Multi-class segmentation of navel orange surface defects based on improved DeepLabv3+

Yun Zhu,[1] Shuwen Liu,[1] Xiaojun Wu,[1] Lianfeng Gao,[1] Youyun Xu[2]

[1]School of Physics and Electronic Information, Gannan Normal University, Ganzhou, Jiangxi; [2]School of Communication and Information Engineering, University of Posts and Telecommunications, Nanjing, Jiangsu, China

Abstract

A multi-class segmentation model based on improved DeepLabv3+ is proposed to detect navel orange surface defects. This model aims to address the problems of the current mainstream semantic segmentation network, including rough edge segmentation of navel orange defects, poor accuracy of small target defect segmentation, and insufficient deep-level semantic extraction of defects, which will result in the loss of feature information. In order to improve semantic segmentation performance, the Coordinate Attention Mechanism is integrated into the DeepLabv3+ network. Additionally, the deformable empty convolution of the Atrous Spatial Pyramid Pooling structure replaces the dilated convolution, improving the network's ability to fit and target irregular defects and shape changes. Furthermore, to achieve multi-scale feature fusion and enhance feature space and semantic information, a Bi-feature pyramid network-based feature fusion branch is added at the DeepLabv3+ encoder side. The experimental findings demonstrate that the improved DeepLabv3+ model improves the extraction capability of navel orange defect features and has better segmentation performance. On the navel orange surface defect dataset, the improved model's average intersection ratio and average pixel intersection ratio accuracies are 77.32% and 86.38%, respectively, which are 3.81% and 5.29% higher than the original DeepLabv3+ network.

Correspondence: Shuwen Liu, School of Physics and Electronic Information, Gannan Normal University, Ganzhou, Jiangxi, China.
E-mail: shuwen0514@163.com

Key words: BiFPN; coordinate attention mechanism; deformable convolution; navel orange surface defects; semantic segmentation.

Contributions: the authors contributed equally.

Conflict of interest: the authors declare no potential conflict of interest.

Received: 2 May 2023.
Accepted: 15 December 2023.

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Introduction

As a fruit with a long hanging period, navel oranges are often damaged by pests and diseases, and mechanical and chemical factors during the fruit growing period, leaving scars on the mature fruit, which greatly reduces its commercial value and affects the profitability of fruit farmers. Fruit surface defect detection (Jin *et al.*, 2021; Soltani Firouz & Sardari, 2022) is a key step for fruits to enter the market, which is based on two parts, defect extraction and defect classification, to determine the merit of the fruit while grading the fruit according to the type and percentage of defects (Ren & Bai, 2013; Zhang *et al.*, 2015; Unay, 2022). This technology also enables fruit farmers to improve the quality of their fruits by allowing them to be more targeted for future orchard maintenance in terms of planting and cultivation. The development of postharvest treatment methods for different types of defective fruits also allows for reducing the waste of resources and maximizing their value.

Initially, the detection of defects on the surface of fruits is mainly manual, relying on the subjective experience of the picker to segment the defects, which is prone to human error and variable results and can consume a lot of labor and time costs (Li *et al.*, 2015; Zhang *et al.*, 2015). Later, defect detection methods based on machine vision became popular, and the early applied methods were mainly image processing and machine learning. Yang *et al.* (2014) analyzed the color information of the navel orange surface in defect detection using image processing techniques and obtained surface defects with appropriate R/B and G/B ratios with an identification accuracy of 93.3%. Rong *et al.* (2017) gave a detailed image processing procedure and proposed a comparative sliding window local segmentation algorithm, which was applied to 1,191 navel orange samples with a defect detection rate reaching 97%. Xie *et al.* (2018) proposed a fast navel orange surface defect detection algorithm by combining wavelet transform and compressed sensing techniques in image processing, but the number of false matches increases if the image contains fruit stalks and nectaries. Bhargava *et al.* (2020) proposed a fully automated detection and categorization mechanism for a wide range of fruits. The mechanism uses four machine learning algorithms to classify and detect fruits after segmenting the fruit region and extracting features, among which the support-vector machines classifier has the best detection performance with 98.48% classification accuracy. The traditional method mainly analyzes the surface defect features of the fruit and manually designs the feature extractor, which can obtain good detection results under specific environments. However, due to the large influence of the environment, the extracted image features are often difficult to generalize to new images (Fan *et al.*, 2020; Nithya *et al.*, 2022). Compared with traditional image processing methods and machine learning methods, deep learning does not require a manual feature selection process, which not only reduces the difficulty of fruit defect segmentation but also has higher accuracy and robustness. The defect detection methods based on deep learning mainly include the image classification method, target detection method, and semantic segmentation method.

At present, most fruit defect detection methods use image classification, and there have been many related research results. Zhou *et al.* (2020) used the stochastic weighted average optimizer and w-softmax loss function to improve the VGG network, and generated a network model for Qingmei defect detection, with an average defect recognition accuracy of 93.8%, but due to the small sample size, the recognition rate of scars and cracks was low. Tian *et al.* (2022) proposed a transfer learning-based classifier for nine tomato diseases and a healthy tomato leaf recognition mechanism, and trained three deep-learning network architectures (VGG16, Inception_v3, and Resnet50) with a test accuracy of 99%. However, image classification cannot distinguish multiple defects on a single image, and the defect classification scene is single. In terms of the application of target detection methods, Yao *et al.* (2021) developed a kiwifruit defect detection model based on improved YOLOv5, which added a small object detection network to the backbone network and embedded squeeze-and-excitation Layer to improve the extraction ability of the model, and the results showed that the mAP@0.5 of the model reached 94.7%. Target detection makes up for the shortcomings of image classification by locating the defect location through a rectangular box, but the localization becomes more difficult in the case of complex defect edges.

Semantic segmentation is based on pixel classification, which can accurately segment defect edges and is more suitable for defect detection with complex features. In recent years, fruit defect detection based on the semantic segmentation method has gained more and more attention. Sun *et al*. (2020) constructed an attention network (FANet) embedded Unet semantic segmentation model to recognize the type of segmented orange defects and distinguish between stem end and flower, and the average recognition accuracy can reach 77.468%. Raman *et al.* (2022) investigated the apple disease classification and segmentation mechanism and improved the standard Unet by using Atrous Convolution for segmentation in step-skipping branches, and this improved Unet model could achieve up to 94.29% accuracy for apple disease recognition. Liang *et al*. (2022) propose a semantic segmentation method based on BiSeNetV2 deep learning network to segment the defective parts of defective apples, and use the model pruning method to optimize the YOLOv4 network structure to help solve the problem of segmentation networks incorrectly segmenting fruit stems. The final mPA of the apple defect detection model based on BiSeNetV2 can be obtained as 99.66% and the average accuracy of the apple classification model based on YOLOv4 is 92.42%.

Selecting an appropriate image segmentation model based on the characteristics of the dataset is a key step in conducting research on fruit defect detection. Since there are multiple defective regions with similar and irregular morphology on the surface of navel oranges, its detection requires a high-precision semantic segmentation network. Considering the current transformer type of segmentation model has more parameters, we mainly select the baseline model from the mainstream CNN models. In the field of defect detection, the semantic segmentation models with better detection effects are Unet (Ronneberger *et al.*, 2015), PSPNet (Zhao *et al.*, 2017), and DeepLabv3+ (Chen *et al*., 2018). Unet is a kind of symmetric U-shape structure of the encoder and decoder network, the encoder gradually reduces the resolution of the feature map, and the decoder gradually restores the feature map resolution, which helps to retain detail information and is suitable for semantic segmentation tasks with small samples and unbalanced data. PSPNet is able to add contextual information by introducing the Pyramid Pooling Module (PPM), which improves the accuracy and robustness of semantic segmentation of images. Moreover, adding the PPM module does not increase the number of parameters too much when the input feature dimensions are small. DeepLabv3+ uses the Atrous Spatial Pyramid Pooling (ASPP) module, which is similar in structure to the PPM module, to effectively deal with the problems of many types of defects in the dataset, irregular regions, and unclear details. Compared with the PPM module, the ASPP module adds Atrous Convolution to expand the receptive field without increasing the number of parameters and computation, thus capturing a wider range of contextual information. By using multiple parallel Atrous Convolution branches, ASPP can process the input feature maps with different sampling rate receptive fields, which can effectively capture semantic information at different scales. To further improve the utilization of features at different scales, DeepLabv3+ introduces a feature fusion module. Feature fusion improves the accuracy of semantic segmentation by fusing low-level feature (LF) maps with high-level feature (HF) maps, which can retain both details and global information. Therefore, DeepLabv3+ is more suitable for navel orange defect detection. The purpose of this paper is to design an umbilical cord orange image defect detection algorithm using image segmentation technology in deep learning to realize fast and accurate real-time detection of defects in umbilical cord orange images and to provide technical reference for umbilical cord orange surface defect detection.

Materials and Methods

DeepLabv3+ semantic segmentation modeling

The structure of DeepLabv3+ is shown in Figure 1. DeepLabv3+ network uses the encoder-decoder structure. For the encoder part, first, the image enters the encoder for feature extraction, and after deep convolutional neural network (DCNN) a shallow feature layer and a deep feature layer are generated, the height and width of the shallow feature layer will be larger, while the deep feature layer will have more downsampling, so the height and width will be smaller. The deep feature layer enters the ASPP structure, and further feature extraction is performed using the Atrous Convolution with different expansion rates, where there are 3x3 convolutions with expansion rates of 6, 12 and 18, which are used to improve the receptive field of the network and make the network have different feature perceptual situations, after which the feature layers are stacked and then adjusted by 1x1 convolution for the number of channels to obtain the fused feature (FF) layer. In the decoder part, the shallow feature layer generated by DCNN enters into the Decoder decoder, and the feature layer with high semantic information generated by the encoder enters into the Decoder for upsampling, after which the results obtained from 1x1 convolution with the shallower features are fused with the features, after which the feature extraction is performed by 3x3 convolution, and, finally, the output image is up-sampled with the input. The image size is the same and the prediction result is obtained.

Coordinate attention mechanism

The DeepLabv3+ network contains multiple feature channel fusion operations, and the features of different channels undergo different convolution operations, and, as the convolution depth increases, the semantic features obtained become more abstract, and their impact on target prediction will be different. The attention mechanism can selectively focus on important information with high weights and ignore irrelevant information with low weights. In addition, the mechanism can adaptively adjust the information weights to select critical information according to the scene requirements, which enhances the scalability and robustness of the model. Squeeze and Energize (SE) attention is one of the most influential attention mechanisms, which learns inter-channel relationships and compresses them into channel importance vectors through global information pooling, squeezing, and excitation operations, and scales them to 0 to 1 through excitation operations, which ultimately achieves attention weighting for different channels. However, the SE module ignores the location information (Hu *et al.*, 2018). To compensate for SE attention, Woo *et al.* (2018) proposed the Convolutional Block Attention Module (CBAM), which introduces spatial information encoding through the convolution of a large-size kernel, but it can only capture local relations and not long-term dependencies that are important for visual tasks. In order to obtain long-distance dependencies with accurate location information, coordinate attention has been further proposed.

Coordinate attention (Hou *et al.*, 2021) allows the attention mechanism to capture long-range dependencies and precise location information in different spatial directions so that the network can focus more on regions or targets of interest. It encodes spatial information into two parallel one-dimensional feature codes and uses the two one-dimensional feature codes to insert coordinate information to avoid the loss of position information caused by two-dimensional global pooling. Coordinates encode accurate position information for channel relations and long-term dependencies in two steps: coordinate information embedding and coordinated attention generation.

Coordinate information embedding

In coordinate attention, to better capture long-range dependencies with precise location information, we use a pair of one-dimensional feature encoding operations. The spatial extent of the pooling kernels (*H*, 1) is used to encode the channel of horizontal coordinates, while the spatial extent of the pooling kernels (1, *W*) is used to encode the channel of vertical coordinates.

Given the input, $X=\{x_1,x_2,\text{K},x_C\}\in \mathrm{R}^{C\times H\times W}$ the output of the *c* channel with height *h* can be expressed as

$$z_c^w(h)=\frac{1}{H}\sum_{0\le j<H} x_c(j,w) \tag{1}$$

where x_c (*h, i*) indicates the value of the height coordinate *h* and width coordinate *j* position feature map of channel *c* and *w* is the width of the feature map.

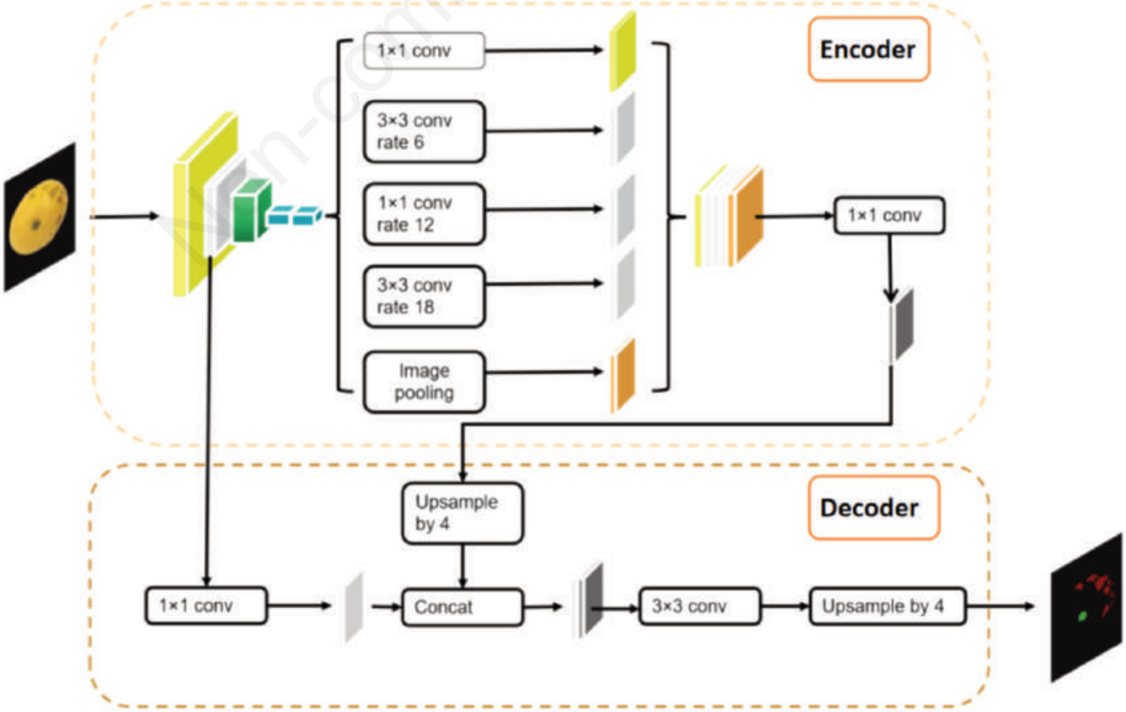


Figure 1. Structure of DeepLabv3+ algorithm.

Similarly, the output of channel c of width w is expressed as

$$z_c^w(h) = \frac{1}{H}\sum_{0\le j<H} x_c(j,w) \quad (1)$$

where x_c (i, w) is the value of the positional feature map of channel c with width w and height coordinate i, H is the height of the feature map.

Coordinate attention generation

The coordinate information embedding module provides a representation of the global receptive field and precise location information. To make better use of these features, the coordinate attention generation operation is required. This operation cascades two feature maps and transforms F_1 using a shared 1x1 convolution to generateas intermediate feature maps for spatial information in the horizontal and vertical directions, with the same downsampling ratio r as the SE module for controlling the module size, which is expressed in the following equation

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \quad (3)$$

where [,] is the splicing operation along the spatial dimension and σ is the nonlinear activation function defined as:

$$\delta(x) = \frac{\mathrm{ReLU6}(x+3)}{6} \quad (4)$$

where ReLU6 is defined as:

$$\mathrm{ReLU6}(x) = \min(6, \max(0, x)) \quad (5)$$

In order to prevent gradient explosion during reverse transmission, unlike the ReLU function, the output of ReLU6 is limited to a maximum value of 6. Then, f is sliced into two separate tensors $f^h \in R^{C/r\times H}$ and $f^w \in R^{C/r\times W}$ along the spatial dimension, and the feature maps f^h and f^w are transformed to the same number of channels as the input X using two 1x1 convolutions F_h and F_w to obtain the following equation

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \quad (6)$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \quad (7)$$

where σ is the sigmoid activation function defined as

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (8)$$

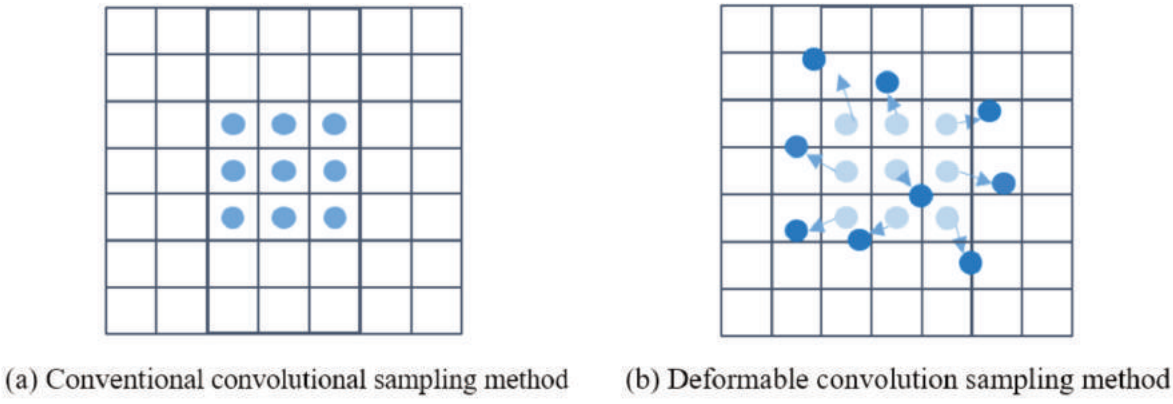
Then g^h and g^w are expanded as attention weights, and the final output of the CA module can be expressed as the following equation.

$$y_c(i,j) = x_c(i,j)\times g_c^h(i)\times g_c^w(j) \quad (9)$$

Deformable convolution

The DeepLabv3+ network uses dilated convolution instead of the down-sampling operation to increase the receptive field of the network. Unlike the traditional down-sampling operation, the Atrous Convolution increases the step length between convolution kernels by adding a dilation rate parameter, thus expanding the receptive field without reducing the resolution of the image. The use of Atrous Convolution can increase the range of information received by each neuron, thus improving the model's understanding of the input data features. However, the convolutional kernel used in this method is square, and the use of a square convolutional kernel does not fully satisfy the need for recognizing objects of different sizes, shapes, and resolutions in all scenes. In order to make the convolutional neural network more efficient in extracting the key semantic information of navel orange surface defects, this paper proposes to introduce deformable convolution from ASPP, which can adaptively adjust the sampling points to precisely locate the objects of different scales and shapes and extract the features. The regular convolution sampling method and the deformable convolution sampling method are shown in Figure 2.

As shown in Figure 2, Figure 2a shows the conventional 3x3 convolution sampling method, where the sampling area is a regular region, and Figure 2b shows the deformable convolution sampling



(a) Conventional convolutional sampling method

(b) Deformable convolution sampling method

Figure 2. Plot of conventional and deformable convolution sampling methods.

method, where the deformable convolution adaptively adjusts the sampling point positions according to the shape of the target and makes each sampling point have different degrees of offset in different directions, thus allowing the network to focus more on the region or target of interest. In traditional convolution, the input and output feature maps are defined as *x* and *y* respectively, and p_0 is the coordinate in the output feature map, which is the coordinate of the convolution kernel in the template. The convolution process of traditional convolution can be expressed as

$$y(p_0)=\sum_{p_n \in T}\omega(p_n)\cdot x(p_0+p_n) \tag{10}$$

where p_0+p_n is the coordinate of the sampling point, w(p_n) is the weight parameter in the convolution kernel, *T* is the kernel template.
The deformable convolution introduces an offset for each point based on the traditional convolution, which can be expressed as

$$y(p_0)=\sum_{p_n \in T}\omega(p_n)\cdot x(p_0+p_n+\Delta p_n) \tag{11}$$

where $\{\Delta p_n \mid n=1, ..., N\}$, $N=\lvert T\rvert$ and Δp is a decimal number.
Let $P_n = p_n + \Delta p_n$ then the deformable convolution formula can be expressed as

$$y(p_0)=\sum_{p_n \in T^*}\omega(p_n)\cdot x(p_0+P_n) \tag{12}$$

where $T^* =\{P_1, P_2, ..., P_n\}$, $N=\lvert T^*\rvert$ and it should be noted thatis a decimal point pair. Since the position after adding the offset is non-integer and does not correspond to the actual existing pixel points on the feature map, it is necessary to use interpolation to get the pixel values after the offset, which can usually be done by bilinear interpolation, expressed by the formula as follows

$$x(p)=\sum_{q}G(q,p)\cdot x(q) \tag{13}$$

where *q* is the coordinate on *x*, *G (,)* is the bilinear interpolation kernel and it is separated into two one-dimensional kernels as

$$G(q,p)=g(q_x,p_x)\cdot g(q_y,p_y) \tag{14}$$

where *g(a,b)* = max(0, |a,b|). p_n can be derived as

$$\frac{\partial y(p)}{\partial \Delta P_n}=\sum_{P_n \in T^*}\omega(P_n)\cdot\frac{\partial y(p_0+P_n)}{\partial P_n}=\sum_{P_n \in T^*}\left[\omega(P_n)\cdot\sum_{q}\frac{\partial G(q,p_0+P_n)}{\partial P_n}x(q)\right] \tag{15}$$

where p_0 is the coordinate in the output feature map and w(P_n) is the weight parameter in the convolution kernel.

Bidirectional feature pyramid network module

In the DeepLabv3+ decoder, only the shallow 1/4 feature map is utilized to fuse with the deep features. This method does not make full use of the features extracted at each stage and thus is less effective in segmenting small targets. To solve the above problem, this paper introduces the bidirectional feature pyramid network module (BiFPN) module (Lin *et al.*, 2017).

BiFPN is a neural network structure for target detection, which is based on the idea of a feature pyramid network (FPN) and bidirectional flow for improvement. In traditional FPNs, multi-scale feature pyramids are built by bottom-up and top-down directions to better handle objects of different sizes. However, this approach may lead to information loss or duplication and affect the model performance. To solve this problem, BiFPN introduces a bidirectional flow mechanism to optimize the feature pyramid network by using a structure consisting of two branches inside each layer: the top branch and the bottom branch. The BiFPN module feature fusion process is shown in Figure 3. As shown in the figure, the left side is a feature map of three different layers with smaller resolutions from bottom to top. The middle part is BiFPN, which upsamples the deep layer features, converts them to the size of the shallow layer feature map, and then fuses them with the shallow layer features. The right side is the feature map obtained after BiFPN, which contains not only the features of the deep layer but also the features of different levels. Here, the feature maps gener-

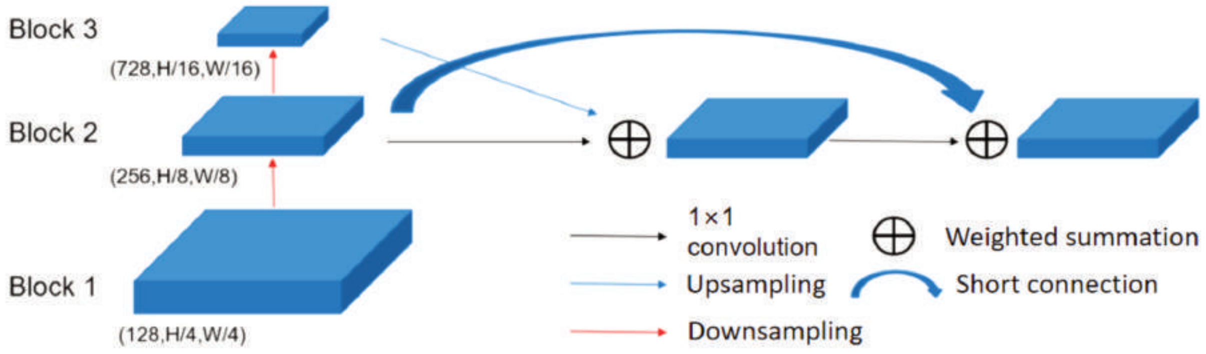


Figure 3. Simplified Bidirectional feature pyramid network feature fusion schematic.

ated by Block3 and Block2 in DeepLabv3+ backbone network Xception are fused, and the feature map sizes of Block3 and Block2 are 1/16 and 1/8 of the input image size and the number of channels are 728 and 256, respectively. In BiFPN, 1×1 convolution dimensionality reduction is performed on the feature maps in Block3 and Block2. The number of channels in Block3 is reduced from 728 to 128, and the number of channels in Block2 is reduced from 256 to 128. Finally, the feature maps of Block3 and Block2 are weighted and summed to obtain the FF map, which is then weighted and summed again with the reduced feature map of Block2 to obtain the final FF map. The FF map contains the features of both levels and has richer semantic and spatial information, which can improve the segmentation effect of the DeepLabv3+ network.

Semantic segmentation model based on improved DeepLabv3+ navel orange surface defects

Although the navel orange surface defect dataset is trained in the original DeepLabv3+ network, it can recognize all kinds of surface defects, but the recognition effect for small target defects and defect edge segmentation is general, and the recognition accuracy needs to be improved. To address the defects of the original DeepLabv3+ network, an improved DeepLabv3+ image segmentation model is proposed in this paper, and the overall structure is shown in Figure 4. CA mechanism is introduced in the encoder and decoder to capture the remote dependencies and retain the accurate position information. Meanwhile, deformable convolution is introduced into the ASPP structure to make the convolutional neural network more efficient in extracting the key semantic information of navel orange surface defects and improve the network's adaptability and fitting accuracy to irregular defect shape changes. In addition, to reduce the feature information loss caused by the deepening of the neural network, this paper introduces the BiFPN structure into the encoder to enhance the fusion of shallow and deep feature information, improve the learning ability of the model on the overall features, and reduce the leakage detection rate.

The encoder module has three outputs, the first one is the LFs output from Block1 in the backbone network, the second one is the FF from Block2 and Block3 output from BiFPN, and the last one is the HF extracted from the backbone network Xception embedded with CA mechanism input to the ASPP structure that introduces deformable The ASPP structure with convolution is sampled in parallel, and the obtained FF are fed into the convolution to obtain 256 channels of HFs. The HFs are first fused with the FF output from BiFPN of 1/8 size of the original image after 2-fold upsampling, and the obtained feature map is again fused with the underlying feature information of 1/4 size by up-sampling, and, finally, the output feature map is operated by attention mechanism and down-sampling, and the output feature map is upsampled by 4-fold to obtain the predicted segmented image.

Datasets and evaluation metrics

Datasets

The image dataset of this paper was obtained from Jiangxi REEMOON Technology Holdings Company. The dataset was collected by placing multiple navel oranges on a row of 360-degree rotatable trays, taking a panoramic photograph containing all the navel oranges at specific intervals, and cropping the panoramic photographs at a later stage to close-up photographs containing only individual navel oranges. We collected 5,290 images of vari-

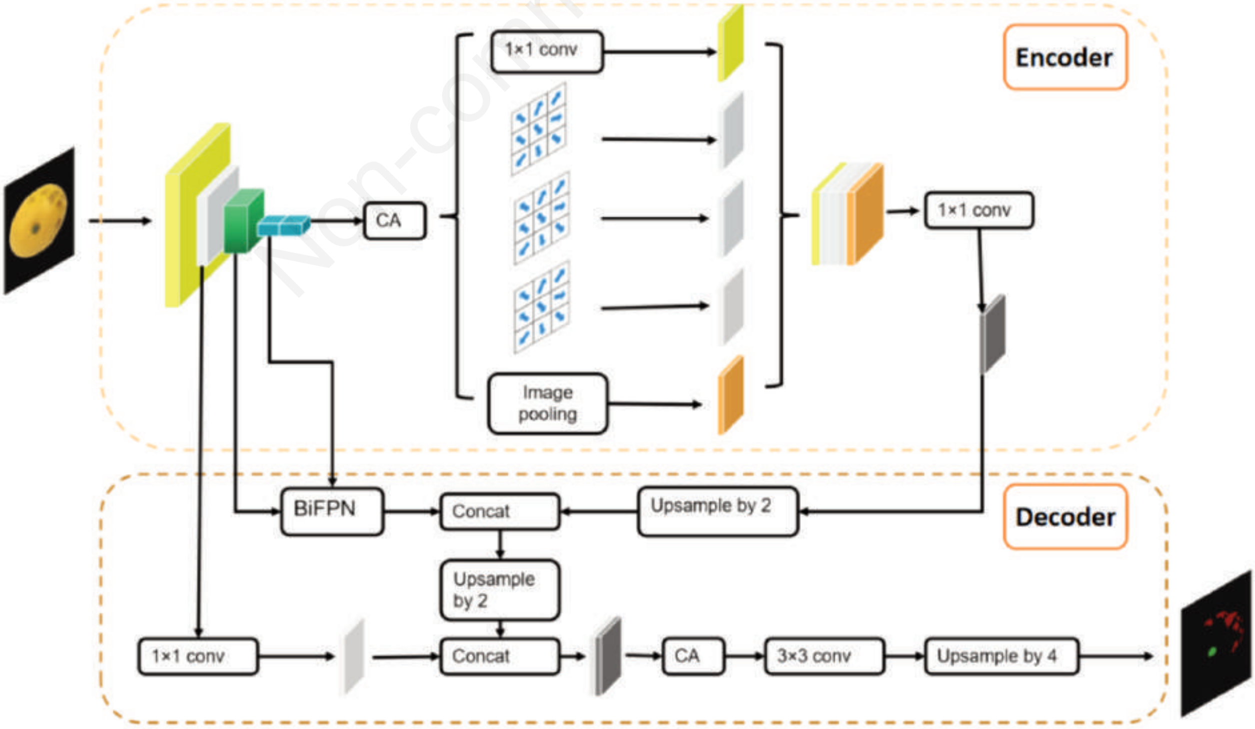


Figure 4. Improved DeepLabv3+ architecture diagram.

ous navel orange defects, which were divided into training, test, and validation sets according to a ratio of 8:1:1, with 4,232 images in the training set and 529 images each in the validation and test sets. The data set was divided into six categories: rotten, navel deformation, mild pitting, severe pitting, severe oil spotting, and background, and some samples of umbilical orange surface defect data set are shown in Figure 5.

To train a deep learning network for supervised image classification and detection tasks, after constructing the navel orange surface defect dataset, this paper uses the open-source tool Labelme to label the dataset. After completing the image annotation, the tool generates a JavaScript object notation (JSON) file with the same name. This JSON file contains information such as the name of the original image, the name of the defect, and the coordinates of the mouse clicks used to generate the defect boundary. The original image and the corresponding Mask annotated image of the dataset in this paper are shown in Figures 6 and 7, and Table 1 shows the number of images owned by each category.

Evaluation indicators

In image segmentation tasks, accuracy is one of the most dominant and popular technical metrics for evaluating model performance. In general, we can classify the accuracy estimation methods into two categories: based on pixel accuracy and based on intersection over union (IoU). Assuming a total of k+1 categories (labeled as L_0 to L_k, which contains a background category), p_{ij} denotes the number of pixels with category i predicted as category j. In this way P_{ii} denotes true positives, P_{ij} and P_{ji} denote false positives and false negatives, respectively.

Pixel accuracy (PA) represents the ratio of the total number of pixels to the predicted correct pixels, and is expressed as:

$$PA = \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{16}$$

Mean pixel Accuracy (mPA) is the average value obtained by summing the total number of correct pixels for each category with

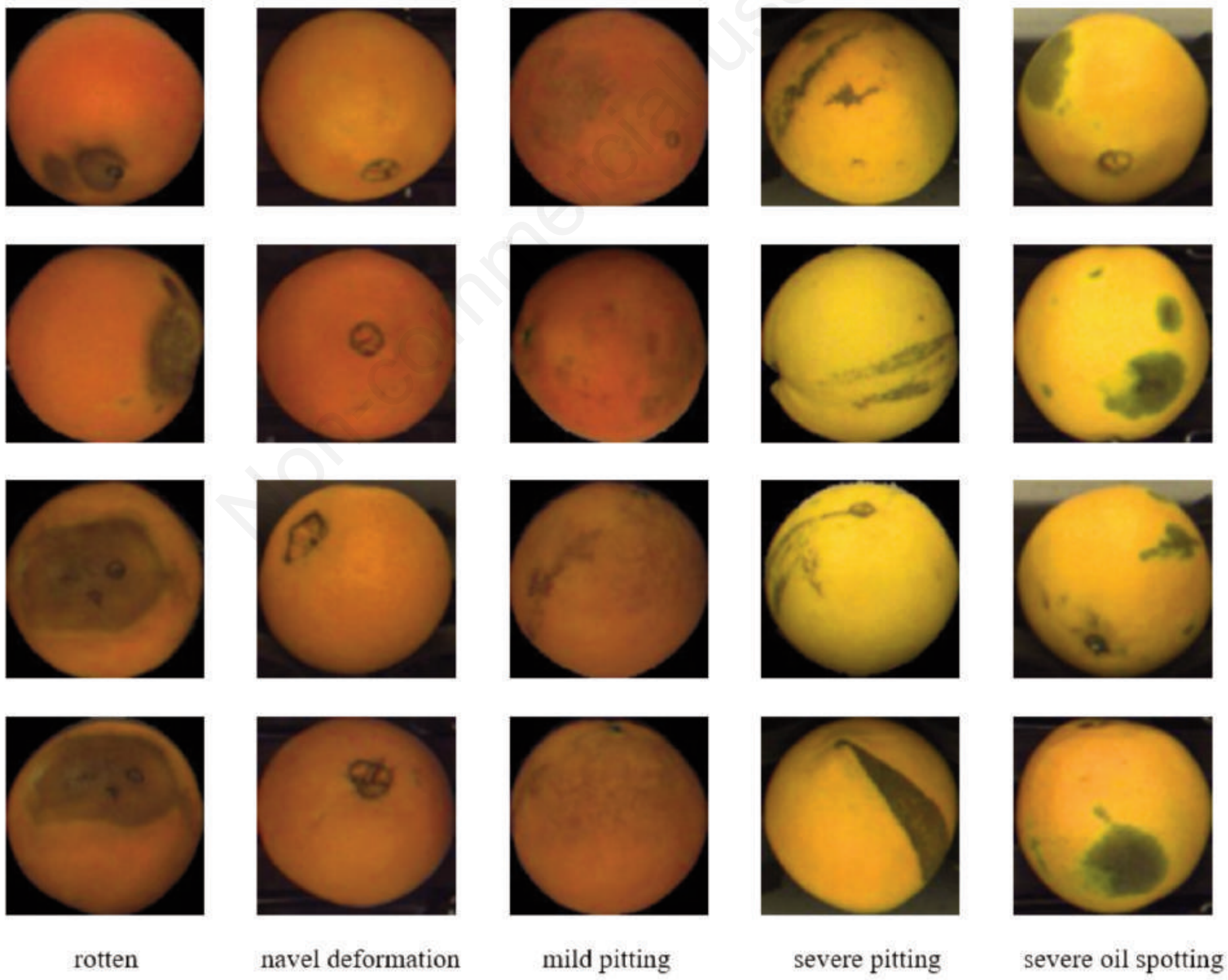


Figure 5. Example of some samples from the navel orange surface defects dataset.

the total contrast ratio for each category, as expressed by

$$mPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}} \tag{17}$$

Mean IoU (mIoU) is a commonly used metric to evaluate the performance of computer vision tasks such as target detection and semantic segmentation. It measures the model prediction accuracy by calculating the degree of overlap between the predicted result and the region of real labels. The specific expression is as follows:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{18}$$

Precision indicates how many of the pixels predicted by the model to be in a particular category are correct, while mean precision (mPrecision) is the average of the precision rates of all the categories, which is given by the following formula:

$$mPrecision = \frac{1}{k+1}\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{j=0}^{k}p_{ji}} \tag{19}$$

Results

Experimental setup

All the processes of training and testing the model described in this work were implemented on one machine, whose configuration parameters were Intel Corei7-7700 3.60 GHz Processor, a NVIDIA GTX 1060 GPU, and 6GB memory. The model in this work was trained in a 64-bit Windows 10 environment using Pytorch 1.8 and CUDA 10.2.

The model is trained by stochastic gradient descent with momentum, which is set to 0.9. The “poly” learning strategy is adopted, and the base learning rate is set to 0.007 as the number of iterations increases, and the input image size is cropped to 144×144. The value of weight decay used to prevent overfitting is 0.0005, the loss function adopts the cross-entropy loss function, and the step size OS is 16, the batch size is 8, and the number of iterations epoch is set to 300, considering the problem of limited video memory resources. Figure 8 shows the training accuracy curves of the improved DeepLabv3+ algorithm. To illustrate the stability of the results, in addition to data not related to training, we give in parentheses after the training results the standard deviation obtained after the model has been trained three times.

Ablation experiments

To verify the performance of each module of this method, Xception is used as the backbone network, and the feature map with an output stride of 16 is extracted. On the navel orange surface defect dataset, the ablation is performed by adding the CA module, introducing deformable convolution in the ASPP module, and adding the BiFPN module, respectively. The number of itera-

Table 1. Number of images owned per category.

	Train set	Test set	Total
Category	4232	529	4761
Rotten	1889	257	2145
Navel deformation	2034	302	2336
Mild pitting	1659	285	1944
Severe pitting	1845	276	2121
Severe oil spotting	1738	268	2006

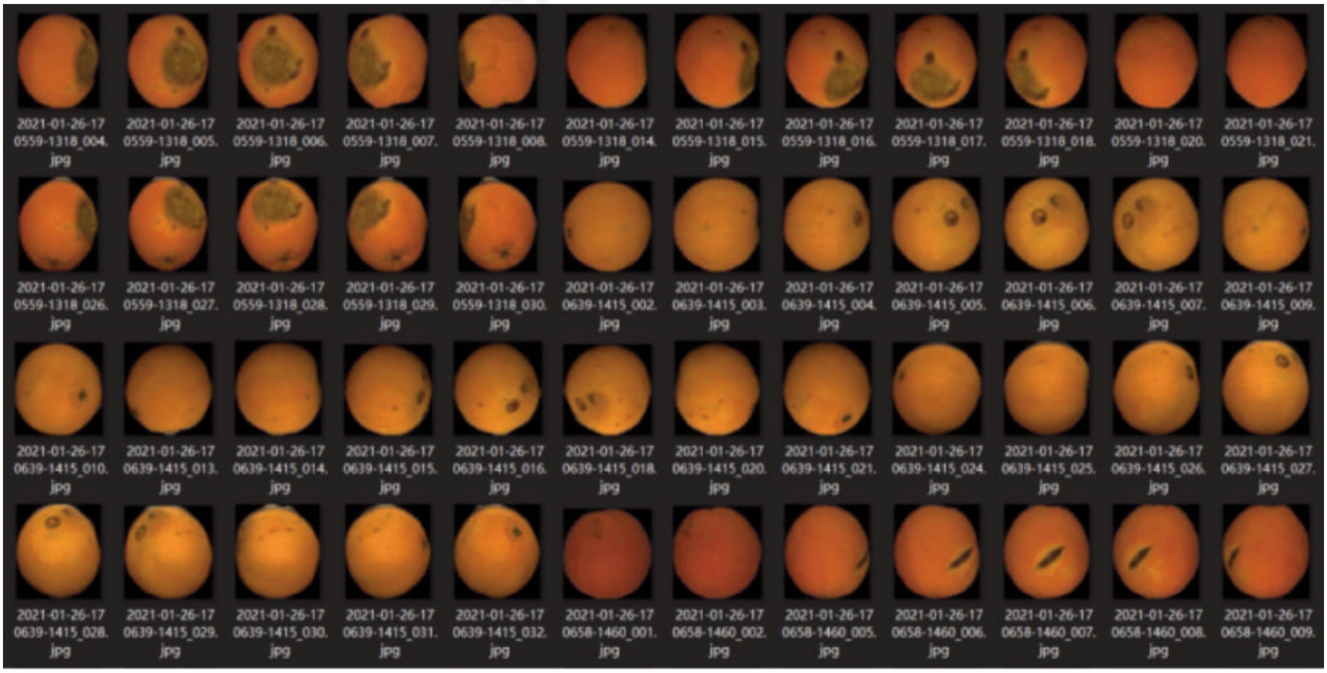

2021-01-26-17 0559-1318_004. jpg
2021-01-26-17 0559-1318_005. jpg
2021-01-26-17 0559-1318_006. jpg
2021-01-26-17 0559-1318_007. jpg
2021-01-26-17 0559-1318_008. jpg
2021-01-26-17 0559-1318_014. jpg
2021-01-26-17 0559-1318_015. jpg
2021-01-26-17 0559-1318_016. jpg
2021-01-26-17 0559-1318_017. jpg
2021-01-26-17 0559-1318_018. jpg
2021-01-26-17 0559-1318_020. jpg
2021-01-26-17 0559-1318_021. jpg
2021-01-26-17 0559-1318_026. jpg
2021-01-26-17 0559-1318_027. jpg
2021-01-26-17 0559-1318_028. jpg
2021-01-26-17 0559-1318_029. jpg
2021-01-26-17 0559-1318_030. jpg
2021-01-26-17 0639-1415_002. jpg
2021-01-26-17 0639-1415_003. jpg
2021-01-26-17 0639-1415_004. jpg
2021-01-26-17 0639-1415_005. jpg
2021-01-26-17 0639-1415_006. jpg
2021-01-26-17 0639-1415_007. jpg
2021-01-26-17 0639-1415_009. jpg
2021-01-26-17 0639-1415_010. jpg
2021-01-26-17 0639-1415_013. jpg
2021-01-26-17 0639-1415_014. jpg
2021-01-26-17 0639-1415_015. jpg
2021-01-26-17 0639-1415_016. jpg
2021-01-26-17 0639-1415_018. jpg
2021-01-26-17 0639-1415_020. jpg
2021-01-26-17 0639-1415_021. jpg
2021-01-26-17 0639-1415_024. jpg
2021-01-26-17 0639-1415_025. jpg
2021-01-26-17 0639-1415_026. jpg
2021-01-26-17 0639-1415_027. jpg
2021-01-26-17 0639-1415_028. jpg
2021-01-26-17 0639-1415_029. jpg
2021-01-26-17 0639-1415_030. jpg
2021-01-26-17 0639-1415_031. jpg
2021-01-26-17 0639-1415_032. jpg
2021-01-26-17 0658-1460_001. jpg
2021-01-26-17 0658-1460_002. jpg
2021-01-26-17 0658-1460_005. jpg
2021-01-26-17 0658-1460_006. jpg
2021-01-26-17 0658-1460_007. jpg
2021-01-26-17 0658-1460_008. jpg
2021-01-26-17 0658-1460_009. jpg

Figure 6. Raw image data for the detection of surface defects on navel oranges.

tions is set to 300 rounds, and after the model converges, the output visualization results are shown in Figure 9. It can be found from Figure 9 that various improvement strategies can improve the model performance to some extent, but the model detection effect is the best after considering various improvement strategies together. The specific evaluation indexes are shown in Table 2.

As can be seen from Table 2, the introduction of the CA mechanism improves the segmentation accuracy by 0.77% and improves the mPA index by 1.59%, indicating that the module can enhance the feature expression ability, effectively capture the relationship between location information and channel information, and use the key information of the feature map to make the features of the important regions in the image more by the weighted summation operation. To improve the segmentation accuracy of navel orange surface defects by using the key information of the feature map to make the features of important regions in the image more obvious through the weighted summation operation; Then replace the dilated convolution in ASPP structure with deformable dilated convolution, the accuracy is improved by 1.14%, while the average pixel accuracy is improved by 1.78%, which proves that deformable convolution can extract precisely for different scales and irregular shape defects, and improve the adaptability and fitting accuracy of the network to irregular defect shape changes; plus the fusion of features of different levels using BiFPN, the accuracy improved by 1.9% and mPA value improved by 1.92%, indicating that the module effectively fuses features at different levels with richer semantic and spatial information to improve the model's ability to extract features of small target objects, and also makes the segmentation results more accurately localized, while the edge details are further optimized.

Model performance comparison experiments

To verify the effectiveness of the improved algorithm, we conducted experiments on the navel orange defect dataset and compared it with models such as PSPNet, UNet, BiSeNetV2 (Yu *et al.*, 2021), Segmenter (Xie *et al.*, 2021), SegFormer (Strudel *et al.*, 2021) and DeepLabv3+ were compared. The above models were trained using the same dataset and the segmentation results under different segmentation networks are shown in Table 3. From the experimental results, it can be seen that the improved model reaches the highest value in mIoU, mPA, and mPrecision, which further indicates that the proposed model possesses the best segmentation performance among all models.

In order to better evaluate the computational cost of the model, we list the specific values of the number of parameters (Params), Giga floating-point operations per second (GFLOPs), and frames per second (FPS) of the proposed model and compare them with

Table 2. Comparison of test results of different improvement options for DeepLabv3+.

Experiment	Different network structures			mIoU (%)	mPA (%)	mPrecision (%)
	Coordinate attention	Deformable convolution	BiFPN			
1	-	-	-	72.51 (±0.06)	81.09 (±0.08)	88.95 (±0.03)
2	√	-	-	74.28 (±0.04)	82.68 (±0.11)	89.68 (±0.09)
3	√	√	-	75.42 (±0.04)	84.46 (±0.10)	90.41 (±0.08)
4	√	√	√	77.32 (±0.05)	86.38 (±0.12)	91.34 (±0.06)

mIoU, mean intersection over union; mPA, mean pixel accuracy; mPrecision, mean precision; BiFPN, Bidirectional feature pyramid network.

Table 3. Comparison of the performance of different models on the navel orange defect test set.

Model	Backbone network	Training weight settings	mIoU (%)	mPA (%)	mPrecision (%)
Unet	VGG-16	Fine-tuning	72.70 (±0.05)	80.23 (±0.15)	89.15 (±0.05)
PSPNet	ResNet101	Fine-tuning	67.95 (±0.03)	77.32 (±0.12)	81.94 (±0.07)
DeepLabv3+	Xception	Fine-tuning	73.51 (±0.06)	81.09 (±0.08)	91.09 (±0.03)
BiSeNetV2	-	From scratch	52.51 (±0.05)	69.95 (±0.10)	77.32 (±0.07)
Segmenter	VIT-S	Fine-tuning	71.24 (±0.06)	83.12 (±0.07)	87.87 (±0.08)
SegFormer	MIT-B2	Fine-tuning	73.82 (±0.07)	85.36 (±0.09)	88.67 (±0.09)
Improved DeepLabv3+	Xception	Fine-tuning	77.32 (±0.05)	86.38 (±0.12)	91.34 (±0.06)

mIoU, mean intersection over union; mPA, mean pixel accuracy; mPrecision, mean precision.

Table 4. Comparison of Params, GFLOPs, and FPS for different models.

Model	Backbone network	Params (M)	GFLOPs (G)	Latency (RTX3060)/FPS (512*512)
Unet	VGG-16	30.92	274.53	13(±1)
PSPNet	ResNet101	46.60	179.23	19(±1)
DeepLabv3+	Xception	54.71	83.44	33(±2)
BiSeNetV2	-	3.35	12.30	109(±2)
Segmenter	VIT-S	25.98	37.39	50(±1)
SegFormer	MIT-B2	24.73	25.26	37(±3)
Improved DeepLabv3+	Xception	56.11	90.56	31(±2)

GFLOPs, giga floating-point operations per second; FPS, frames per second.

other models, as shown in Table 4. From the results, we can find that although the Params and GFLOPs of the proposed model are much larger than those of BiSeNetV2, and the FPS is also much different, considering that in terms of detection performance, the proposed model outperforms BiSeNetV2 by 46.69%, 23.49%, and 18.13% in mIoU, mPA, and mPrecision, respectively, which is a very significant improvement. From the practical application point of view, the proposed model is more suitable for navel orange surface defect detection. Moreover, numerically, besides BiSeNetV2, Segmenter can also be said to be a computationally inexpensive and fast inference model, which is 31.13M and 53.17G lower than the proposed model in terms of Params and GFLOPs, respectively, and 19 higher than the proposed model in terms of FPS. However, Segmenter is 6.08%, 3.26%, and 3.47% lower than the improved model in terms of mIoU, mPA, and mPrecision, respectively. This suggests that although the improved model has some disadvantages in terms of computational cost, the advantages brought by its detection performance can compensate for this. In terms of commercial value, the proposed model can bring more practical benefits compared to other models.

Comparison of segmentation accuracy for different defect types

In order to better measure the segmentation performance of the

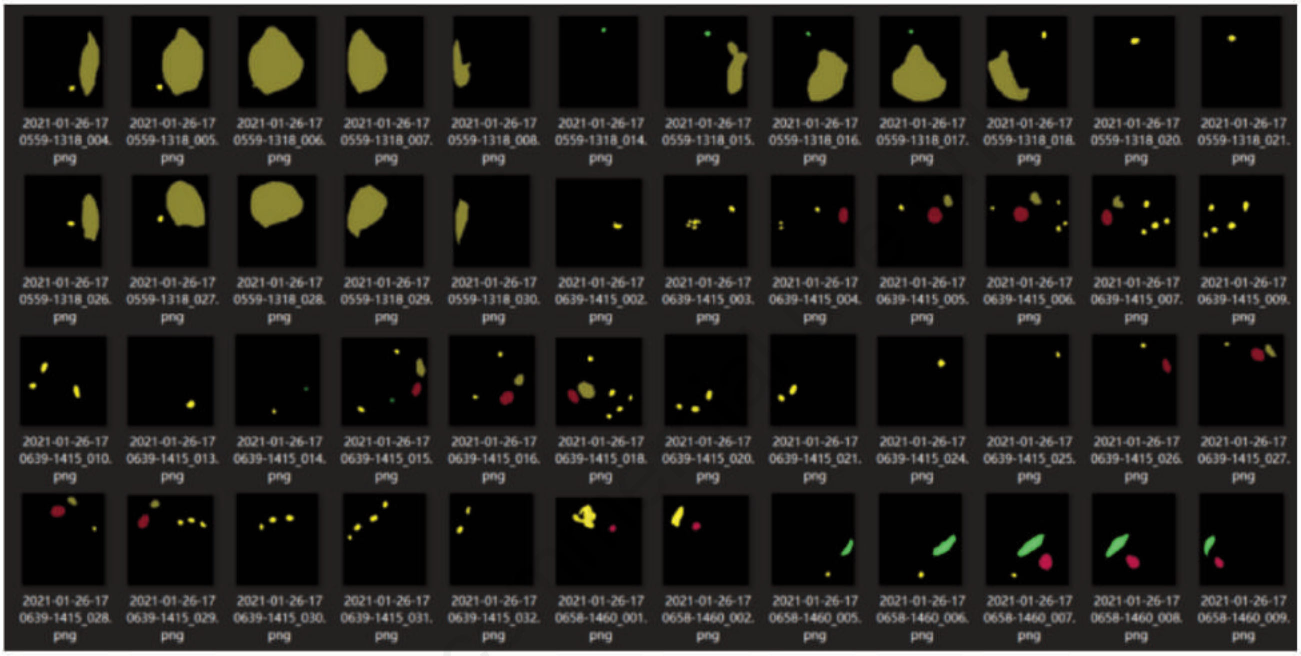

2021-01-26-17 0559-1318_004.png
2021-01-26-17 0559-1318_005.png
2021-01-26-17 0559-1318_006.png
2021-01-26-17 0559-1318_007.png
2021-01-26-17 0559-1318_008.png
2021-01-26-17 0559-1318_014.png
2021-01-26-17 0559-1318_015.png
2021-01-26-17 0559-1318_016.png
2021-01-26-17 0559-1318_017.png
2021-01-26-17 0559-1318_018.png
2021-01-26-17 0559-1318_020.png
2021-01-26-17 0559-1318_021.png
2021-01-26-17 0559-1318_026.png
2021-01-26-17 0559-1318_027.png
2021-01-26-17 0559-1318_028.png
2021-01-26-17 0559-1318_029.png
2021-01-26-17 0559-1318_030.png
2021-01-26-17 0639-1415_002.png
2021-01-26-17 0639-1415_003.png
2021-01-26-17 0639-1415_004.png
2021-01-26-17 0639-1415_005.png
2021-01-26-17 0639-1415_006.png
2021-01-26-17 0639-1415_007.png
2021-01-26-17 0639-1415_009.png
2021-01-26-17 0639-1415_010.png
2021-01-26-17 0639-1415_013.png
2021-01-26-17 0639-1415_014.png
2021-01-26-17 0639-1415_015.png
2021-01-26-17 0639-1415_016.png
2021-01-26-17 0639-1415_018.png
2021-01-26-17 0639-1415_020.png
2021-01-26-17 0639-1415_021.png
2021-01-26-17 0639-1415_024.png
2021-01-26-17 0639-1415_025.png
2021-01-26-17 0639-1415_026.png
2021-01-26-17 0639-1415_027.png
2021-01-26-17 0639-1415_028.png
2021-01-26-17 0639-1415_029.png
2021-01-26-17 0639-1415_030.png
2021-01-26-17 0639-1415_031.png
2021-01-26-17 0639-1415_032.png
2021-01-26-17 0658-1460_001.png
2021-01-26-17 0658-1460_002.png
2021-01-26-17 0658-1460_005.png
2021-01-26-17 0658-1460_006.png
2021-01-26-17 0658-1460_007.png
2021-01-26-17 0658-1460_008.png
2021-01-26-17 0658-1460_009.png

Figure 7. Segmented label dataset for navel orange surface defect detection.

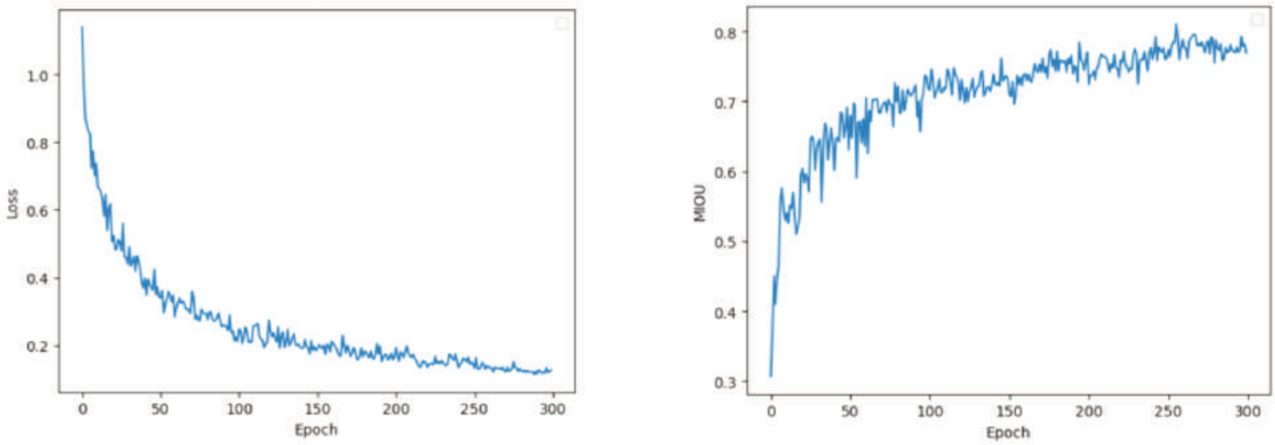


Figure 8. Improving DeepLabV3+ loss and mean intersection over union variation curves.

models for different defects, the segmentation effectiveness of different models for different navel orange defects was also compared. As shown in Table 5, the proposed model achieves the best IoU values for all categories of navel orange defects compared to other models. This indicates that the model is more capable of learning features for different classes of defects and has the best segmentation performance compared to PSPNet, UNet, BiSeNetV2, Segmenter, SegFormer, and DeepLabv3+.

In order to visualize the performance difference between this method and other methods, we show the segmentation results for the navel orange defective dataset in Figure 10. In order to better illustrate the importance of mIoU, mPA, and mPrecision on the defect detection effect, we do not consider BiSeNet here, which has too large a difference in the values of the evaluation metrics with the other models, and only show PSPNet, Segmenter, Unet, DeepLabv3+, SegFormer, Improved DeepLabv3+ visualization results on the navel orange defect test set and analyze them with the values of their evaluation metrics.

PSPNet may lead to loss of information due to the pooling operation with fixed size only, so its mPA and mPrecision are low, and the probability of detecting defects and detection accuracy are not strong, basically, it can only segment the approximate shape of the defects on the surface of navel oranges. Segmenter allows the global context to be modeled in the first layer and throughout the network, so there is a large increase in the ability to capture information, and many small target defects can be detected, and the mIoU increases by 3.29%, but there are more cases of misidentification as can be seen from the figure, especially the harder to distinguish pitting type defects. UNet uses skip connections, which can be used to fill in missing information using LF, improving defect recognition to some extent. mPrecision improves by 1.28% over Segmenter, but not all skip connections have a positive effect. DeepLabv3+ introduces ASPP and feature fusion module, which enhances the ability to learn different defective features, so

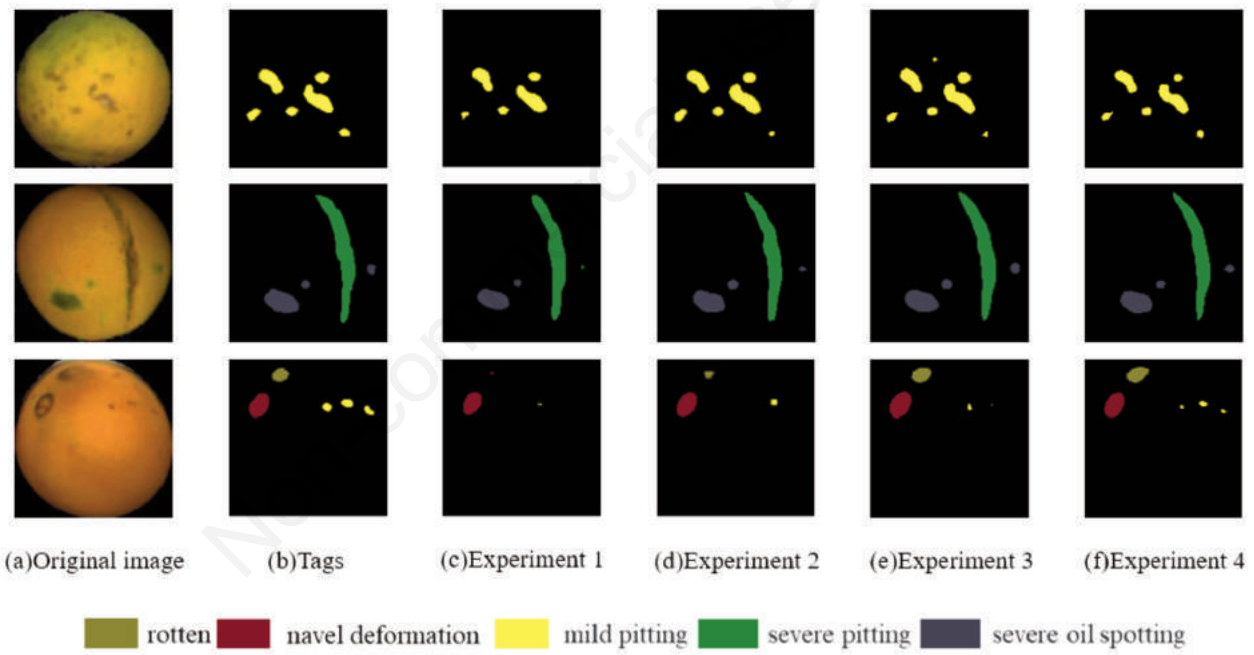


Figure 9. Segmentation results from different experiments.

Table 5. Comparison of the performance indicators of the models for different navel orange surface defects.

Evaluation indicators	Model	Rotten	Navel deformation	Mild pitting	Severe pitting	Severe oil spotting
IoU (%)	Unet	71.32 (±0.05)	72.16 (±0.06)	65.53 (±0.07)	74.90 (±0.02)	79.74 (±0.04)
	PSPNet	71.76 (±0.04)	69.22 (±0.05)	60.31 (±0.09)	70.41 (±0.03)	68.05 (±0.06)
	DeepLabv3+	71.24 (±0.05)	74.86 (±0.07)	70.60 (±0.08)	72.31 (±0.02)	78.54 (±0.08)
	BiSeNetV2	57.75 (±0.05)	65.13 (±0.09)	26.47 (±0.06)	37.85 (±0.04)	75.36 (±0.07)
	Segmenter	71.26 (±0.07)	53.82 (±0.08)	87.13 (±0.04)	60.85 (±0.06)	82.51 (±0.03)
	SegFormer	74.18 (±0.09)	83.55 (±0.11)	56.69 (±0.06)	65.66 (±0.06)	86.80 (±0.08)
	Improved DeepLabv3+	76.89 (±0.04)	78.05 (±0.04)	74.19 (±0.12)	75.74 (±0.02)	81.73 (±0.06)

IoU, intersection over union.

mPrecision is higher than that of Segmenter and Unet, and more defective regions can be recognized while the recognition accuracy is also improved. SegFormer's mIoU is similar to DeepLabv3+, with the difference that its detection focuses on the ability to detect defects, while there is a gap in accuracy with DeepLabv3+. In contrast, the improved DeepLabv3+ has the strongest defect detection capability, with better detection of small spots with minor and severe flaking, and segmented defect edges that are more similar to the original label. In addition, from the detection effect, the segmentation error of navel orange defects is mainly divided into two cases. The first case is the missed or wrong detection of the edge area of the defects, as well as the complete omission of small areas of defects. The second case is the wrong judgment of the defect type. In these two cases, we can improve the model's ability to extract locally important information as a way to improve the segmentation accuracy.

In addition, from the segmentation effect graphs of the improved DeepLabv3+ shown in Figures 9 and 10, we find that the presentation of umbilical orange skin defects also affects the segmentation effect of the model. For defects with darker color or obvious outer contour features, such as severe oil spots and navel deformation, the segmentation of the defective parts is relatively simple, and thus the defective segmentation region has the highest overlap rate with the manually labeled region, such as the severe oil spotting (gray part) in the second figure of experiment 4 in Figure 9, and the navel deformation (red part) in the third figure. As for the defects with unclear outer contours, such as rotting and severe pitting, the segmentation effect has a certain gap compared with the first two defects because the segmentation area becomes irregular and the segmentation is more difficult, as shown in the rotten (brown part) of the third graph of experiment 4 in Figure 9 and the severe pitting (green part) of the third graph of Improved DeepLabv3+ in Figure 10. Finally, for mild pitting, the defective part is lighter in color and irregular in shape, which not only makes it difficult to distinguish from normal fruit epidermis, but also makes the defective region more difficult to be segmented, and thus the difference between the segmented region of the defect and the manually labeled region is more obvious, as shown in mild pitting (yellow part) in the third graph of Improved DeepLabv3+ in Figure 10.

Conclusions

This paper introduces the mainstream semantic segmentation network DeepLav3+, and analyzes the shortcomings and deficiencies of the original model applied to the navel orange surface defect dataset. Considering the features of navel orange surface defects, the CA attention mechanism is embedded into the DeepLabv3+ network with better semantic segmentation performance in order to strengthen the feature extraction ability of the network and reduce feature loss. Meanwhile, the Atrous Convolution of ASPP structure is replaced with deformable Atrous Convolution to improve the network's adaptability to irregular defect shape changes and fitting accuracy. In addition, the BiFPN-based feature fusion branch is introduced at the DeepLabv3+ encoder end to realize multi-scale feature fusion and enrich the feature space and semantic information. Through experiments, the proposed model effectively improves the accuracy of navel orange surface defect detection compared with the DeepLabv3+ algorithm on the navel orange surface defect test set.

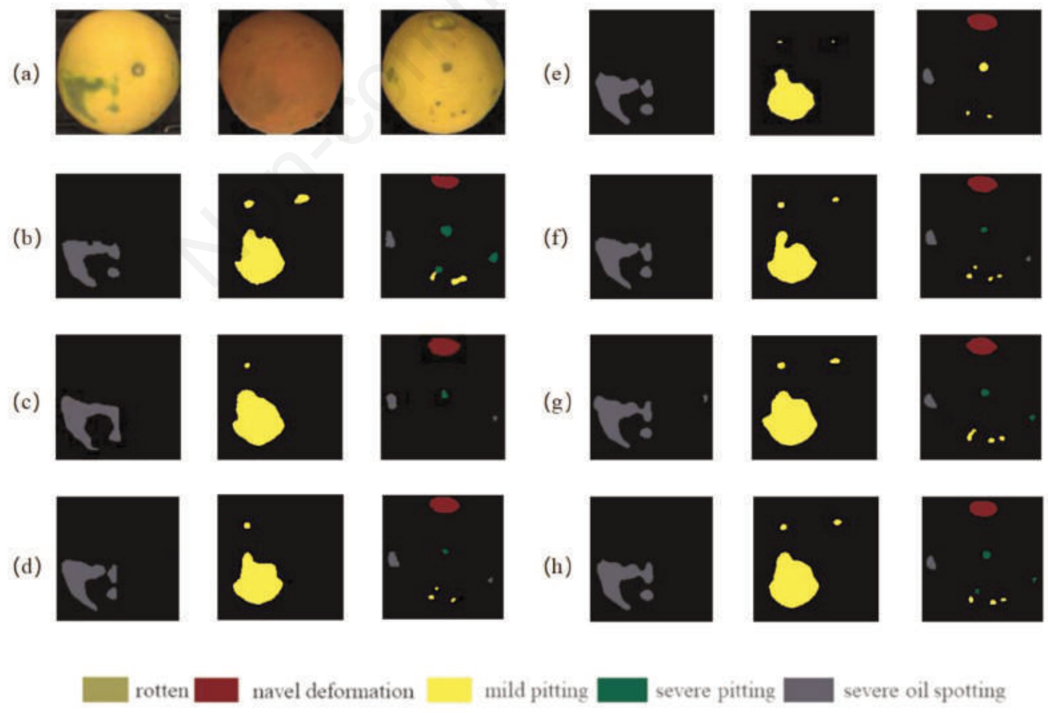


Figure 10. Graph of predicted results from different models. a) Original; b) tag; c) PSPNet; d) Segmenter; e) Unet; f) DeepLabv3+; g) SegFormer; h) improved DeepLabv3+.

References

Bhargava, A., Bansal, A. Automatic detection and grading of multiple fruits by machine learning. 2020. Food Anal. Methods 13:751-61.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV). pp. 801-18.

Fan, S., Li, J., Zhang, Y., Tian, X., Wang, Q., He, X., Zhang, C., Huang, W. 2020. Online detection of defective apples using computer vision system combined with deep learning methods. J. Food Eng. 286:110102.

Hou, Q., Zhou, D., Feng, J. 2021. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13713-22.

Hu, J., Shen, L., Sun, G. 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132-41.

Jin, L.I., Renyong, Z.H.A.O., Boxue, D.U., Liucheng, H., Kai, B. 2021. Research progress of nondestructive detection methods for defects of electrical epoxy insulators. Transact. Chin. Electrotechn. Soc. 36:4598-607.

Li J.B., Huang W.Q., Zhao C.J. 2015. Machine vision technology for detecting the external defects of fruits - A review. Image. Sci. J. 63:241-51.

Liang, X., Jia, X., Huang, W., He, X., Li, L., Fan, S., Li, J., Zhao, C., Zhang, C. 2022. Real-time grading of defect apples using semantic segmentation combination with a pruned YOLO V4 Network. Foods. 11:3150.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. 2017. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117-25.

Nithya, R., Santhi, B., Manikandan, R., Rahimi, M., Gandomi, A.H. 2022. Computer vision system for mango fruit defect detection using deep convolutional neural network. Foods. 11:3483.

Raman, S., Chougule, A., Chamola, V. 2022. A low power consumption mobile based IoT framework for real-time classification and segmentation for apple disease. Microprocess. Microsyst. 94:104656.

Ren, H.-E., Bai, J.-Y. 2013. Color grading based on dynamic clustering in lab color space. Jisuanji Gongcheng/Comput. Eng. 39.

Rong, D., Rao, X., Ying, Y. 2017. Computer vision detection of surface defect on oranges by means of a sliding comparison window local segmentation algorithm. Comput. Electron. Agr. 137:59-68.

Ronneberger, O., Fischer, P., Brox, T. 2015. U-net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing. pp. 234-41

Soltani Firouz, M., Sardari, H. 2022. Defect detection in fruit and vegetables by using machine vision systems and image processing. Food Eng. Rev. 14:353-79.

Strudel, R., Garcia, R., Laptev, I., Schmid, C. 2021. Segmenter: transformer for semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262-72.

Sun, X., Li, G., Xu, S. 2020. Fastidious attention network for navel orange segmentation. arXiv preprint arXiv:2003.11734.

Tian, K., Zeng, J., Song, T., Li, Z., Evans, A. Li, J. 2022. Tomato leaf diseases recognition based on deep convolutional neural networks. J. Agric. Eng. Res. 54.

Unay, D. 2022. Deep learning-based automatic grading of bi-colored apples using multispectral images. Multimed. Tools Appl. 81:38237-52.

Woo, S., Park, J., Lee, J.Y., Kweon, I.S. 2018. Cbam: Convolutional block attention module. Proceedings of the European Conference on Computer Vision. pp. 3-19.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. Luo, P. 2021. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neur. In. 34:12077-90.

Xie, X., Ge, S., Xie, M., Hu, F., Jiang, N., Cai, T., Li, B. 2018. Image matching algorithm of defects on navel orange surface based on compressed sensing. J. Amb. Intel. Hum. Comp. pp. 1-9.

Yang, G.L., Luo, L., Feng, Y.Q., Zhao, H.S. 2014. Research of navel orange defect and color detection based on machine vision. Appl. Mech. Mater. 513:3442-5.

Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., Li, X. 2021. A real-time detection algorithm for Kiwifruit defects based on YOLOv5. Electronics. 10:1711.

Yu, C., Gao, C., Wang, J., Yu, G., Shen, C. Sang, N. 2021. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. Int. J. Comput. Vision. 129:3051-68.

Zhang, B., Huang, W., Gong, L., Li, J., Zhao, C., Liu, C., Huang, D. 2015. Computer vision detection of defective apples using automatic lightness correction and weighted RVM classifier. J. Food Eng. 146:143-51.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. 2017. Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881-90.

Zhou, H., Zhuang, Z., Liu, Y., Liu, Y., Zhang, X. 2020. Defect classification of green plums based on deep learning. Sensors. 20:6993.