Coordinate Attention) and CBAM (Convolutional Block Attention Module) [23]. The core idea of the SE module is to enhance the feature map by learning the importance of each channel. The SE module first uses a global average pooling operation to capture the global information of each channel, then learns the weight of each channel through two fully connected layers, and finally uses the sigmoid function to normalize the weight of each channel. The ECA module aims to reduce the computational costs by improving the SE model and capturing inter-channel relationships over a larger spatial range. It captures global information by adaptively determining the size of the one-dimensional convolution kernel through the channel dimension function. Since it does not involve global average pooling, it can reduce the computational cost. The CBAM module connects channel attention and spatial attention in series, allowing the model to dynamically focus on important information in the image in both the channel and spatial dimensions. The structures of the ECA and CBAM models are illustrated in Fig. 3.

Hard-sigmoid is used for the ECA attention mechanism instead of Sigmoid linear activation function as the function offers higher computational efficiency and effectively addresses the vanishing gradient problem. The improved H-ECA attention mechanism is shown in Fig. 4. After feature extraction, the Hard-sigmoid activation function is used to obtain the weight *w*

of each channel, and finally multiplied with the corresponding element of the original feature map to obtain the final output feature map.



(a) ECA attention mechanism.



(b) CBAM attention mechanism.

Fig. 3. ECA and CBAM model structure diagram.

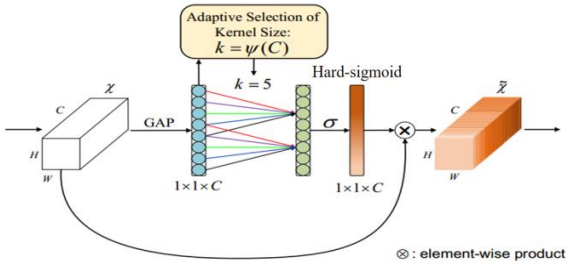


Fig. 4. Improved H-ECA attention mechanism.

To ensure the efficiency of navel oranges sorting, high recognition speed of images during the sorting process is required. The SE attention mechanism with high complexity is used in MobileNetV3, which will affect the response speed and real-time performance of the model. Therefore, to reduce the computational complexity and the number of parameters, the H-ECA mechanism replaces the SE mechanism in *Bneck*, and the feature information of different scales is better captured through local cross-channel convolution operations. The *Bneck* structure diagram of the improved backbone network HECA-MobileNetV3 is shown in Fig. 5.

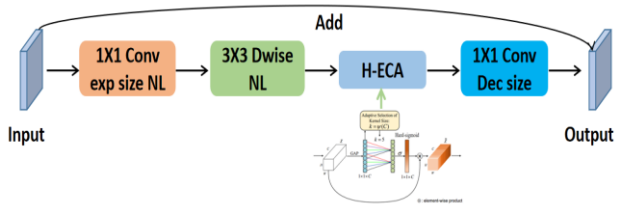


Fig. 5. Improved Bneck structure diagram.

The number of parameters occupied by the attention mechanisms are quantified in both the original MobileNetV3 and the improved HECA-MobileNetV3 networks, as well as their proportion of the total parameters as shown in Table I. The H-ECA attention mechanism replaced the SE attention mechanism in layers 3, 5, 6, 8, 12, 13, 14, and 15. Since H-ECA is not affected by the number of input and output channels, the size of its convolution kernel K is 5, Params = K*1+1, and the calculated Params is 6, accounting for almost 0.00%. It can be seen that the H-ECA attention mechanism occupies only a very small number of parameters.

TABLE I. STATISTICS OF ATTENTION MECHANISM PARAMETERS IN MOBILENETV3 NETWORK BEFORE AND AFTER IMPROVEMENT

Layer	SE		H-ECA	
	Params	Params Proportion	Params	Params Proportion
3	0.003M	0.107%	6	0.000%
5	0.007M	0.228%	6	0.000%
6	0.007M	0.228%	6	0.000%
8	0.105M	3.526%	6	0.000%
12	0.231M	6.722%	6	0.000%
13	0.231M	6.722%	6	0.000%
14	0.460M	13.166%	6	0.000%
15	0.460M	13.166%	6	0.000%
Total	1.504M	43.865%	48	0.000%

C. Improved DeeplabV3+ Network Model

In the actual grading and sorting of navel oranges, a large amount of image data needs to be collected. Even for the same navel orange, images need to be collected from multiple angles, and the detection speed of the results needs to be controlled within the millisecond. At the same time, although the shape of the navel orange does not change much, the details inside the peel are random and variable. The location, size and shape of the defects, the size of the navel and the thickness of the head are all different. Therefore, this study proposes an improved DeeplabV3+ lightweight network model that integrates the attention mechanism. The main improvements are in the following parts:

1) The improved lightweight backbone network HECA-MobileNetV3 is used to replace Xception, which not only reduces the model calculation amount and improves the real-time detection, but also is very easy to add to embedded systems or mobile devices, improving the applicability of the research.

2) In the ASPP structure of the DeeplabV3+ model, the CBAM attention mechanism and the Channel Space Parallel Mechanism (CSPM) were introduced to redesign the ASPP structure. CBAM uses the output of the channel mechanism as the input of the spatial mechanism, with a faster calculation speed and can gradually enhance the expressiveness of the feature map. At the same time, in order to address the key information loss problem that may exist in the CBAM mechanism, a parallel mechanism of channel attention and spatial attention is added to obtain the global dependency

information of the input navel orange image in channels and space, and finally the splicing and fusion are passed to the decoding layer.

3) CBAM attention mechanism is added to the backbone network to extract low-level feature information of the image and adaptively adjust the feature map weights that can enhance key low-order features, suppress noise and redundant information, improve the model's generalization ability, and provide clearer and more effective features as shown in Fig. 6.

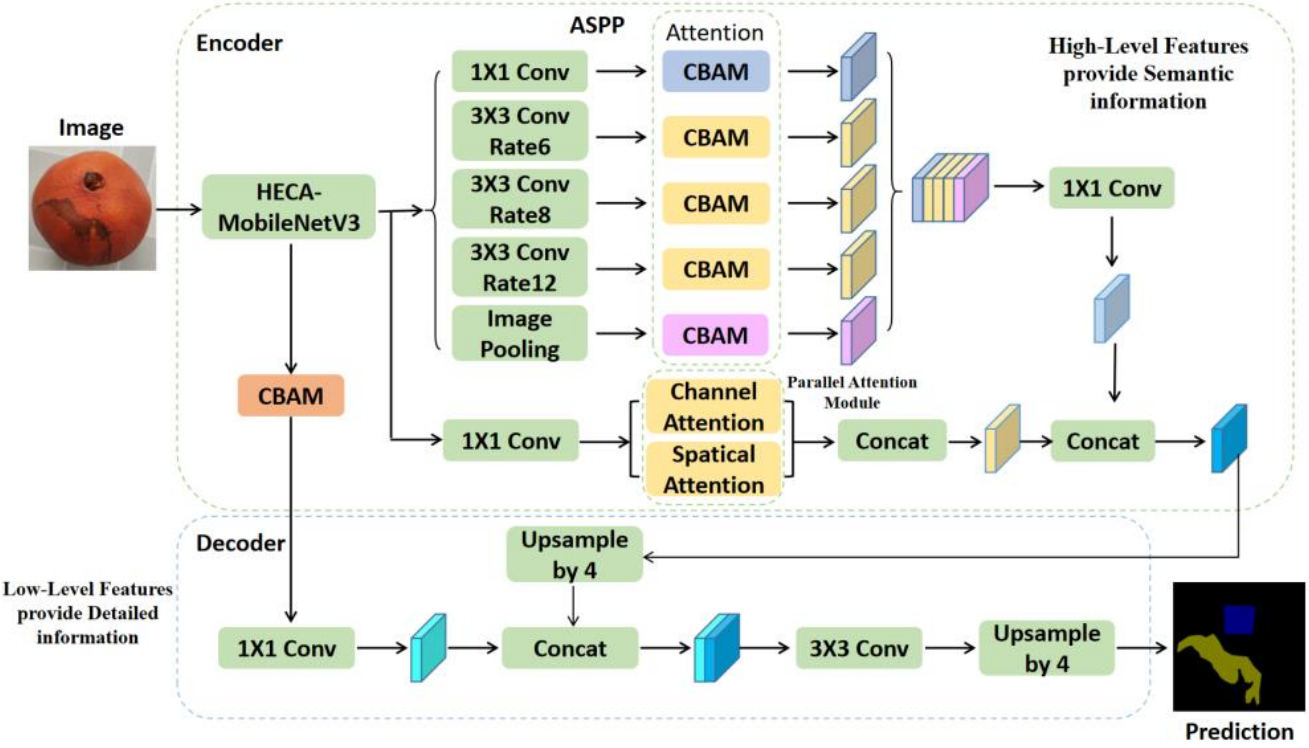


Fig. 6. Improved deeplabV3+ network model.

III. Data Collection and Experimental Evaluation Indicators

A. Data Collection

November to December every year is the harvest season for navel oranges in southern Jiangxi. The shelf life of navel oranges is as long as three to six months. The fruit is available everywhere in the market, which facilitates data collection. In actual fruit and vegetable grading and sorting equipment, multiple industrial depth cameras are often used to collect videos of navel oranges from different angles, convert the videos into pictures, and send them to the controller for intelligent processing [24]. In this study, a smartphone is used to collect image data with a resolution of 3200 x 1440 and 64 mega pixels. All images are collected in a static state, and the size of the collected images is 3472 x 4624. A high-performance computer is used to process image data and optimize, train and analyze deep learning network models. The computer specification is tabulated in Table II.

Various image of navel oranges placed under natural light in the living environment and artificial lighting after harvesting are recorded for database. A total of 800 original images were obtained, including 100 spotted, 100 navel, 100 moldy, 100 thick-heads, 200 damaged, and 200 other defects. Considering the limited number of images, data augmentation techniques were applied using Python's OpenCV and Pillow libraries to rotate, crop, and adjust the brightness of the images to enhance the generalization ability of the model. This process increased the dataset to 2400 images, with the image size is adjusted to 512 x 512.

The collected images were labeled using the *labelme* tool. By segmenting the details of the navel orange peel, six semantic labels were obtained, including spots, mildew, damaged, navel, head hypertrophy and other defects. The *json* file generated by the labeling tool is created as a dataset, and the data was divided into training set and test set in a ratio of 8:2, resulting in 1920 training sets and 480 validation sets.

TABLE II. Computer Hardware and Software Configuration

Hardware and Software	Configuration
Hardware	CPU：Intel(R) Core(TM) i7-14700KF CPU @ 3.4GHz
	GPU：NVIDIA RTX4070Ti Super 16G
	Memory：32G
	Operating System：Windows 11
Software	Deep Learning Frameworks：pytorch 2.2.0
	Image processing software：Open CV 14.2
	Compiled Language：Python 3.12.1

B. Experimental Evaluation Metric Parameters

This study uses the intersection over union (IoU), mean intersection over union (MIoU), and mean pixel accuracy (MPA) parameters to evaluate the accuracy of the model, and uses the parameter quantity and detection speed (FPS) of the model indicators to evaluate the capacity and real-time performance of the model. In all segmentation, recognition, and classification experiments, four types of results exist: true positive (TP), where the actual positive sample is correctly predicted as positive; true negative (TN), where the actual negative sample is correctly predicted as negative; false negative (FN), where the actual positive sample is incorrectly predicted as negative; and false positive (FP), where the actual negative sample is incorrectly predicted as positive. By evaluating the proportions of these outcomes, the effectiveness of the model's predictions can be determined [25]. MIoU is the average IoU of all different semantic categories, where IoU is the ratio of the intersection to the union between the predicted and the ground truth annotations. The formulas for calculating IoU and MIoU are as follows:

$$\mathrm{IoU} = \frac{\text{Predictive Value} \cap \text{True Value}}{\text{Predictive Value} \cup \text{True Value}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} = \frac{\mathrm{P_{ii}}}{\mathrm{P_{ij}} + \mathrm{P_{ji}} + \mathrm{P_{ii}}}$$

$$MIoU = \frac{1}{K+1}\sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}}$$

MPA is the mean pixel accuracy, which calculates the proportion of correctly classified pixels for each semantic category and determine the average value. Due to the imbalanced distribution of positive and negative samples in this study, this indicator is can be used to measure the proposed method performance. The calculation formula is as follows:

$$\mathrm{MPA} = \frac{1}{\mathrm{K}+1}\sum_{\mathrm{i=0}}^{\mathrm{k}} \frac{\mathrm{p_{ij}}}{\sum_{\mathrm{j=0}}^{\mathrm{k}} \mathrm{p_{ij}}}$$

where *i* represents the true value, *j* represents the predicted value, and *k* represents the number of semantic categories.

The size and complexity of the model have a great impact on the requirements of the device's hardware performance. The number of model parameters is an important measurement indicator, which is related to the number of input and output channels and the size of the convolution kernel; the number of frames per second represents the images that the model can detect per second, and its calculation formula is as follows:

$$\mathrm{Params} = \sum_{\mathrm{l=1}}^{\mathrm{D}} \mathrm{K_l^2 C_{l-1} C_l} + \sum_{\mathrm{l=1}}^{\mathrm{D}} \mathrm{C_l}$$

$$\mathrm{FPS} = \frac{n}{time}$$

where K is the convolution kernel size, C is the number of channels, n is the number of images segmented by the model, and time is the total time required for model segmentation.

IV. Experimental Results

This study uses stochastic gradient descent for training. After repeated optimization, the initial learning rate was set to 0.005 and the batch size was set to 8. The data size was 512 × 512, with a batch size of 4, and the number of training iterations is 100. The first 50 iterations were used for frozen training, during which the backbone feature extraction network was frozen to accelerate training, while the remaining 50 iterations were used for unfrozen training to fine-tune the parameters. The Adam optimizer was selected for this study as this optimizer uses the first-order momentum and the second-order momentum, which can dynamically adjust the learning rate and make the model converge faster.

A. Ablation Experiment

To validate the effectiveness of the various improvements made to the DeeplabV3+ model in enhancing the segmentation accuracy for navel orange defects, ablation studies were conducted by modifying the backbone network, adding the CBAM mechanism, and incorporating the CSPM mechanism. The experimental results of these segmentation effects are presented in Table III.

As shown in Table III, the DeeplabV3+ model with the HECA-MobileNetV3 backbone exhibits the lowest parameter count and the highest frame rate after improvements; after incorporating the CBAM attention mechanism into the basic model, MIoU increased by 3.73% and MPA increased by 2.47%. As CSPM attention mechanism is incorporated into the model, MIoU increased by 3.41% and MPA increased by 1.99%; and after adding both CBAM and CSPM mechanisms to the model, MIoU reached 89.50%, an increase of 8.01%, and MPA reached 94.02%, an increase of 3.86%. These results indicate that the improvements proposed in this study effectively enhance the accuracy and precision of navel orange defect segmentation.

B. Cross-Entropy Loss Function

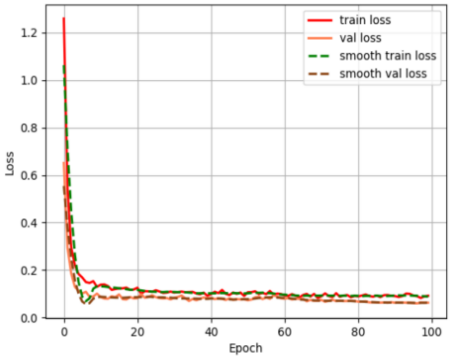
In per-class segmentation experiments, the cross-entropy loss function is often used to check each pixel one by one, and the predicted value is compared with the actual value to average the loss. Since the individual differences of some defective navel oranges are small, it is easy to cause semantic recognition errors of defects. Therefore, the multi-classification cross-entropy loss function is used to measure the segmentation effect. The loss curve for the DeeplabV3+ model only with the HECA-MobileNetV3 backbone is illustrated in Fig. 7(a), while the loss curve for the DeeplabV3+ model incorporating the attention mechanisms is shown in Fig. 7(b).

The horizontal axis represents the number of iterations, and the vertical axis represents the calculated loss value. The train loss represents the loss calculated during training; the val loss represents the loss calculated in the confirmation; the smooth train loss and smooth val loss represent the smooth loss values during training and verification respectively. It was found that the loss values of the two models gradually stabilize with the increase of the number of iterations during the training process. The Deeplabv3+ model improved by the fusion attention mechanism with the smallest loss value, stronger convergence and more stability.

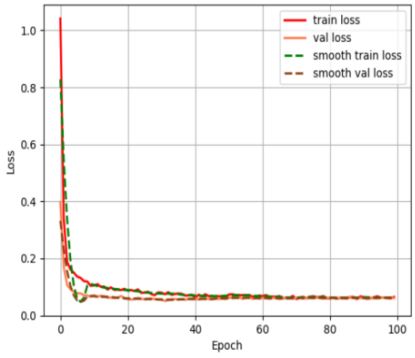
TABLE III. SEGMENTATION ACCURACY OF DIFFERENT BACKBONE NETWORKS AND MODELS WITH ATTENTION MECHANISM

Model	MIoU/%	MPA/%	Params/M	FPS/fps
D+X	82.23	90.51	56.14	15.2
D+M	80.71	89.27	6.61	72.3
D+H	81.49	90.16	**5.83**	**76..5**
D+H+CB	85.22	92.63	6.58	74.1
D+H+CS	84.95	92.15	6.47	74.9
D+H+CB+CS	**89.50**	**94.02**	6.92	70.8

D：DeeplabV3+，X：Xception，M：MobileNetV3，H：HECA-MobileNetV3，CB：CBAM，CS：CSPM



(a) Change HECA-MobileNetV3 loss curve.



(b) Fusion attention mechanism loss curve.

Fig. 7. Loss curve of multi-classification cross entropy loss function.

C. Comparative Experiments

To ensure the validity of the research, several classical networks were applied to this navel orange dataset for comparative experiments. These experiments aim to verify the effectiveness of the improved DeeplabV3+ model incorporating the CBAM attention mechanism in segmenting navel oranges defects. The experimental results are shown in Table IV.

As shown in Table IV, the DeeplabV3+ model with MobileNetV3 as the backbone network has the fewest parameters and the highest frame rate, but lower MIoU and MPA. The improved DeeplabV3+ model significantly outperforms the other five models in terms of MIoU and MPA. Specifically, the MIoU of the improved DeeplabV3+ model is 21.74%, 16.06%, 13.01%, 7.27%, and 8.79% higher than those of Unet, SegNet, PSPNet, DeeplabV3+ with Xception backbone, and DeeplabV3+ with MobileNetV3 backbone, respectively. In terms of MPA, it is 29.31%, 14.35%, 7.57%, 3.51%, and 4.75% higher, respectively. The improved DeeplabV3+ model reduces the number of parameters by 49.42MB and increases the frame rate by 55.6fps compared to the DeeplabV3+ model with Xception backbone. Although its parameter count and frame rate are slightly lower than those of the DeeplabV3+ model with MobileNetV3 backbone, its MIoU and MPA are significantly higher than those of the unimproved DeeplabV3+ model.

To provide a more intuitive comparison of the segmentation performance of different models, this study selected five representative images for visual contrast. The effectiveness of the improved DeeplabV3+ model incorporating the attention mechanism is compared with Unet, SegNet, PSPNet, DeeplabV3+ with Xception backbone, and DeeplabV3+ with MobileNetV3 backbone in navel orange defect segmentation, as shown in Fig. 8.

The result shows the improved DeeplabV3+ model provides the clearest segmentation boundaries and accurately identifies small defects in navel oranges, achieving higher detection performance compared to other models. The unimproved DeeplabV3+ model performs better than the other networks but still exhibits some issues with fuzzy boundary segmentation and misidentification of small targets. Unet, SegNet, and PSPNet networks also suffer from varying degrees of recognition errors and fuzzy boundary segmentation.

TABLE IV. COMPARISON OF EXPERIMENTAL RESULTS WITH THE CLASSIC NETWORK

Model	Backbone network	MIoU/%	MPA/%	Params/MB	FPS/fps
U-Net	ResNet50	67.76	64.71	23.9	38.9
SegNet	VGG16	73.44	79.67	21.8	45.8
PSP-Net	ResNet101	76.49	86.45	27.6	42.3
DeeplabV3+	Xception	82.23	90.51	56.14	15.2
DeeplabV3+	MobileNetV3	80.71	89.27	**6.61**	**72.3**
Improved DeeplabV3+	HECA-MobileNetV3	**89.50**	**94.02**	6.72	70.8

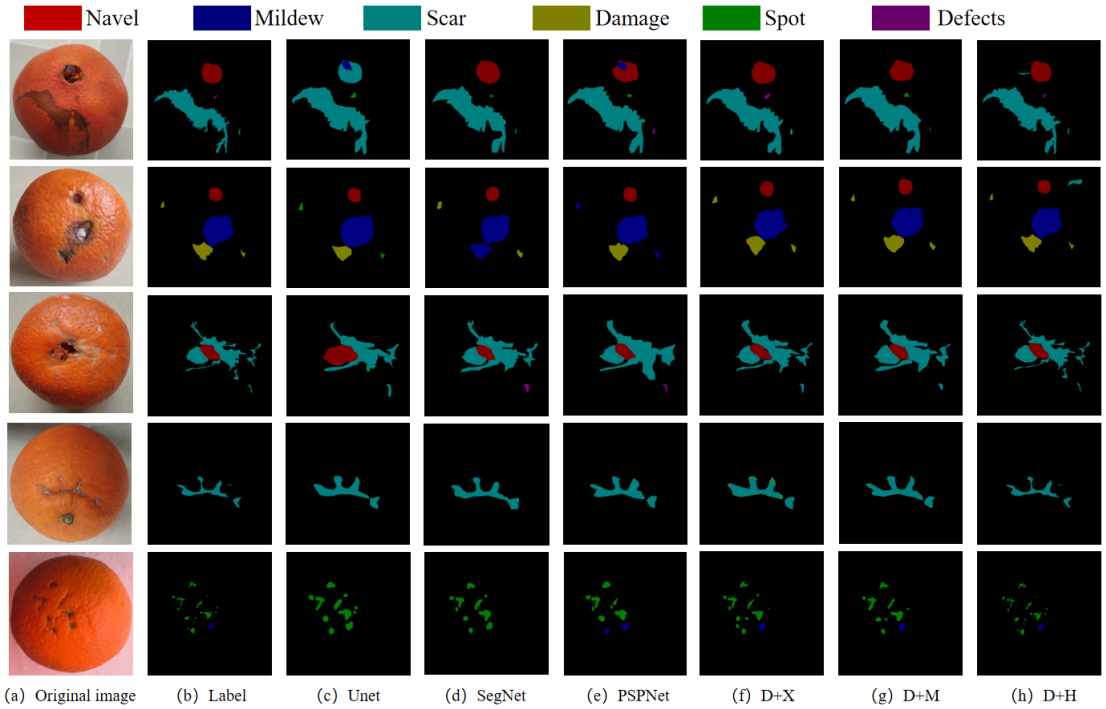


Fig. 8. Effects of navel orange defect segmentation detection using different models.

V. Discussion

Unlike other improved DeeplabV3+ methods, this study does not simply replace the complex Xception network with a lightweight backbone network MobileNetV3, but uses the improved H-ECA mechanism to replace the SE attention mechanism in the MobileNetV3+ structure. The Hard-sigmoid activation function is applied to the ECA structure, which can effectively improve the model calculation efficiency and improve the gradient disappearance problem. ECA is an improved structure based on the SE mechanism. The Hard-sigmoid activation function is combined with ECA and applied to the MobileNetV3 structure, which not only makes the backbone network lighter and ensures the real-time detection, but also effectively improves the ability to extract image features. At the same time, the CBAM and CSPM attention mechanisms are flexibly integrated into the shallow and deep feature extraction networks of DeeplabV3+, and the weights of the feature maps are adaptively adjusted in the two dimensions of channel and space, which improves the sensitivity to navel orange defects, focuses more on high-level semantic information, captures key information that is easily lost in model up and down sampling, integrates global and local features, improves the model's feature representation and generalization capabilities, and achieves more accurate semantic segmentation. Experimental results show that DeeplabV3+ with integrated attention mechanism has faster segmentation speed and higher accuracy.

The improved DeeplabV3+ model integrates attention mechanisms at multiple levels, which plays an important role in the convolution of each layer of the model. It has good segmentation performance for multiple categories of defects on the surface of navel oranges, but for some defects without obvious boundaries and light colors, the segmentation performance of this model is not as good as other defects. In future research, the DeeplabV3+ model will be further improved, such as using a more powerful backbone network, introducing a residual structure, improving the loss function, applying adaptive feature pyramid technology, and proposing a new model structure, which will be applied to navel orange defect segmentation in order to obtain better results, which can further applied to other fruit and vegetable defect detection and other image segmentation fields.

VI. Conclusion

In this paper, an improved new semantic segmentation model DeeplabV3+ model is proposed that incorporates an attention mechanism to solve the problems of low recognition accuracy and slow detection speed of similar defects and small targets in the navel orange defect grading and sorting task. By employing the improved HECA-MobileNetV3 backbone network, the model reduces parameters and enhances real-time detection. The CBAM mechanism is integrated into the ASPP structure and an additional CSPM mechanism is introduced to improve distinguishing and recognition capabilities for similar defect features. Furthermore, CBAM is incorporated into the low-level feature extraction structure to enhance segmentation of small target boundary features. Comparative study with DeeplabV3+ model was conducted resulting in improvement of MIoU of 89.50% and MPA of 94.02%, while reducing parameters by 49.42M and increasing detection speed by 55.6fps. In comparison to other semantic segmentation networks, the proposed model achieves higher detection accuracy and segmentation effectiveness while maintaining advantages in parameter efficiency and speed. The algorithm presented in this paper effectively meets the precision and speed compatibility requirements for navel orange defect grading and sorting in industrial applications.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Jiangxi Province (Grant No. 20224BAB214056), The Science and Technology Project of Jiangxi Province (Grant No. GJJ214901) and Universiti Teknologi MARA.

REFERENCES

[1] P. Zhou, J. Wei, T. Zhong and H. Zheng, "The Research on Navel Oranges Detection Systems of Harvesting Robots Based on an Improved YOLOv5," 2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR), Tianjin, China, 2023, pp. 537-542.

[2] H. Javadikia, S. Sabzi, and H. Rabbani, "Machine vision based expert system to estimate orange mass of three varieties," International journal of agricultural and biological engineering, vol. 10, no. 2, pp. 132–139, Mar. 2017.

[3] D. Rong, Y. Ying, and X. Rao, "Embedded vision detection of defective orange by fast adaptive lightness correction algorithm," Computers and Electronics in Agriculture, vol. 138, pp. 48–59, Jun. 2017.

[4] M. Zhang, T. Wang, P. Li, and Y. Zheng. "Surface defect detection of navel orange based on region adaptive brightness correction algorithm," Chinese Agricultural Science, vol. 52, no. 2, pp. 2360-2370, Jan. 2019.

[5] W. Luo, G. Fan, P. Tian, W. Dong, H. Zhang, B. Zhan. "Spectrum classification of citrus tissues infected by fungi and multispectral image identification of early rotten oranges". Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 279, 121412, Oct. 2022.

[6] S. Md. Iqbal, A. Gopal, P. E. Sankaranarayanan, and A. B. Nair, "Classification of Selected Citrus Fruits Based on Color Using Machine Vision System," International Journal of Food Properties, vol. 19, no. 2, pp. 272–288, May 2015.

[7] D. M. Asriny, S. Rani, and A. F. Hidayatullah. "Orange Fruit lmages Classification using Convolutional Neural Networks," IOP Conference Series Materials Science and Engineering, Vol. 803. No. 1, pp. 12-20, Apr. 2020.

[8] X. Cai, Y. Zhu, S. Liu,and Y. Xu. "FastSegFormer: A knowledge distillation-based method for real-time semantic segmentation of surface defects in navel oranges,Computers and Electronics in Agriculture - X-MOL," X-mol.com, 2024.

[9] A. Z. Da Costa, H. E. H. Figueroa, and J. A. Fracarolli. "Computer vision based detection of external defects on tomatoes using deep learning," Biosystems Engineering, pp. 131-144, Feb. 2020.

[10] X. Liang et al., "Real-Time Grading of Defect Apples Using Semantic Segmentation Combination with a Pruned YOLO V4 Network," vol. 11, no. 19, pp. 3150–3150, Oct. 2022.

[11] J. Hao, Y. Zeng, X. Wang,et al."Research on kiwifruit feature extraction and automatic grading based on DeeplabV3+," Agricultural Machinery and Agronomy, vol. 55, no. 03, pp.49-54, Feb. 2024.

[12] W. Gu, J. Wei, Y. Yin, X. Liu, and C. Ding. "Multi-category segmentation method of tomato images based on improved Deeplabv3+," Transactions of the Chinese Society of Agricultural Machinery, vol. 54, no. 12, pp.261-271, 2023.

[13] S. Fan et al., "Real-time defects detection for apple sorting using NIR cameras with pruning-based YOLOV4 network," Computers and Electronics in Agriculture, vol. 193, p. 106715, Feb. 2022.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," arXiv.org, 2014.

[15] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 1 April 2018.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," arxiv.org, Jun. 2017, Available: https://arxiv.org/abs/1706.05587.

[17] H. Peng, S. Xiang, M. Chen, H. Li and Q. Su, "DCN-Deeplabv3+: A Novel Road Segmentation Algorithm Based on Improved Deeplabv3+," in IEEE Access, vol. 12, pp. 87397-87406, 2024.

[18] R. Liu and D. He, "Semantic Segmentation Based on Deeplabv3+ and Attention Mechanism," 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2021, pp. 255-259,

[19] T. Zhang, R. Zhou, L. Zhang and M. Liang, "Research on Multi-scale Feature Fusion Method for Target Detection Based on IN-FPN," 2023 8th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 2023, pp. 94-98.

[20] T. Lei et al., "Ultralightweight Spatial–Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–14, Jan. 2023.

[21] A. Howard et al., "Searching for MobileNetV3," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1314-1324.

[22] S. Qian, C. Ning and Y. Hu, "MobileNetV3 for Image Classification," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 490-497.

[23] H. Yang, L. Lin, S. Zhong, F. Guo and Z. Cui, "Aero Engines Fault Diagnosis Method Based on Convolutional Neural Network Using Multiple Attention Mechanism," 2021 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), Weihai, China, 2021, pp. 13-18.

[24] P. Nirale and M. Madankar, "Analytical Study on IoT and Machine Learning based Grading and Sorting System for Fruits," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 2021, pp. 1-6.

[25] L. Yu et al., "A Lightweight Complex-Valued DeepLabv3+ for Semantic Segmentation of PolSAR Image," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 930-943, 2022.