2023 IEEE 6th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) | 979-8-3503-0562-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/AUTEEE60196.2023.10408693

Using Improved DeepLabV3+ for Complex Scene Segmentation

Yuchen Wu[#]
Anhui University of Science and Technology
Huainan, China

Jin Li[#]
Hangzhou Dianzi University
Hangzhou, China

Junkai Yang[#,*]
Beijing Jiaotong University
Beijing, China
21722036@bjtu.edu.cn
[#]These authors contributed equally.

***Abstract*—This paper presents a lightweight convolutional neural network model based on DeepLabV3+. By collecting and annotating an image dataset, sufficient sample support is provided for the model's training. In order to enhance the model's training speed and applicability, this study proposes replacing the backbone network Xception with MobileNetv2. This successfully reduces the number of model parameters, making the training process more rapid and efficient. Additionally, by refining the model structure and incorporating Coordinate Attention (CA) and Cascaded Feature Fusion (CFF) for feature integration, the feature map processing is further optimized. This allows the model to more accurately identify in complex background environments. Verification shows that the improved model performs excellently in metrics such as mIoU, mPA, mPrecision, and mRecall. During testing, mIoU reached 82.66%, mPA reached 88.20%, mPrecision reached 92.26%, and mRecall reached 87.67%. These results fully demonstrate the feasibility, accuracy, and robustness of the proposed method, providing strong support in practical applications.**

Keywords—DeepLabV3+, CA Attention Mechanism, CFF Feature Fusion Module, MobileNetv2

I. INTRODUCTION

With the rapid development of deep learning, significant breakthroughs have been made in the field of semantic segmentation [1,2]. Compared with the traditional method based on hand-designed features, deep convolutional neural networks are able to learn richer features from the original image through multi-level convolutional operations, and are no longer limited by feature representation ability and generalization ability, which greatly improves the accuracy and robustness of semantic segmentation. FCN (Fully Convolutional Networks) achieves pixel-level prediction for the first time by replacing the fully connected layer with a fully convolutional layer[3], and the use of this architecture allows the network to directly output semantic segmentation results of an image, leading to important applications in computer vision tasks. And then, Ronneberger et al. proposed U-Net in 2015.As a network based on encoding-decoding structure, U-Net realizes high-resolution semantic segmentation through the combination of encoder and decoder. The U-Net++ and U-Net+++, which are based on it, are widely used in processing with the advantage of high accuracy. Presented by the Google team in 2016, DeepLab is a model that utilizes atrous convolution to expand the sensory field and uses atrous spatial pyramid pooling to capture more contextual information[4,5], and is therefore able to correctly classify objects in an image at the pixel level. Since it was proposed by Google in 2017, MobileNet can be regarded as the Inception of lightweight networks, which significantly reduces computation and storage requirements while maintaining high accuracy, and is suitable for image classification and target detection tasks in resource-constrained environments such as mobile devices. It realizes efficient feature extraction and model compression through techniques such as depth-separable convolution and width multiplication, with small model size and low computational complexity. The above networks have not only achieved remarkable results in academia, but have also been widely used and validated in practical applications[6].

Meanwhile, through the efforts of many scholars, a series of network optimization schemes such as Attention Mechanism, Multi-Supervision Mechanism, Data Enhancement, Weight Decay, etc. have been successively raised. Attention Mechanism makes the model pay more attention to important information by assigning different attention weights to different parts of the input data. This mechanism can help the model to better handle long sequential data or handle tasks with variable length inputs. Multi-supervised mechanisms train models to provide richer information by using supervised signals from multiple labels or multiple tasks simultaneously[7]. And this paper presents a network based on DeepLabV3+ and replaces the backbone from the original Xception to MobileNetv2. As a result, the number of model parameters plummets, which makes the model run more efficiently, converge faster, and be more suitable for the mechanical tomato picking task. Also because of the use of Coordinate Attention on the final downsampled feature maps[8], the model automatically assigns importance to different channels and focuses on spatial location information. In addition, this network introduces a CFF (Cascade Feature Fusion) structure as a contextual information fusion module and takes feature maps with different resolutions and combines them through up-sampling and addition or other operations to achieve the purpose of fusing semantic information from different layers of the network without loss of spatial resolution[9].

In the current study, datasets from different sources were selected and processed in a uniform manner, and a relatively comprehensive dataset was compiled after labeling them one by one. It contains image data with different model, different numbers, different weather (rainy, cloudy, sunny), and different backgrounds. Using this dataset to train the model increases the robustness and generalization ability of the model to be suitable for dealing with the recognition and segmentation [10]. Due to the use of the homemade dataset, this study can target hyper-parameter tuning during the training process of the model, and select appropriate optimizers and learning strategies. Data augmentation

methods can also be used in a targeted manner to finalize the goal of enhancing the segmentation accuracy. In summary, this is achieved through the use of homemade datasets, the adoption of MobileNet_V2 as a backbone in DeepLab_V3+, and the introduction of CA and CFF modules.

II. DEEPLABV3+ BASIC FRAMEWORK

As shown in Fig. 1, DeepLab_V3+ is a convolutional neural network with encoding and decoding mechanism, which mainly contains four parts: feature extraction, feature processing, feature fusion, and prediction.

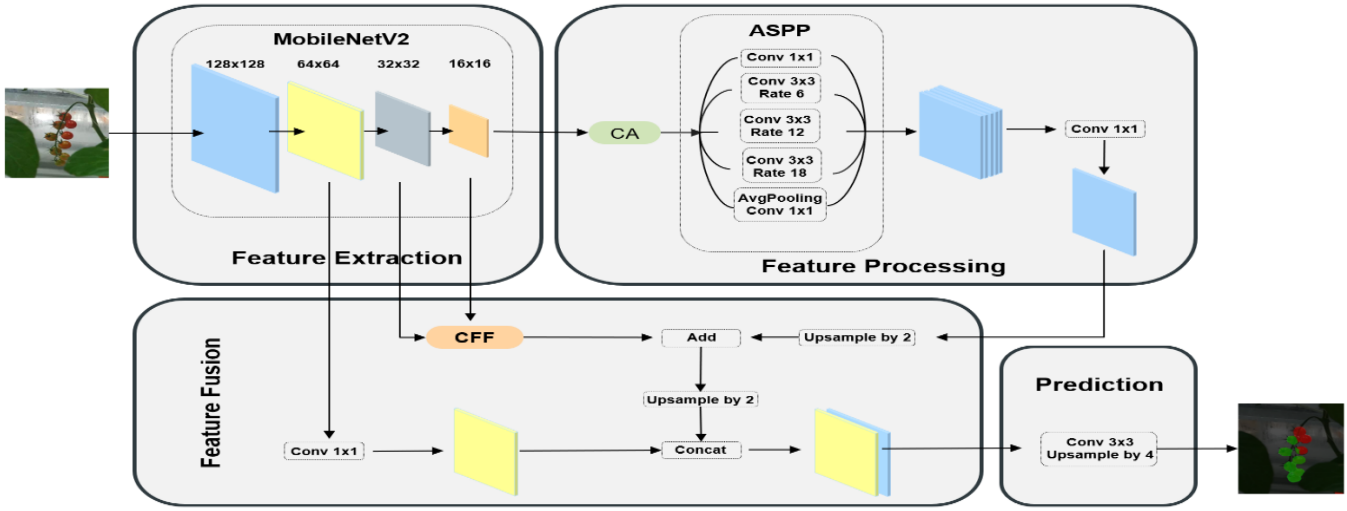


Fig. 1. Improved network architecture diagram of DeepLabV3+

A. Feature extraction component

Feature extraction is to extract features from the image, as shown in Fig. 1 the feature extraction part of this paper uses the structure of MobileNetv2, and four convolution operations are performed using depth-separable convolution to obtain feature maps that are 4 times, 8 times, 16 times, and 32 times downsampled from the size of the original image as out1, out2, out3, and out4, respectively.

B. Feature processing component

Feature processing is to get more image features and improve the accuracy of the task. In this paper, two feature processing methods, CA (Coordinate Attention) and ASPP (Atrous Spatial Pyramid Pooling), are mainly used for out4. CA not only has the ability of automatically weighting the channels of the channel attention mechanism, but also takes into account the positional information, which unites the channel attention and spatial attention; while ASPP mainly focuses on the spatial information in the feature map, and obtains the features of different sensory fields through multi-scale convolution operation. The combination of these two methods can improve the model's ability to perceive different scales and different channel information, thus improving the performance of the pixel-level classification task. The out4 is successively processed by CA and ASPP to obtain the feature map out_ASPP.

C. Feature fusion component

Feature fusion, by effectively integrating features from different levels, scales, or even sources within a model, can significantly enhance both the perceptual and expressive abilities of the model, leading to improved performance in various tasks. In this paper, three different feature fusion strategies, namely Cascade Feature Fusion (CFF), Addition (Add), and Concatenation (Concat), are employed. Firstly, features from out3 and out4 are fused using the CFF module to generate out_CFF. The architecture of the CFF module is depicted in the figure below. Compared with traditional single-level target detectors, CFF is able to realize the fusion of information from different scale feature maps, thereby improving the performance of multi-level detectors.

This concatenation process successfully captures the contextual relationships within the input data, leading to more comprehensive and richly detailed representations that are crucial for accurate model performance. The CFF feature fusion module is shown in Fig. 2.

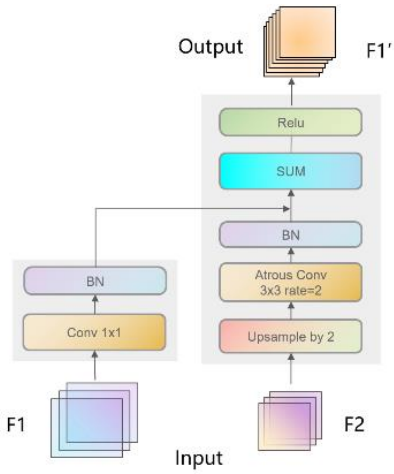


Fig. 2. The CFF feature fusion module

D. Forecasting component

The predictions in this paper are made using a convolutional kernel of size 3 for convolution followed by quadruple upsampling, and the final prediction is obtained by a fully connected layer.

III. DATA PROCESS

Training a deep learning model based on the tomato harvesting dataset is an important and challenging task. This dataset comprises 219 JPG format images captured within agricultural greenhouses and has been divided into a 9:1 ratio

for training and testing purposes, facilitating model training and evaluation. To meet the training requirements of the DeepLabV3+ network, all images in this project have been resized to a uniform size of 512x512 pixels.

Furthermore, in order to improve the model's accuracy in identifying ripe tomatoes and enhance its robustness in various scenarios, the dataset for this project includes tomatoes in different states, including ripe, unripe, and nearly ripe, as shown in Fig. 3. Additionally, factors such as reflections and occlusions that may occur during mechanical harvesting have been considered during the dataset creation process. This diversity contributes to a better understanding of tomato recognition tasks under different conditions by the model.

Moreover, considering the potential variations in environmental lighting conditions during real harvesting, this project's dataset also incorporates images from different levels of ambient brightness. This diversity helps boost the model's resistance to interference and its ability to adapt to image segmentation tasks under varying lighting conditions.

In summary, this dataset offers a wide range of tomato states and environmental factors, making it invaluable for training a deep learning model for tomato recognition in real-world agricultural settings. The dataset diversity allows the model to learn various characteristics of tomato fruits and handle complexities that may arise during tomato recognition tasks under different conditions. The dataset can be further extended to include more images and tomato states for even better accuracy and robustness of the models.

Fig. 3. Dataset examples

During the dataset annotation process, this project employed the Python-based annotation tool labelme, which enables the creation of various shapes for semantic segmentation annotations and the generation of VOC files. The VOC file format is widely used in computer vision tasks such as object detection and image segmentation. By utilizing the labelme tool, precise segmentation annotations can be generated easily, thereby assisting in training and evaluating machine learning models more effectively. Fig. 4. illustrates an example of dataset annotation using labelme.

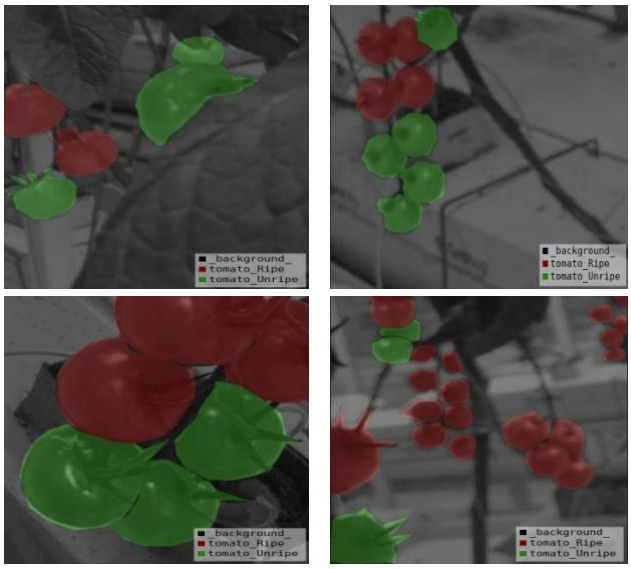


Fig. 4. Annotate examples

IV. Experimental Results

A. Experimental platform

The experimental configuration used by the research team is as follows: The central processing unit (CPU) is an AMD EPYC 9654 with 96 cores and 16 virtual CPUs (vCPU). The graphics processing unit (GPU) is an RTX 4090 (24GB). The system has 60GB of RAM and runs on the Windows 11 operating system. The experiment implemented the DeepLabV3+ model on the deep learning framework PyTorch 2.0.0. The runtime environment is Python 3.8, with the PyTorch version being torch2.0.0 and the Cuda version being 11.8. The experimental software used is PyCharm. Specific experimental parameter settings are as follows: image size is 512 x 512; the training process spans a total of 100 training epochs, with a batch size of 16 samples each time; other parameters are set to their default values experimental data.

The dataset used in this study consists of 219 manually captured photographs of tomato plants in orchards, which include both mature and immature tomatoes. The detailed division of the dataset is shown in TABLE I.

TABLE I. Division of Datasets

Train		Validation		Total
Number	Percent (%)	Number	Percent (%)	Number
219	90	22	10	219

B. Evaluation indicators of experimental results

To quantitatively assess the performance of the model, this paper adopts Loss, mIoU, mPA, mPrecision, and mRecall as evaluation metrics.

Loss (Loss Function): The loss function is used to measure the discrepancy between the model's predictions and the actual labels. Different tasks and models may employ various loss functions, such as the cross-entropy loss used in this paper.

Calculation formula:

$$\text{Loss} = -\sum_{i} y_i \log(p_i) \tag{1}$$

Among them, y_i is the real label, p_i is the model prediction probability.

mIoU (Mean Intersection over Union): Commonly used in image segmentation tasks, it measures the overlap between the predicted area and the actual area.

Calculation formula:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{2}$$

Among them, TP, FP, and FN stand for True Positives,False Positives, and False Negatives, respectively.

mPA (Mean Pixel Accuracy): Measures pixel-level accuracy, primarily used for image segmentation and pixel classification tasks.

Calculation formula:

$$\text{mPA} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{3}$$

mPrecision (Mean Precision): Represents the average proportion of True Positives in the model's predicted positives, typically used for classification and detection tasks.

Calculation formula:

$$\text{mPrecsion} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{4}$$

mRecall (Mean Recall) is one of the commonly used evaluation metrics in multi-class tasks, employed to measure the average recall rate of the model for each category. Recall is utilized to describe the model's ability to identify positives, that is, out of all actual positives, what proportion the model can correctly identify.

Calculation formula:

$$\text{mRecall} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{5}$$

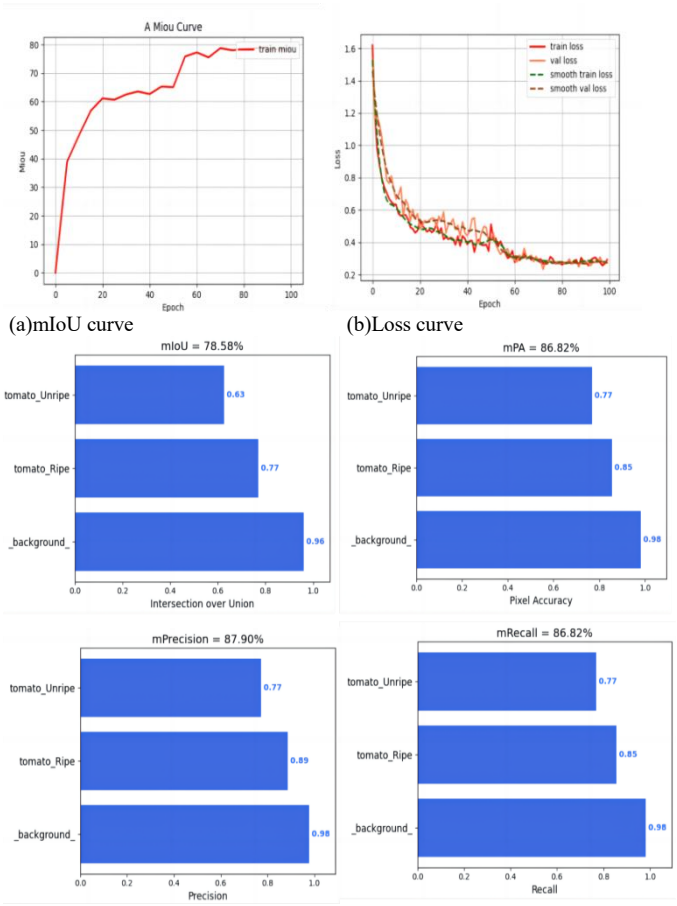
C. Analysis of experimental results

This paper places emphasis on a thorough study of two backbones based on DeepLabV3+, namely Xception and MobileNetv2, and compares their performance metrics. Objective evaluation metrics were visualized in the subsequent sections, and the results show that the combination of MobileNetv2 and DeepLabV3+ performs better. Furthermore, this study also delves into the improved methods of introducing channel attention and the CCF feature fusion module on top of MobileNetv2. Throughout the training process, various performance metrics of the DeepLabV3+ model were obtained, including Loss, mIoU, mPA, mPrecision, and mRecall. Empirical research shows that these performance metrics excel in terms of convergence speed; the Loss value decreases rapidly, while other metrics exhibit a favorable upward trend. These experimental results amply demonstrate the effectiveness and superiority of the proposed methods.

1) Using Xception as the backbone

The curve changes for Loss and mIoU are shown in Fig. 5 (a) and 5(b) respectively.

The Loss, mIoU, mPA, mPrecision, and mRecall when the number of classifications is 3 are depicted in Fig. 5(c).



(a)mIoU curve (b)Loss curve

(c)Loss, mIoU, mPA, mPrecision, and mRecall when the number of classifications is 3

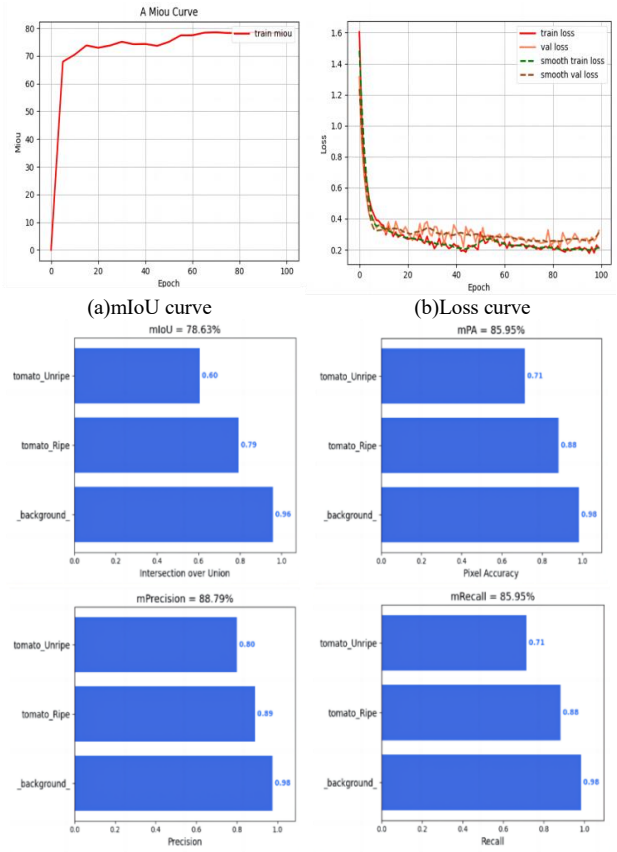
Fig. 5. Performance metric images with Xception as the backbone

2) Using MobileNetv2 as the backbone

The curve changes for Loss and mIoU are shown in Fig. 6 (a) and 6(b) respectively.

The Loss, mIoU, mPA, mPrecision, and mRecall when the number of classifications is 3 are depicted in Fig. 6(c).

This study found that using MobileNetv2 as the backbone allows for the convergence of Loss and mIoU with fewer training epochs, without a decline in performance metrics.



(a)mIoU curve (b)Loss curve

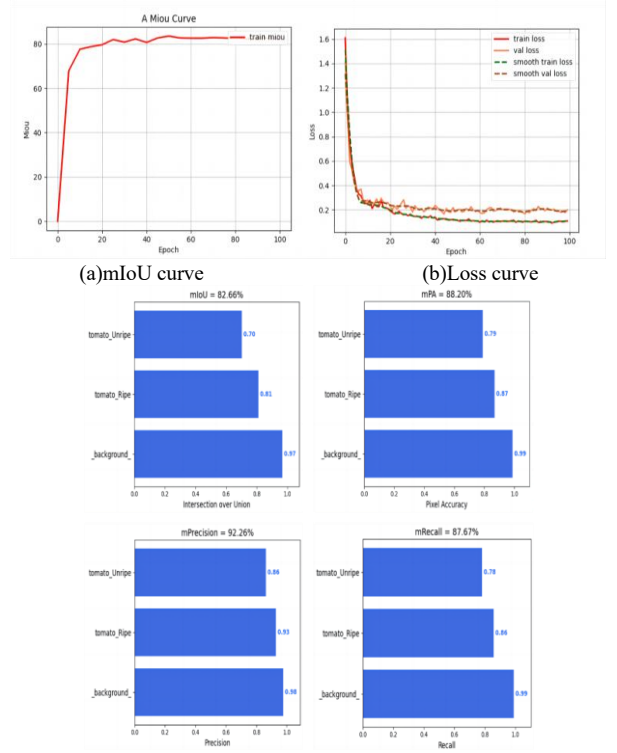
(c)Loss, mIoU, mPA, mPrecision, and mRecall when the number of classifications is 3

Fig. 6. Performance metric images with MobileNetv2 as the backbone

3) Incorporating the Coordinate Attention Mechanism (CA) and Cascade Feature Fusion Module (CFF) while using MobileNetv2 as the backbone

The curve changes of Loss and mIoU are shown in Fig. 7(a) and 7(b) respectively.

The metrics of Loss, mIoU, mPA, mPrecision, and mRecall for a classification count of 3 are illustrated in Fig. 7(c).



(a)mIoU curve (b)Loss curve

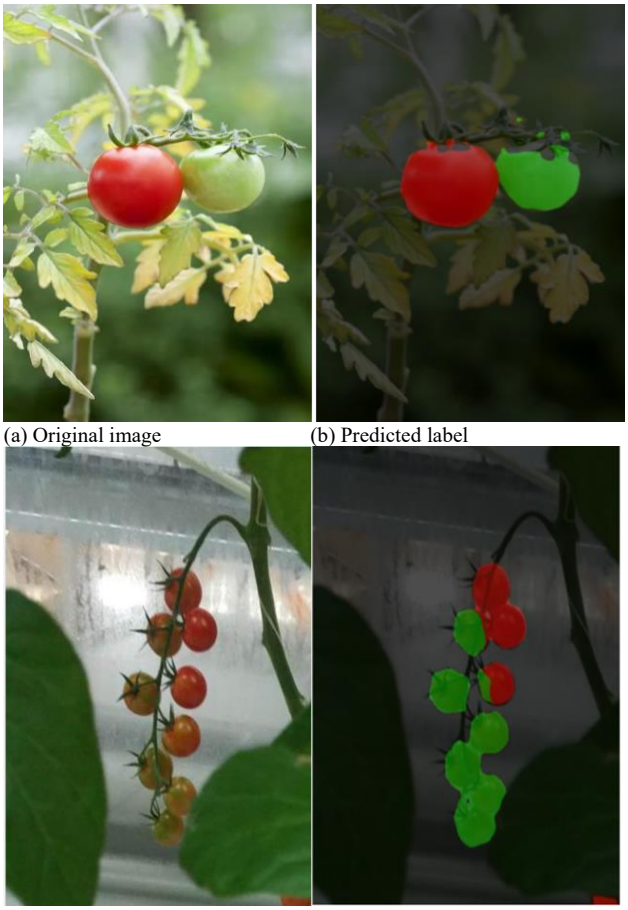
(c)Loss, mIoU, mPA, mPrecision, and mRecall when the number of classifications is 3

Fig. 7. Performance metric images incorporating CA and CFF on top of the model in 2

This study discovered that by incorporating the CA attention mechanism and the CFF feature fusion module, it was possible to achieve convergence of Loss and mIoU within fewer training epochs. Simultaneously, there was an observed improvement in metrics including mIoU, mPA, mPreciosn, and mRecall. Specifically, the mIoU reached 82.66%, mPA was 88.20%, mPreciosn stood at 92.26%, and mRecall was 87.67%.

4) Using the model from 3) as a benchmark for semantic segmentation of input images

The study selected two images outside of the dataset for testing. The results are shown in Fig. 8.



(a) Original image (b) Predicted label

(c) Original image (d) Predicted label

Fig. 8. Semantic segmentation on instances outside of the dataset

V. Conclusion

This paper provides a detailed introduction to a convolutional neural network semantic segmentation model based on DeepLabV3+, which demonstrates exceptional performance, effectively facilitating the task of autonomous tomato harvesting. The model is mainly divided into two key parts: the Encoder and the Decoder.

In the Encoder segment, there are two sub-components: feature extraction and feature processing. An innovative lightweight approach has been introduced in this study, which replaces the Xception in the original DCNN module with MobileNetv2. This successfully reduces the number of parameters and accelerates training speed. This improvement not only eases the model's computational burden and boosts training efficiency but also guarantees real-time performance. Furthermore, a Coordinate Attention Mechanism (CA) was introduced to the feature map output by the Deep Convolutional Neural Network (DCNN) module, allowing it to capture both local and global image information more effectively, thus providing precise guidance for subsequent segmentation. With the help of the Atrous Spatial Pyramid Pooling (ASPP) module, multi-scale information is integrated, enabling the model to better understand and process features of various scales in the input image.

For the Decoder, it consists of feature fusion and prediction. This paper introduces a Cascaded Feature Fusion (CFF) technique that merges the shallow and deep features from the encoder's DCNN and fuses them with the deep features output by ASPP. Additionally, the output results are fused with shallower features from the DCNN. This strategy effectively combines features from different levels, forming a unified feature representation. The richness and comprehensiveness of this representation supply crucial information for the final segmentation results.

In experiments, the model in this study exhibits a faster convergence and training speed compared to the original DeepLab v3+. The improved model performed excellently, with outstanding results in mIoU, mPA, mPrecision, and mRecall. In tests, mIoU was 82.66%, mPA 88.20%, mPrecision 92.26%, and mRecall 87.67%. This indicates that the model proposed in this paper possesses a high-precision identification and segmentation capability in tomato harvesting tasks and achieves satisfactory results in categorizing tomatoes of different maturity levels.

References

[1] Peng H, Xue C, Shao Y, et al. Semantic segmentation of litchi branches using DeepLabV3+ model[J]. IEEE Access, 2020, 8: 164546-164555.

[2] Li K, Zhang L, Li B, et al. Attention-optimized DeepLab V3+ for automatic estimation of cucumber disease severity[J]. Plant Methods, 2022, 18(1): 1-16.

[3] Barth R, IJsselmuiden J, Hemming J, et al. Data synthesis methods for semantic segmentation in agriculture: A Capsicum annuum dataset[J]. Computers and electronics in agriculture, 2018, 144: 284-296.

[4] Cai M, Yi X, Wang G, et al. Image segmentation method for sweetgum leaf spots based on an improved DeeplabV3+ network[J]. Forests, 2022, 13(12): 2095.

[5] Milioto A, Lottes P, Stachniss C. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs[C]//2018 IEEE international conference on robotics and automation (ICRA). IEEE, 2018: 2229-2235.

[6] Cheng H, Zhang J, Gong Y, et al. Semantic segmentation method for myocardial contrast echocardiogram based on DeepLabV3+ deep learning architecture[J]. Mathematical Biosciences and Engineering: MBE, 2022, 20(2): 2081-2093.

[7] Polat H. A modified DeepLabV3+ based semantic segmentation of chest computed tomography images for COVID-19 lung infections[J]. Int J Imaging Syst Technol. 2022; 32(5):1481-1495.

[8] Shoaib M, Hussain T, Shah B, et al. Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease[J]. Frontiers in Plant Science, 2022, 13: 1031748.

[9] Yang Z, Wu Q, Zhang F, et al. A New Semantic Segmentation Method for Remote Sensing Images Integrating Coordinate Attention and SPD-Conv[J]. Symmetry, 2023, 15(5): 1037.

[10] Chen S, Qiu C, Yang W, et al. Multiresolution aggregation transformer UNet based on multiscale input and coordinate attention for medical image segmentation[J]. Sensors, 2022, 22(10): 3820.