Green Tomato Segmentation Model Based on Optimized Swin-Unet Algorithm Under Facility Environments

Ru Jiang, Huichuan Duan, Jingyu Yan, Weikuan Jia

Abstract—**In facility-based agricultural environments, accurately identifying green tomatoes presents a significant challenge for machine vision systems due to the color similarity between green fruits and background branches and leaves as well as the overlapping occlusion between fruits. To solve this problem, this study constructs and optimizes the Attention Gate (AG) module using Swin-Unet as the baseline model, so that the model can focus on the features related to green tomatoes, suppress irrelevant regions in the background, and effectively enhance the representation of target features. Additionally, in order to optimize the edge smoothing of green tomato segmentation, this study further introduces a Atrous Spatial Pyramid Pooling (ASPP) module, which significantly improves the segmentation accuracy by expanding the feature sensing field and enhancing the multi-scale feature extraction capability. Experimental results on the specially constructed green tomato dataset show that the model achieves 97.5%, 92.4% and 85.9% for Pixel Accuracy (PA), Dice similarity coefficient (Dice) and Intersection over Union (IOU), respectively. The new model outperforms existing partial semantic segmentation models in several key metrics, proving its effectiveness in complex facility environments. This research not only addresses the technical difficulties in recognizing green fruits, but also provides solid technical support for the development and application of intelligent agricultural equipment. The model can be applied to segmentation and recognition of other types of fruits to meet the accuracy and efficiency requirements of green fruit recognition in smart agricultural equipment, which has a broad application prospect.**

Index Terms—**Green tomatoes, Swin-Unet, Semantic segmentation, ASPP, AG**

I. INTRODUCTION

Tomatoes are highly valued in agriculture due to their rich nutritional content. With the continuous expansion of planted areas and production, the ease of transportation and storage of green tomatoes further enhances their market appeal [1-2]. However, the cultivation and management of green tomatoes face numerous challenges in the complex environment of facility-based agriculture [3]. Among these challenges, harvesting is particularly time-consuming and labor-intensive, making the development of automatic picking robots a focal point of interest. The primary task of automatic picking involves accurately locating the fruits using computer vision [4-6]. Traditional machine methods struggle to recognize green fruits accurately due to their color similarity with the background foliage and the overlapping occlusion among fruits [7]. Therefore, the use of advanced semantic segmentation techniques to accurately segment and localize green tomatoes is of paramount importance. This technique not only contributes to efficient agricultural management and planning but also optimizes resource allocation. For instance, water sources can be more precisely targeted to tomato-growing areas, significantly reducing water wastage. Similarly, fertilizer application can be adjusted according to the distribution of tomatoes, effectively minimizing fertilizer waste. In automated harvesting, robots equipped with semantic segmentation capabilities can detect and locate fruits, thereby reducing reliance on human labor [8]. Thus, enhancing the visual system's segmentation precision is paramount not only for the effective management of green tomato cultivation but also as a pivotal element in augmenting the efficacy of robotic fruit and vegetable harvesters.

In recent years, image semantic segmentation has emerged as a prominent focus in the field of deep learning. The integration of deep learning-based image semantic segmentation techniques with agricultural applications has gradually evolved into a significant area of development [9]. Studies show that image semantic segmentation offers a significant advantage in accurately segmenting fruit targets, particularly when there is a pronounced color difference between the fruit and the background. Häni [10] found that apple target segmentation can be effectively achieved using the U-Net model, especially when the color distinction of the fruit is clear. The U-Net model demonstrates superior segmentation performance when there is high similarity in dataset characteristics. In images with a dense accumulation of fruits, missing fruit detection is common. For instance, Bargoti [11] employed a multi-scale multilayer perceptron combined with a convolutional neural network (CNN) to segment apple images. While this method demonstrated high

Manuscript received April 17, 2024; revised September 24, 2024.

This work is supported by National Nature Science Foundation of China (No.: 62072289); Young Innovation Team Program of Shandong Provincial University (No.: 2022KJ250); New Twentieth Items of Universities in Jinan (2021GXRC049).

R.Jiang is a postgraduate student of School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (superjr1102@163.com);

H.C. Duan is a professor of School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (Corresponding author, hcduan@sdnu.edu.cn);

J.Y.Yan is a postgraduate student of School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (1543670705@qq.com)

W.K. Jia is an associate professor of School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (Corresponding author, phone: +86-531-86181755; fax: +86-531-86181750;
e-mail: jwk_1982@163.com)

detection accuracy under ordinary conditions, it was unable to completely recognize all the fruits in images taken in complex environments. Barth et al. [12] used DeepLab to segment bell pepper fruits and plants. Kang et al. [13] proposed the DaSNet-V2 network architecture for real-time detection and semantic segmentation of apples and branches in orchard environments using visual sensors. This network enhances feature extraction capabilities through spatial pyramid pooling and a gated feature pyramid structure. Experimental results demonstrate that the optimal model achieves a segmentation accuracy of 87.6% and an F1 score of 77.2%. Mo et al.[14] proposed a method for the semantic segmentation of apples based on an improved DeepLabV3+ architecture. The encoder utilizes a lightweight MobileNet module for feature extraction and employs depthwise separable convolution instead of standard convolution. This model achieves a pixel accuracy (PA) of 95.3% and a mean intersection over union (MIoU) of 87.1%. Semantic segmentation is also commonly used to segment rotten parts of fruits. For instance, Matsui [15] trained and validated a U-net++ model on X-ray avocado images to detect internal fruit rot, achieving an accuracy of 98%. Roy [16] constructed a semantic segmentation model based on En-UNet to segment rotten parts in apple RGB images, achieving training and validation accuracies of 97.46% and 97.54%, respectively. These studies highlight the significance of image semantic segmentation technology in the agricultural domain, particularly in cases where there is a pronounced color difference between fruits and the background. Accurate image segmentation can improve agricultural automation efficiency, enhance fruit quality assessment, and optimize disease detection and management strategies.

In complex situations where the target and the background colors are similar, deep learning-based image semantic segmentation techniques face significant challenges, primarily in distinguishing between the target and the background. To address this issue, several studies have made significant progress. For example, Li [17] proposed an optimized U-Net model by integrating residual blocks and gated convolutions to develop the Edge structure. They also used Atrous Spatial Pyramid Pooling (ASPP) to merge Edge features with the high-level features of U-Net, significantly improving the segmentation accuracy for green apples and enhancing the model’s generalization ability. Subsequently, He [18] enhanced the DeepLabV3+ model by replacing its backbone with MobileNetV2, introducing the Shuffle Attention Mechanism, and replacing the activation function with Meta-ACONC. This enhancement increased the MIoU metric for green banana crown segmentation to 85.75% and the MPA to 91.41%. Yan [19] proposed a lightweight convolutional neural network based on an improved DeepLabV3+ for segmenting and locating picking points of tea leaves, achieving an MIoU of 91.85%. Additionally, Bai [20] achieved fine pixel-level segmentation of cucumbers by improving the U-Net model, with mIOU and mean pixel accuracy reaching 94.24% and 97.46%, respectively, This improvement enhanced the recognition of features such as the shape and texture of green cucumbers in complex agricultural environments. Liu et al. utilized a complex number neural network (cNN) to segment bell peppers among green leaves using hyperspectral inputs, demonstrating the effectiveness of this method in generating more stable and less noisy segmentation results [21]. Although deep learning has made significant progress in image segmentation, applications involving similarly colored targets and backgrounds in specific scenarios continue to face inherent challenges, such as target misclassification, omission, and poor segmentation of fruit edges. Nonetheless, these advances have facilitated research into the semantic segmentation of specific targets, such as green tomatoes, and proposed new directions for designing and optimizing deep network architectures.

However, the complexity and unstructured nature of facility environments present challenges, including variations in lighting angles, occlusion or overlap of fruits, collection angles, and the similarity of green fruits to the background. These factors impact segmentation accuracy and necessitate further research and improvements. This paper proposes a semantic segmentation model for green tomatoes based on an improved Swin-Unet [22]. The main contributions of this study include:

(1) To more accurately segment green tomatoes, which are similar in color to the background, this study incorporates the Attention Gate (AG) module within the skip connections. Attention coefficients are designed to evaluate the importance of each feature, allowing the model to focus on features associated with green tomatoes while suppressing irrelevant background regions.

(2) To achieve multi-scale extraction of green tomato features and optimize the smoothness of their segmentation edges, this study proposes the ASPP module in the bottleneck section, which enlarges the feature receptive field while keeping the parameter quantity unchanged, enhancing the model's ability to handle contextual information at different scales.

(3) Experiments conducted on a custom green tomato dataset demonstrate that this method outperforms other state-of-the-art techniques in accuracy and robustness. proving to be more suitable for segmenting green tomatoes in facility environments.

II. Materials and Methods

A. Green Tomato Fruit Dataset

This study aims to address the challenge of green tomato segmentation in facility agricultural environments, where the color similarity between green tomato fruits and background leaves complicates recognition, often resulting in fruit omission or confusion with branches and leaves.

1) Image Collection

Image Acquisition Location: Greenhouse, Hetong Village, Shangkou Town, Shouguang City, Weifang, Shandong Province, China.

Image Acquisition Equipment: The images were captured using a Canon EOS 80D DSLR camera and subsequently resized to a resolution of 640 × 640 pixels.

Image Acquisition Environment: To enhance the model's robustness during training, green tomato images were collected from diverse environments, varying in time of day, lighting conditions, and occlusion scenarios.

a Daytime green tomato images b Frontlighting green tomato image c Backlighting green tomato image

d Nighttime green tomato image e Overhead shot of green tomato image f Close-up image of green tomato image

g Distant view green tomato image h Block the green tomato image i Overlapping green tomato image

Fig.1 Images of green tomato fruits in different environments

The captured images are presented in Fig. 1. Figures 1a to 1d display green tomatoes under different lighting conditions, including natural daytime illumination (both front and back lighting) and nighttime illumination by LED lights. Figures 1e to 1g illustrate tomato images taken from various proximities and viewpoints, simulating perspectives typical of picking equipment in a real orchard environment. Figures 1h and 1i show examples of significant shading and overlapping in facility-based agricultural environments, with fruits obscuring one another and branches and leaves causing visual obstructions.

2) Dataset Creation

To accurately reflect the complexity of the facility-based agricultural environments, the images captured in this study were designed to emphasize randomness and expressiveness. This approach ensures that the images closely align with the visual processing requirements of mechanical equipment in real-world operations. Considering the efficiency of mechanical equipment in handling low-resolution images, the captured images were uniformly scaled and compressed to 640 × 640 pixels. This adjustment aims to optimize the green tomato segmentation network, making it more adaptable to the segmentation of low-resolution images.

Existing datasets for green tomato image classification are primarily designed for classification tasks and lack the labels necessary for semantic segmentation. To address this issue, the LabelMe [23] software was employed to manually annotate the datasets in detail. This process involved generating category labels and annotation points for each image, thereby providing the essentation ground truth information for semantic segmentation. All annotation data were meticulously recorded and saved in JSON files. Additionally, corresponding labeled images were generated based on these JSON files to ensure that each original image had a matching labeled version, as illustrated in Fig. 2.

The dataset was divided into two subsets at an 8:2 ratio, the training set, which includes 1066 images for model training, and the validation set, comprising 267 images for evaluating the model's performance. This careful partitioning enables a more accurate assessment of the model's efficacy.

Fig. 2. Original image and annotated image

TABLE I
THE QUANTITY DISTRIBUTION OF DIFFERENT MASK SIZES IN THE GREEN TOMATO DATASET.

Dataset	Class	Images	Instances	Target Amount		
				Small ($0<area<32^2$)	Medium ($32^2<area<96^2$)	Large ($96^2<area$)
Train	tomato	1066	4752	3（0.06%）	232(4.88%)	4517(95.06%)
Test	tomato	267	1442	5 (0.35%)	187(12.97%)	1250(86.68%)

Green tomato fruits were categorized into small-scale, medium-scale, and large-scale classes according to the criteria used in the Microsoft COCO [24] dataset. The area of each fruit instance was determined by the number of pixels in its corresponding mask to evaluate the performance of the algorithm. Table I provides the relevant details. It is noteworthy that the number of small-scale fruits is relatively low, with only three small targets in the training set, representing 0.06%, and small targets in the validation set accounting for 0.35%.

B. Optimization of Swin-Unet Segmentation Model

In facility-based agricultural environments, image acquisition is challenging due to complex backgrounds, occlusion, overlapping branches and foliage, and variable lighting conditions (including downlight, backlight, and nighttime environments), all of which can adversely affect image quality. Specifically, for green fruits with colors similar to the background, their boundaries are often indistinct, which significantly complicates accurate segmentation [25]. Additionally, a notable issue in the agricultural field is the insufficiency of samples, especially for the segmentation of specific types of fruits, such as green fruits. The lack of adequate labeled samples to train high-performance segmentation models not only limits the effectiveness of model training but also increases the risk of overfitting. Consequently, this can diminish detection accuracy and present challenges in meeting the operational requirements of actual mechanical equipment.

To address the challenge of accurately segmenting green tomatoes, this paper presents an optimized model based on the Swin-Unet architecture. The Swin-Unet architecture integrates the robust feature extraction capabilities of the Swin Transformer with the high-precision segmentation capabilities of U-Net, resulting in an innovative framework. Through its unique design, the model effectively addresses long-range dependency issues while preserving spatial information, which is crucial for segmenting green tomato images with complex backgrounds. The optimized model consists of four main components: Encoder, Bottleneck, Decoder, and Skip Connection (Fig. 3). In the Encoder stage, the model adjusts channel numbers using Patch Partition and Linear Embedding techniques to achieve feature extraction and downsampling through multiple Swin Transformer Blocks and Patch Merging layers. The Bottleneck stage incorporates an Atrous Spatial Pyramid Pooling (ASPP) module [26] to capture image information at various scales and expand receptive fields. In the Decoder stage, multiple Swin Transformer modules, along with Patch Expanding layers, are employed for upsampling and restoring feature map sizes. Additionally, an Attention Gate (AG) module [27] is introduced in skip connections to enhance target feature information while suppressing irrelevant details for improved segmentation accuracy. This entire process effectively integrates multi-scale information, enhancing the segmentation results of green tomato images.

1) Target Feature Enhancement Module

In complex scenarios where green tomatoes closely resemble the background color, model performance is often compromised by interference from non-target areas, leading to reduced segmentation accuracy. To address this issue, the Attention Gate (AG) was introduced to better identify and emphasize image regions relevant to the task. The overall structure of the Attention Gate module is illustrated in Fig. 4. The AG utilizes attention mechanisms to automatically focus on the image regions most pertinent to the task while

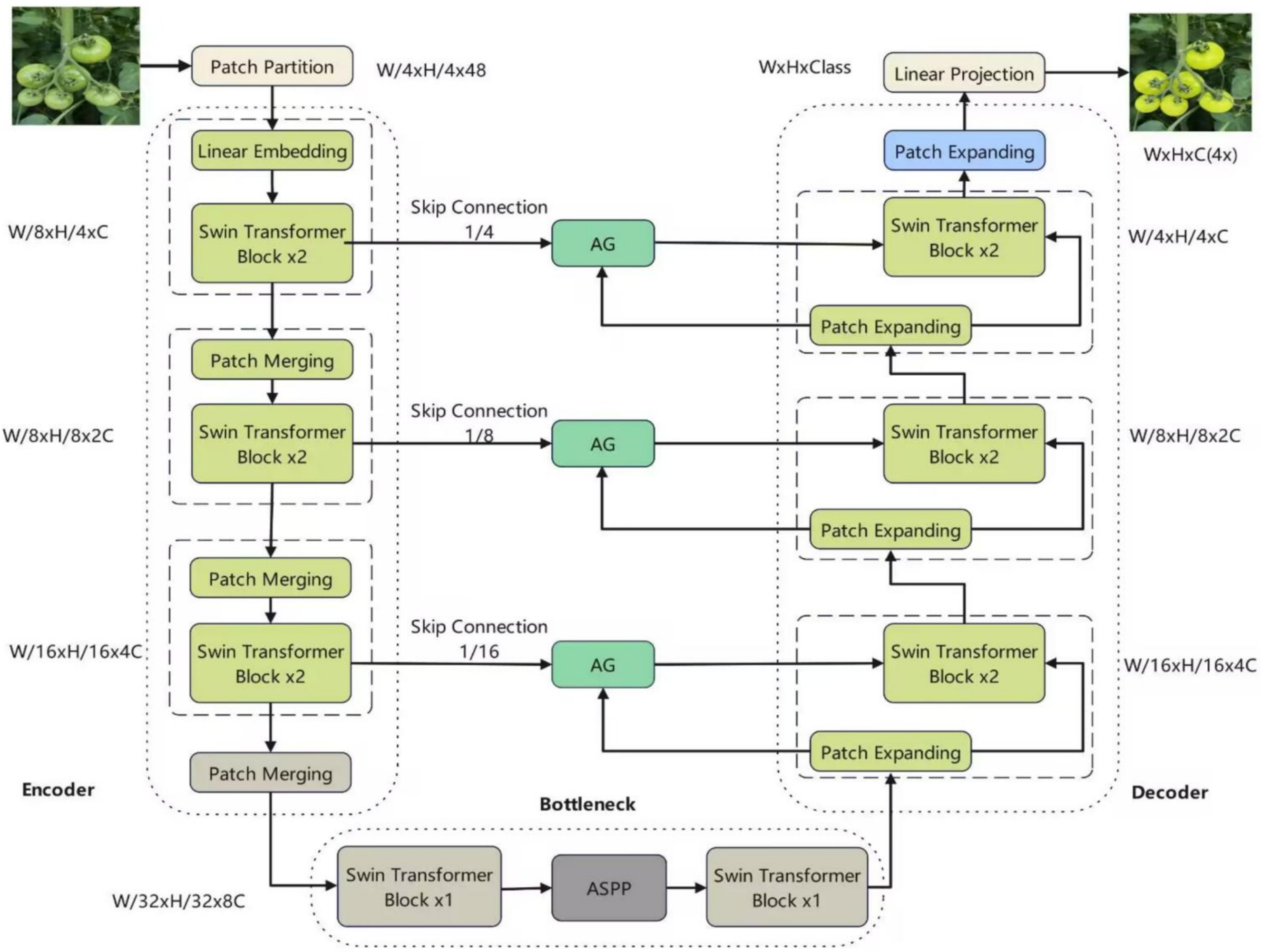


Fig. 3. Structure of the green tomato segmentation model optimized based on Swin-Unet

suppressing less important areas. By calculating attention coefficients for each feature, the model adjusts the weights of the feature maps, directing the network's focus towards the target area rather than the background. This optimization not only improves the model's learning process, making it more efficient in extracting key features, but also significantly enhances overall accuracy. Specifically, in the segmentation of green tomatoes, it notably increases diagnostic accuracy.

Skip connections merge features from the encoder and decoder, preserving the spatial integrity of the image while enhancing the model's ability to recognize details, such as edges, which are crucial for the segmentation of green tomatoes. The incorporation of the Attention Gate (AG) module into skip connections further amplifies this advantage by enabling dynamic feature weight allocation. This allows the model to flexibly adjust its focus on features based on image content, concentrating on the most relevant image regions through attention mechanisms. By focusing on these relevant image regions, sensitivity to important features is heightened, significantly improving the model's overall accuracy and enabling more precise differentiation between green tomatoes and similarly colored backgrounds. Specifically, two input features, x (from the encoder) and g (from the decoder) are transformed into new features $x1=W_x x$ and $g1=W_g g$, through their respective linear transformations.

These parts are then weighted and merged, and an attention coefficient ζ , ranging from 0 to 1, is obtained via the Sigmoid activation function by adding g_1 and x_1.When ζ is close to 1, the corresponding feature has a higher weight in the fusion, whereas when ζ is close to 0, the feature's weight is lower.This dynamic feature weight allocation enables the model to flexibly adjust its focus on features based on the content of the image, thus determining the importance of each feature.

$$\zeta = Sigmoid(g_1 + x_1) \tag{1}$$

Subsequently, through another linear transformation φ , the attention coefficients are adjusted to match the dimensions of the feature map x. These coefficients are then multiplied by the encoder feature map x to obtain the weighted feature map $x^{\sim}$, thereby accomplishing feature selection and enhancement.

$$x^{\sim} = \varphi(\zeta) \odot x \tag{2}$$

Where Wx and Wg serve as the weight matrices for the linear transformations applied to x and g, respectively. ζ signifies the attention coefficients post-Sigmoid function processing, denotes another linear transformation, and $\odot$ represents element-wise multiplication (hadamard multiplication), leading to the generation of the final weighted output feature map.

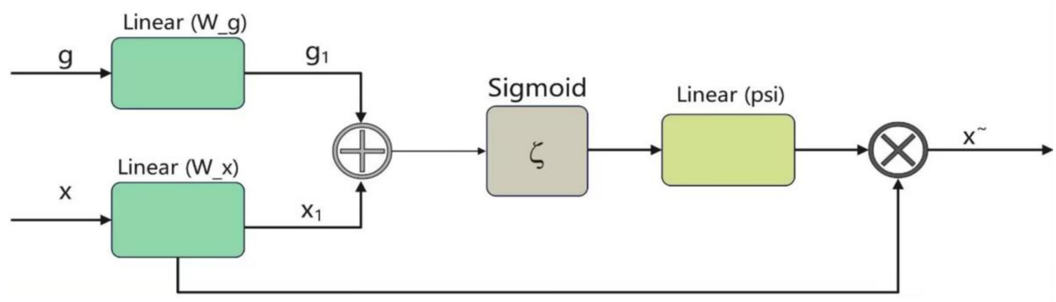


Fig.4. Attention Gates Module

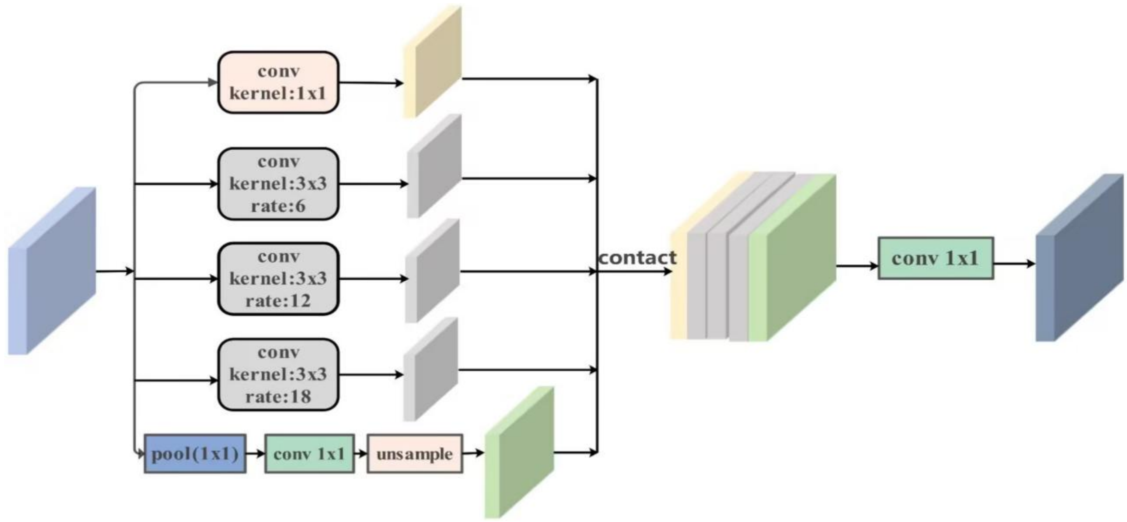


Fig.5. Atrous spatial pyramid pooling Module

2) Multiscale Edge Restoration Module

In the encoder-decoder architecture used for green tomato image segmentation, the encoder expands the receptive field through downsampling, while the decoder restores the image to its original size via upsampling. However, this process can result in the loss of semantic features at the edges of green tomatoes and the neglect of contextual information, which adversely affects segmentation accuracy.

ASPP is a spatial attention mechanism for image segmentation modeling designed to improve the capture of contextual information. This module effectively captures the features of green tomato images through dilated convolutions at different scales, thereby integrating these features to strengthen the model's understanding of the semantic content of the images. The ASPP module consists of five parallel branches: one 1x1 convolution primarily for extracting local information and reducing the number of parameters; three 3x3 dilated convolutions with different dilation rates (6, 12, 18), allowing the convolution kernels to cover a broader input area without increasing the number of parameters, thereby aiding in capturing wider contextual information. ASPP enlarges the receptive field through convolutions with varying dilation rates, better capturing the detail features of green tomatoes, especially at different scales. Additionally, a global average pooling branch generates a global feature descriptor by applying global average pooling to the feature map, aiding the model in grasping image-level contextual information. This structure allows the model to extract multi-scale features while maintaining the same number of parameters, enlarging the receptive field, and thereby enhancing the expressive capability of feature maps [28].

To address this issue, the Atrous Spatial Pyramid Pooling (ASPP) module is employed. The structure of the ASPP module is illustrated in Fig. 5. This module effectively captures detailed features through atrous convolution different scales, enhancing the model's performance in handling the details of green tomato edges, and significantly preserving edge details and significantly improving segmentation accuracy.

3) Loss Function

The goal of green tomato image segmentation is to accurately recognize green tomatoes through pixel-level classification, ensuring a clear distinction from the background. However, a significant size difference between the green tomatoes and the background, resulting in an imbalance in the number of pixels between the two categories. This imbalance makes it difficult for the model to adequately learn the features of the green tomatoes, often leading to an increase in false-negative predictions. Consequently, this issue seriously impacts the accuracy of the semantic segmentation of green tomatoes [29].

The Cross Entropy Loss (CE Loss) function is employed to calculate the prediction accuracy for each pixel and then averages these calculations to obtain an overall loss value. This approach treats the prediction of each pixel as independent and equally important. However, in the context of green tomato segmentation, where the target and background pixel categories are imbalanced, the loss function becomes dominated by the background pixels. This dominance biases the model heavily toward the background, leading to poor training outcomes and inaccurate predictions for the green tomatoes. The cross-entropy loss formula is as follows:

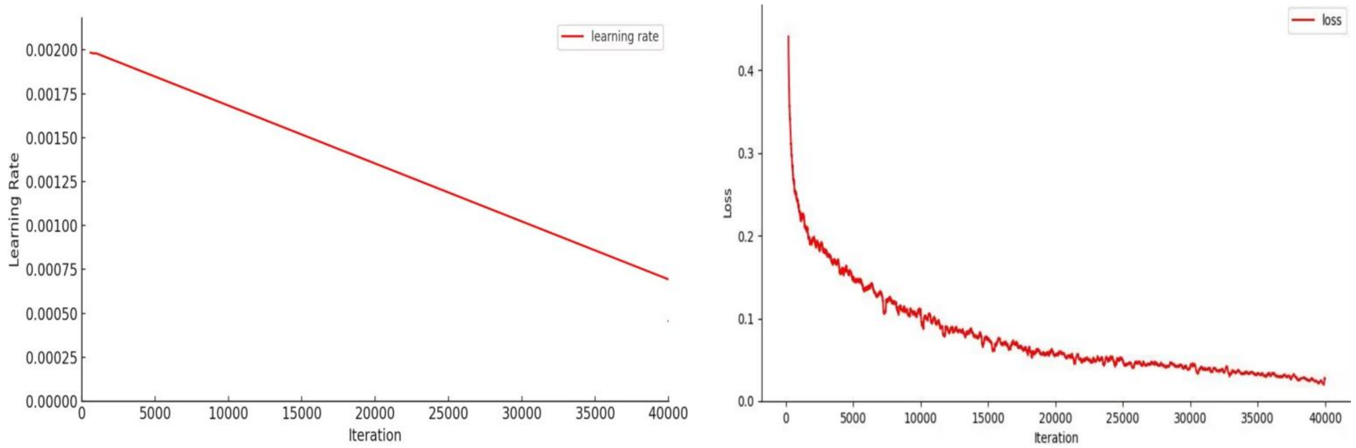


Fig.6. Left: Learning rate change during training.
Right: Changes in loss on datasets in the model training phase.

TABLE II
COMPARATIVE RESULTS OF THE IMPACT OF ASPP AND AG MODULES ON SWIN-UNET

Base Model	ASPP	AG	PA(%)	Dice(%)	IoU(%)
Swin-Unet	×	×	93.0	86.5	76.3
	×	✓	96.0	87.6	77.9
	✓	×	96.8	90.2	82.2
Swin-Unet	✓	✓	97.5	92.4	85.9

$$Loss_{ce} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log(y_i')+(1-y_i)\log(1-y_i')] \qquad (3)$$

To effectively address the imbalance problem, Dice Loss is adopted. This loss function calculates the loss by comparing the similarity between the predicted probabilities and the true labels, making it particularly suitable for addressing category imbalance. Dice Loss ensures that the model optimizes the prediction of frequent categories while also paying attention to infrequent categories. Consequently, it effectively mitigates the model's bias toward the background in green tomato segmentation.

$$Loss_{dice} = 1 - \frac{2\sum_{i=1}^{N} y_i' y_i}{\sum_{i=1}^{N} y_i' + \sum_{i=1}^{N} y_i} \qquad (4)$$

The gradient form of the Dice Loss is complex, and its formula is as follows:

$$\frac{\partial Loss_{dice}}{\partial y_i'} = -\frac{2y_i^2}{(y_i'+y_i)^2} \qquad (5)$$

Based on (5), it can be inferred that, in extreme scenario when the values of andare very small, the gradient values may become very large, potentially leading to more unstable training. To comprehensively consider the effects of class imbalance and training instability on segmentation edge accuracy, this study employs a weighted composite loss function, combining cross entropy loss and dice loss. The specific loss function is as follows:

$$Loss = \alpha Loss_{ce} + (1-\alpha) Loss_{dice} \qquad (6)$$

Herein, represents the true label of the ith pixel, denotes the probability of predicting the ith pixel as the target category, and α is the weight coefficient to balance the two loss terms. In the experiments, α is set to 0.4 to ensure an effective balance between Cross Entropy Loss and Dice Loss.

III. RESULTS AND ANALYSIS

To better validate the effectiveness of the model for green tomato segmentation, a series of experiments were conducted in this study. The experimental details were meticulously described, and the results were compared and analyzed. During the training process, the optimal model was selected and applied to the validation set to facilitate a comparative evaluation of the experimental outcomes. Comparative experiments were conducted under identical experimental configurations to assess the performance of the proposed model in this study.

A. Experimental Environment

The experimental setup is based on the Ubuntu 18.04 64-bit system, utilizing the deep learning framework PyTorch. The GPU used for the experiments is a 24GB NVIDIA A30, with CUDA version 11.4. All models were run using Python version 3.7 and PyTorch version 1.12.

B. Parameter Settings

Prior to inputting into the training network, image sizes were uniformly fixed at (640,640). Pre-trained weights from the ImageNet dataset, specifically swin_tiny_patch4 _window7_224.pth, were used, with an initial learning rate of 0.01, momentum of 0.9, weight decay of 0.0001, and stochastic gradient descent (SGD) as the optimization algorithm. The model was trained for 50 epochs, with a batch size of 2, window size of 7, and patch size of 4x4.The learning rate curve variation is shown in Figure 6. Using the above training parameters, the training loss variation curve of the model is based on the Green Tomato dataset as shown in Fig 6.

Fig.7. Comparative Visualization of Ablation Study Results.

C. *Evaluation Metrics*

To evaluate the segmentation accuracy of the optimized Swin-Unet algorithm for green tomatoes, metrics such as Precision, Recall, Pixel Accuracy (PA), Dice coefficient, and Intersection over Union (IoU) are commonly used in semantic segmentation methods [30].

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

PA denotes the ratio of the number of correct predictions for all pixel classes to the total number of pixels.

$$PA = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

The Dice coefficient is used to calculate the similarity between two sets, as shown in equation (11):

$$Dice = \frac{2TP}{FP + 2TP + FN} \quad (11)$$

The Intersection over Union (IoU) represents the ratio of the intersection to the union of two sets, as illustrated in equation (12):

$$IoU = \frac{TP}{FP + TP + FN} \quad (12)$$

This paper defines the metric using a confusion matrix, categorizing green tomato samples based on the relationship between predicted values and actual values into four categories: true positive (TP): actual positive samples; false positive (FP): false positive samples; true negative (TN): actual negative samples; false negative (FN): False negative samples.

D. Ablation Study

To verify the effectiveness of these two structures, the optimized Swin-UNet algorithm and the original Swin-UNet algorithm were evaluated in the Green Tomato dataset. To ensure fairness and comparability the experimental settings and hyperparameter configurations were kept consistent across all algorithms, as detailed in Table II.

As shown in Table II, the introduction of the AG module into the base model improves the PA, Dice, and IoU of the model by 3.0, 1.1, and 1.6 percentage points, respectively, compared to the original model. This significant enhancement in segmentation accuracy for green tomato images confirms the AG module's effectiveness in focusing more on relevant regions while suppressing irrelevant areas.Furthermore, the incorporation of the Atrous Spatial Pyramid Pooling (ASPP) module into the base model increases PA, Dice, and IoU by 3.8, 3.7, and 5.9 percentage points, respectively. These results demonstrate that the ASPP module substantially improves the model's ability to process contextual information across different scales without adding extra parameters. This enhancement is reflected in the model's improved performance in capturing edge details and overall semantic understanding, leading to more precise edge segmentation of green tomatoes.

Finally, by introducing both ASPP and AG modules into the base model, compared to the original model, the model's PA, Dice, and IoU increased by 4.5, 5.9, and 9.6 percentage points, respectively. This further validates that both proposed modules effectively enhance accuracy in green tomato image segmentation.

In order to compare the effect of each module on the segmentation results more intuitively, the results of the ablation experiments are visualized in this paper, and the visualization results are shown in Fig. 7. As can be seen from Fig 7, the original Swin-Unet model exhibits issues such as unclear segmentation edges and target leakage, etc. However, with the gradual introduction of modules such as ASPP and AG, the target contour becomes more accurate and clearer, and the segmentation effect of the model is much closer to that of the real labels, and the phenomenon of unclear segmentation edges of the target leakage is reduced, which fully proves that the model proposed model effectively improves green tomato image segmentation.

E. Segmentation Results

To further analyze the performance of the algorithm, the optimized Swin-Unet is compared with several contemporary and advanced semantic segmentation algorithms using the Green Tomato dataset. The comparative algorithms include DeepLabv3+ [31], DeepLabv3 [32], PSPNet [33], DANet [34], KNet [35], ISA-Net [36], DPT [37], OCRNet [38], and BEiT [39]. All experiments were conducted under identical conditions, with consistent parameter settings, datasets, and evaluation criteria.

TABLE III
EXPERIMENTAL RESULTS COMPARING DIFFERENT ALGORITHMS

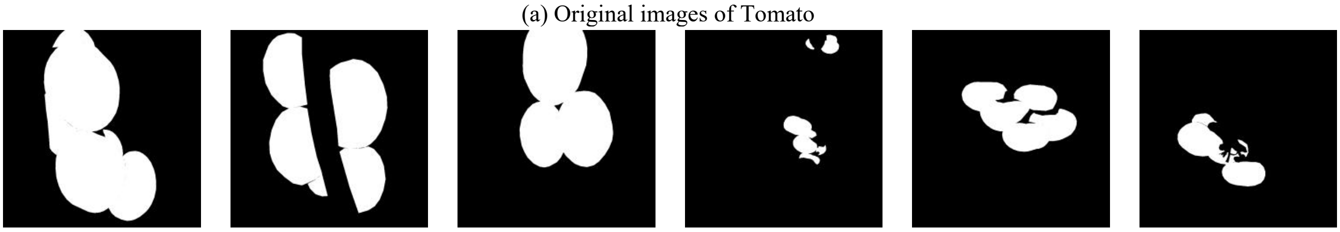
Segmentation model	Precision(%)	Recall(%)	Accuracy(%)	Dice(%)	IoU(%)
Deeplabv3+	97.88	85.23	85.23	91.12	83.68
Deeplabv3	97.5	84.13	84.13	90.32	82.35
Pspnet	96.67	89.83	86.83	91.48	84.3
Danet	97.73	86.65	86.65	**92.46**	84.94
Knet	**98.27**	87.15	87.15	92.38	85.84
Isanet	97.01	87.15	87.15	91.82	84.87
Dpt	93.14	88.8	88.8	90.92	83.35
Ocrnet	97.66	87.74	87.74	92.34	85.84
Beit	93.62	83.74	83.74	88.4	79.22
Ours	98.0	89.9	97.5	92.4	85.9

TABLE IV
COMPARISON OF THE NUMBER OF PARAMETERS AND FLOPS COMPUTATIONAL COMPLEXITY OF MODELS. INPUT SIZE: (640,640).

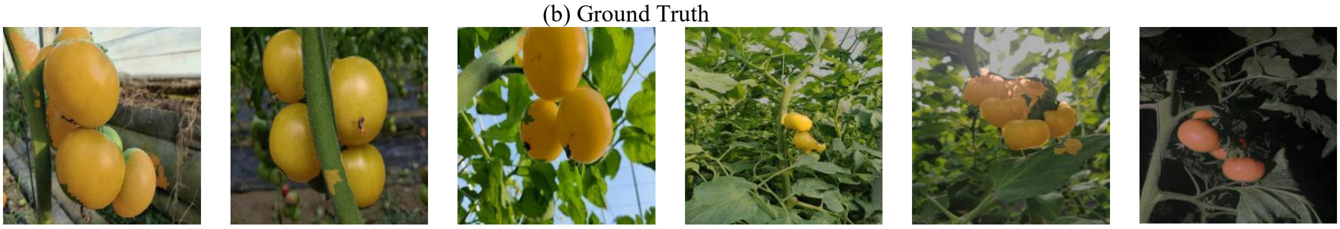
Method	Deeplabv3+	Deeplabv3	Pspnet	Danet	Knet	Isanet	Dpt	Ocrnet	Beit	Ours
Params/M	41.216	65.74	46.602	47.485	60.412	35.344	110	12.067	72.137	27.55
GFLOPs/G	276	422	279	338	320	235	360	82.902	437	116.15

Ⅰ Overlapping tomato image | Ⅱ Block the tomato image | Ⅲ Overhead shot of tomato image | Ⅳ Distant view tomato image | Ⅴ Backlighting tomato image | Ⅵ Nighttime tomato image

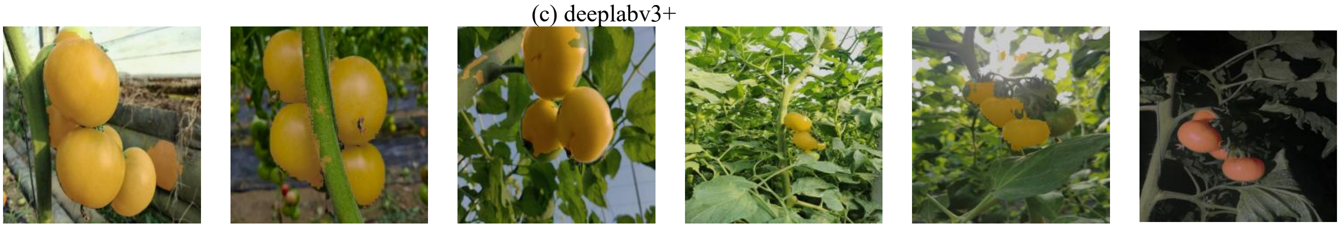
(a) Original images of Tomato



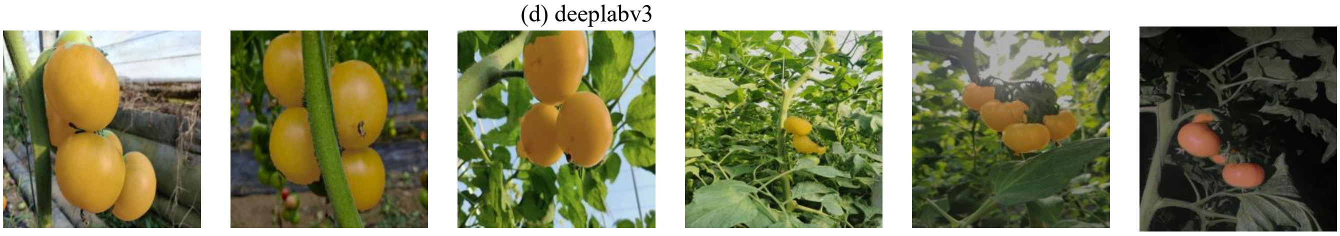
(b) Ground Truth



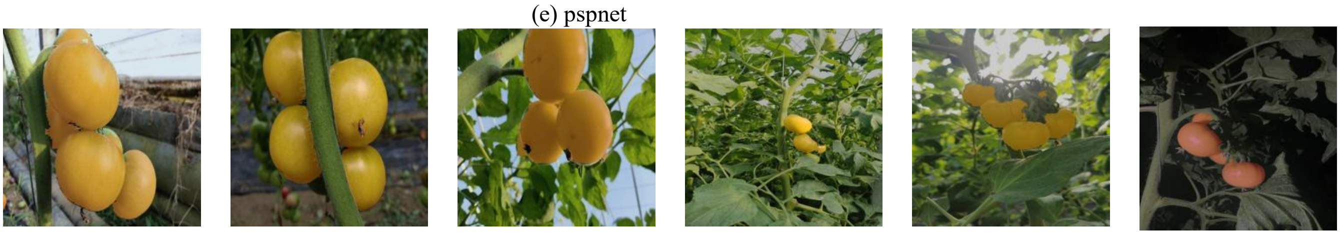
(c) deeplabv3+



(d) deeplabv3

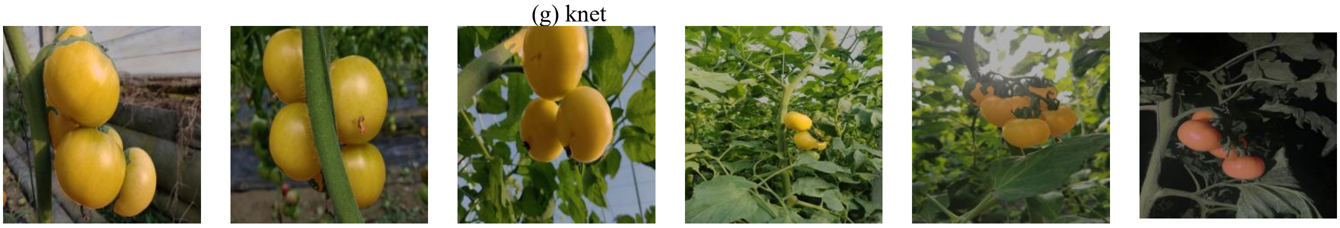


(e) pspnet

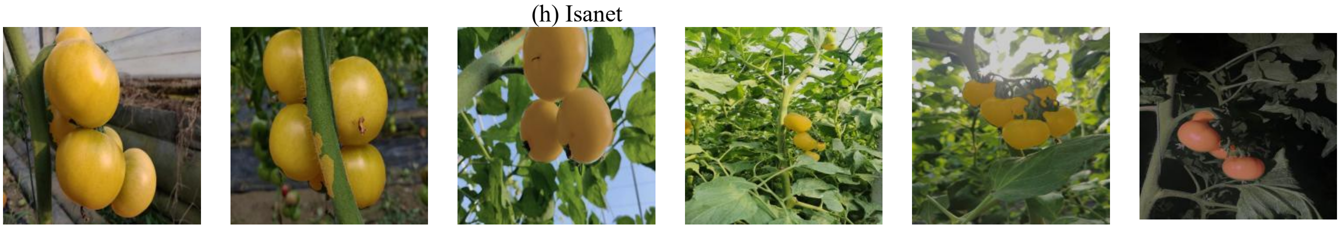


(f) panet

(g) knet

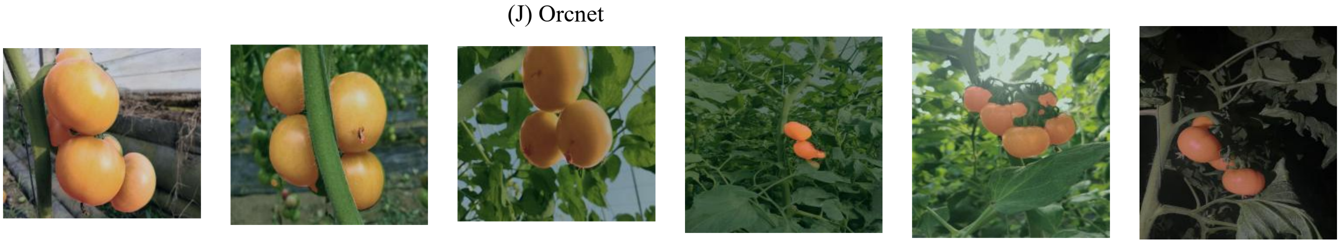


(h) Isanet



(i) Dpt

(J) Orcnet



(k) Beit

(l) Ours

Fig.8. Comparative Visualization of Experimental Result

All comparison algorithms were trained and tested on the green tomato dataset, and the comparison experiments were conducted using MMsegmentation version 1.2.2. The segmentation results of each model are presented in Table III. It is observed that the optimized Swin-Unet algorithm demonstrates strong competitiveness across various evaluation metrics compared to the other algorithms.

Although segmentation accuracy is a key factor, the number of model parameters and computational complexity are also crucial for the overall quality of the model. For an input image size of 640 × 640, the number of parameters and computational complexity of each segmentation model are presented in Table IV. While the Precision and Dice metrics are slightly inferior to those of KNet and DANet, and the number of parameters and complexity are slightly higher than those of OCRNet, the optimized Swin-Unet model demonstrates strong overall performance , maintaining a balance between model capacity and computational efficiency.

From this analysis, it is evident that the optimized Swin-Unet algorithm demonstrates significant improvements across all assessment metrics. Although issues such as target miss-detection and unclear segmentation edges are encountered, this method effectively achieves precise segmentation of green tomato images compared to other semantic segmentation algorithms, thereby enhancing overall segmentation accuracy. To provide an intuitive comparison of the impact of different algorithms on segmentation results, this study conducted comparative experiments on green tomatoes under various conditions, including overlapping, occlusion, low-angle, long-shot, and backlighting scenarios. The specific visualization results are shown in Fig. 8.

While other algorithms face challenges such as target missing and misclassification when dealing with fruit parts under occlusion and overlapping scenarios, the improved Swin-Unet algorithm effectively addresses these issues and surpasses them in delineating fruit edge details. The visualization of comparative experiments shows that the improved algorithm excels in achieving clear boundary segmentation and maintaining the integrity of target fruits when capturing green tomatoes from various angles, such as aerial and long-distance views. Compared to other algorithms, it more accurately resolves these issues. Additionally, in backlighting scenarios, the improved algorithm demonstrates significant advantages over other models. The optimized Swin-Unet algorithm provides remarkable improvements by delivering more precise edge segmentation, reducing errors, and greatly enhancing overall performance. These experimental results validate the effectiveness of incorporating ASPP and AG structures to optimize the Swin-Unet algorithm specifically for green tomato segmentation in controlled agricultural environments.

This study demonstrates the refinement of the optimized Swin-Unet algorithm, which addresses complex challenges in facility agriculture applications. These challenges include variable lighting conditions, diverse shooting angles, and transitions between day and night environments. Green tomato images were captured under various conditions, such as occlusions, overlaps, and different lighting scenarios, to evaluate the algorithm's adaptability to these practical issues. The experimental results indicate that the algorithm achieves excellent segmentation accuracy across different scenarios, effectively managing images with varying lighting conditions and backgrounds while accurately identifying and segmenting occluded or overlapping fruits. These tests collectively confirm the algorithm's robustness, generalization capability, and theoretical sophistication, demonstrating its effectiveness in real-world applications. Beyond its impressive performance in green tomato segmentation, the optimized Swin-Unet algorithm shows potential for broader applications, including fruit segmentation in similar environments, thus supporting the wider adoption of automation techniques in facility-based agriculture.

IV. Conclusion

Addressing the challenges posed by the similar color features of green fruits and the background of branches and leaves, as well as issues of occlusion, overlapping, and varying lighting conditions that complicate segmentation, this thesis utilizes a custom dataset of green tomatoes. The semantic segmentation algorithm Swin-Unet is optimized by integrating the Atrous Spatial Pyramid Pooling (ASPP) module into the Bottleneck, which allows for the integration of features at different scales from the green tomato images, increases the receptive field, and significantly enhances the model's performance in processing edge details of green tomatoes. Additionally, the Attention Gate (AG) module is introduced at the Skip connection, enabling the model to focus on regions relevant to green tomatoes while suppressing irrelevant areas. Experimental results demonstrate that the optimized algorithm achieves relatively high accuracy in facility environments, with notably clearer edge segmentation of green fruits, thereby effectively improving the success rate of picking in real scenarios. Overall, the optimized Swin-Unet algorithm exhibits superior segmentation performance and stronger generalization ability, offering a theoretical reference for the segmentation of other green fruits. However, despite the significant progress made by the optimized Swin-Unet algorithm compared to other semantic segmentation algorithms, its increased model complexity remains a notable issue. This complexity is particularly prominent in application scenarios requiring fast processing or limited computational resources. Future research should focus on further reducing the complexity of Swin-Unet networks and exploring more lightweight network designs. Such efforts aim to mitigate the model's dependence on computational resources without compromising segmentation accuracy, thereby enabling the optimized Swin-Unet model to better adapt to various application requirements, especially those demanding high real-time performance and resource efficiency.

References

[1] Perveen R, Suleria H A R, Anjum F M, et al. Tomato (Solanum Lycopersicum) carotenoids and lycopenes chemistry; metabolism, absorption, nutrition, and allied health claims — a comprehensive review. Critical Reviews in Food Science and Nutrition, 2015, 55(7): 919-929.

[2] Tiwari J K, Singh A K, Behera T K. CRISPR/Cas genome editing in tomato improvement: advances and applications. Frontiers in Plant Science, 2023, 14: 1121209.

[3] Ugonna C U, Jolaoso M A, Onwualu A P. Tomato value chain in nigeria: issues, challenges and strategies. Journal of Scientific Research and Reports, 2015, 7(7): 501-515.

[4] Arad B, Balendonck J, Barth R, et al. Development of a sweet pepper harvesting robot. Journal of Field Robotics, 2020, 37, 1027 – 1039.

[5] Xiong Y, Ge Y, Grimstad L, et al. An autonomous strawberry - harvesting robot: design, development, integration, and field evaluation. Journal of Field Robotics, 2020, 37(2): 202-224.

[6] Jia W, Zhang Y, Lian J, et al. Apple harvesting robot under information technology: A review. International Journal of Advanced Robotic Systems, 2020, 17(3): 25310.

[7] Saranya N, Srinivasan K, Pravin Kumar S K, et al. Fruit classification using traditional machine learning and deep learning approach. Computational Vision and Bio-Inspired Computing: ICCVBIC, 2020: 79-89.

[8] Fujinaga T, Nakanishi T. Semantic segmentation of strawberry plants using DeepLabV3+ for small agricultural robot. International Symposium on System Integration. IEEE, 2023: 1-6.

[9] Marizuana Mat Daud, Zulaikha Kadim, and Hon Hock Woon, Detection of oil palm tree and loose fruitlets for fresh fruit bunch's ready-to-harvest prediction via deep learning approach. IAENG International Journal of Computer Science, vol. 50, no.4, pp1183-1193, 2023.

[10] Häni N, Roy P, Isler V. A comparative study of fruit detection and counting methods for yield mapping in apple orchards. Journal of Field Robotics, 2020, 37(2): 263-282.

[11] Bargoti S, Underwood J. Deep fruit detection in orchards. 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017: 3626-3633.

[12] Barth R, IJsselmuiden J, Hemming J, et al. Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation. Computers and Electronics in Agriculture, 2019, 161: 291-304.

[13] Kang H, Chen C. Fruit Detection, Segmentation and 3D visualisation of environments in apple orchards. Computers and Electronics in Agriculture, 2020, 171: 105302.

[14] Mo L, Fan Y, Wang G, et al. DeepMDSCBA: An improved semantic segmentation model based on DeepLabV3+ for apple images. Foods, 2022, 11(24): 3999.

[15] Matsui T, Sugimori H, Koseki S, et al. Automated detection of internal fruit rot in hass avocado via deep learning-based semantic segmentation of X-ray images. Postharvest Biology and Technology, 2023, 203: 112390.

[16] Roy K, Chaudhuri S S, Pramanik S. Deep learning based real-time industrial framework for rotten and fresh fruit detection using semantic segmentation. Microsystem Technologies, 2021, 27: 3365-3375.
[17] Li Q, Jia W, Sun M, et al. A novel green apple segmentation algorithm based on ensemble U-Net under complex orchard environment. Computers and Electronics in Agriculture, 2021, 180: 105900.
[18] He J, Duan J, Yang Z, et al. Method for segmentation of banana crown based on improved DeepLabv3+. Agronomy, 2023, 13(7): 1838.
[19] Yan C, Chen Z, Li Z, et al. Tea sprout picking point identification based on improved DeepLabV3+. Agriculture, 2022, 12(10): 1594.
[20] Bai Y, Guo Y, Zhang Q, et al. Multi-network fusion algorithm with transfer learning for green cucumber segmentation and recognition under complex natural environment. Computers and Electronics in Agriculture, 2022, 194: 106789.
[21] Liu X, Yu J, Kurihara T, et al. Hyperspectral imaging for green pepper segmentation using a complex-valued neural network. Optik, 2022, 265: 169527.
[22] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. European conference on computer vision, 2022: 205-218.
[23] Russell B C, Torralba A, Murphy K P, et al. LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, 2008, 77: 157-173.
[24] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]. European Conference on Computer Vision Zurich: Springer, 2014.
[25] Zhang W, Zhao Y, Guan Y, et al. Green apple detection method based on optimized yolov5 under orchard environment. Engineering Letters, 2023, 31(3): 1104-1113.
[26] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.
[27] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999, 2018.
[28] Liu R, Tao F, Liu X, et al. RAANet: a residual ASPP with attention framework for semantic segmentation of high-resolution remote sensing images. Remote Sensing, 2022, 14(13): 3109.
[29] Zhang J, Qin Q, Ye Q, et al. ST-unet: Swin transformer boosted U-net with cross-layer feature enhancement for medical image segmentation. Computers in Biology and Medicine, 2023, 153: 106516.
[30] Zheng Z, Liang E, Zhang Y, et al. A segmentation-based algorithm for classification of benign and malignancy thyroid nodules with multi-feature information. Biomedical Engineering Letters, 2024: 1-16.
[31] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 801-818.
[32] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.
[33] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2881-2890.
[34] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.
[35] Zhang W, Pang J, Chen K, et al. K-net: Towards unified image segmentation. Advances in Neural Information Processing Systems, 2021, 34: 10326-10338.
[36] Huang L, Yuan Y, Guo J, et al. Interlaced sparse self-attention for semantic segmentation. arXiv preprint arXiv:1907.12273, 2019.
[37] Chen Z, Zhu Y, Zhao C, et al. Dpt: Deformable patch-based transformer for visual recognition. ACM International Conference on Multimedia. 2021: 2899-2907.
[38] Yuan Y, Chen X, Chen X, et al. Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation. 2021[J]. 1909.
[39] Huang L, Yuan Y, Guo J, et al. Interlaced sparse self-attention for semantic segmentation. arXiv preprint arXiv:1907.12273, 2019.