Segmentation and recognition of filed sweet pepper based on improved self-attention convolutional neural networks

Weidong Zhu
of Jiangsu University
Jun Sun (sun2000jun@sina.com)
of Jiangsu University
Simin Wang
of Jiangsu University
Kaifeng Yang
of Jiangsu University
Jifeng Shen
of Jiangsu University
Xin Zhou
of Jiangsu University

Research Article

Keywords: Sweet pepper, Semantic segmentation, Recognition, Convolutional neural networks, Machine vision

Posted Date: March 30th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1484696/v1

Abstract

Automatic and accurate recognition of the parts to be picked is the key to realize the intelligent picking of sweet pepper. However, pepper fruits are always covered by other organs, and small objects like stems and shoots are difficult to be recognized by machines or cameras under certain extreme conditions. To accurately segment and recognize all kinds of objects in sweet pepper images captured at night, three experiments were performed in this paper, and an enhanced model based on convolutional neural networks (CNNs) was eventually achieved. In experiment I, several semantic segmentation networks were trained on a small dataset, and the full-resolution residual network (FRRN) was taken as a primary network. Then, the impact of resolution of input images on the segmentation effect was investigated in experiment II. In order to strengthen the feature presentation of inconspicuous objects, the position attention module (PAM) was appended on top of the FRRN in experiment III. This architecture was further trained to provide more precise segmentation results. The experimental result shows that the mean intersection over union (IoU) is 78.88%, which is at least 1.94 percentage points higher than other models, and the mean pixel accuracy (PA) is 97.94% on the test set. The proposed method has higher generalization performance when facing unforeseen picking situations, meanwhile it is generic and can be applied to other fruits and vegetables.

1 Introduction

Sweet pepper is one of the most favored vegetables in the world, which has rich nutritive value and edible value [1]. According to data from internet, the annual production of sweet pepper is more than 10 million tons worldwide [2]. Since the economic benefits of sweet peppers are substantial, it is essential to maximize production efficiency [3] and product quality. And now, the automated harvesting of sweet peppers [4, 5] is emerging as a key instrument to achieve high production efficiency in precision agriculture.

In practical production, most of the sweet peppers are cultivated in greenhouses, and the high temperature and humidity in greenhouses make it difficult to work in them. Replacing human beings with machines for picking work has been the major trend in modern precision agriculture [6]. With the fast development of robot technology, the research of greenhouse fruit and vegetable picking robot [7] has become a hot spot. In recent decades, picking robot technology has developed rapidly, and many countries have made a profound study on this technology and made some achievements [8]. However, the problems of weak adaptability, slow picking speed, and low recognition accuracy in the greenhouse environment still need to be solved. In particular, image recognition technology is essential for robot to identify targets quickly and accurately.

At present, researchers have made numerous researches on image recognition of fruits and vegetables. The main methods are based on the combination of machine learning algorithms and machine vision techniques. Wang et al. [9] used an improved Otsu segmentation algorithm to locate the cotton peach region and sampled RGB values of the pixels in different regions. Thus, the problem of image segmentation was transformed into the problem of pixel classification. As a result, they trained the ELM classification model and realized accurate segmentation of the cotton peach images. In the work of Dhingra et al. [10], the recognition of leaf diseases was studied. An extended segmentation technique based on neutron logic was used to extract the region of interest (ROI) of leaves. Then several features, including texture, color, and histogram, were employed to detect diseased leaves. Nine different classifiers were constructed, and the best classification accuracy of 98.4% was eventually obtained. To recognize and locate cucumber accurately, Bao et al. [11] proposed a multi-template matching method and

established a multi-template matching library containing 65 cucumber images. The robot vision system can utilize the matching library to calculate the normalized correlation coefficient matrix of the target image one by one, and judge whether there is a target cucumber in the image. The aforementioned researches prove that image recognition technology has been widely used in the agricultural field to identify and locate vegetables accurately. However, most of the above methods adopt traditional machine learning, which mainly depends on multiple feature selection and application under specific background. It is difficult to play a role in the environment of the greenhouse.

With the development of high-performance computing systems and the sharing of massive datasets, deep learning has gradually become one of the most important technologies in the field of intelligent agriculture. The basic deep learning tool used in this work is convolutional neural networks (CNNs), which have a strong feature extraction capability and are suitable for processing a large amount of input images in parallel. Therefore, CNN is one of the most powerful technologies in the field of image recognition, which is widely used in image classification [12, 13], object detection [14, 15] and image segmentation [16, 17]. In recent years, CNN technology has been introduced into the field of fruit and vegetable image recognition. Zhao et al. [18] proposed an apple location method based on YOLOv3, which realized apple detection in a complex environment. This work provided a theoretical basis for the research and development of apple harvesting robot. Fu et al. [19] studied the problem of fast identification of multiple clusters of Kiwifruit in the field. They trained LeNet as a classification model to identify occluded fruits, overlapping fruits, adjacent fruits, and independent fruits, and finally achieved an average recognition rate of more than 85%. In the work presented by Yang et al. [20] based on the Mask-RCNN, Resnet-50 was adopted as a backbone network. On this basis, the feature pyramid network (FPN) architecture was integrated to construct the feature extraction network. Their method accurately located strawberry picking points, which was helpful to improve the working performance of strawberry picking robot. The aforementioned literatures demonstrate that it is feasible to recognize fruit and vegetable by convolutional neural networks instead of traditional machine learning.

In this paper, an improved CNN architecture based on the FRRN was designed and verified, which was applied to segment sweet pepper images captured at night greenhouse. Specifically, PAM was appended on top of the FRRN to improve the feature representation, which contributes to more precise segmentation results. Furthermore, the influence of image resolution on segmentation results was explored. This study aimed to provide a theoretical basis for the development of image recognition performance of picking robot in precision agriculture.

The rest of this paper is organized as follows. The dataset and experiment arrangements were introduced in Section 2. Section 3 describes the model architecture of the FRRN combined with the PAM in detail. The training details of the model are listed in Section 4. The experimental results are analyzed and discussed in Section 5. The conclusions are drawn in Section 6.

2 Materials And Experimental Scheme

2.1 Dataset description

The sweet pepper dataset, used in this paper, was obtained from 4TU. 10 500 greenhouse sweet pepper images and their corresponding ground truth labels can be obtained at ResearchData (https://data.4tu.nl). These synthetic images were rendered through Blender based on 21 empirically measured plant properties. They highly

simulated empirical images, which obtained by sweet pepper harvest robot from different angles at night greenhouse [21]. Besides, the sweet pepper dataset also contains 50 empirical images of the crop obtained from a high-tech commercial greenhouse. These empirical images were used in experiment III. The sweet pepper dataset was publicly released with the intention of comparing the performance of agricultural computer vision methods. There are 8 classes of segmentation objects in the picture, including background, leaves, peppers, peduncles, stems, shoots, wires, and cuts for picking pepper. In the labeled images, 8 classes were annotated on a per-pixel level, corresponding to the color of black, blue, yellow, red, pink, green, white, and light blue. In Fig. 1, examples of synthetic images are shown.

2.2 Overview of experiment arrangement

To obtain the optimal segmentation model of the sweet pepper image, three main experiments were designed and carried out in this study. Overview of performed main experiments I through III are shown in Table 1.

Experiments I: A suitable CNN architecture was picked out for sweet pepper segmentation. Since the same network structure may perform differently on different datasets, in a specific task, it is necessary to select the appropriate network according to the dataset. Existing semantic segmentation network models are mostly applied to segmentation of street scenes, and rarely used in agricultural scene. In this experiment, several model structures were applied on the same sweet pepper dataset for training and testing, and the differences in segmentation performance of different models were analyzed. Each network structure is verified with the same number of images, 400 images were used as training set, and 100 images were used as validation set and test set respectively. Since different network configurations has different requirements for the size of the input images, the size of the input images was set to 384×384 to balance the training time and model performance.

Experiments II: The effect of image resolution on segmentation results was investigated. After experiment I, a suitable model architecture for the sweet pepper segmentation task was obtained. Then the impact of the input image size on the segmentation performance of the model was further explored. Images with different sizes were input into the model for training. Unlike experiment I, experiment II was performed on a larger dataset. In this experiment, 10 500 sweet pepper images were divided into training set, validation set and test set according to 3:1:1, specifically, 6 300 images in training set, 2 100 images in validation set and test set respectively.

Experiments III: The self-attention mechanism was introduced into the network to improve the segmentation performance. The self-attention mechanism can effectively establish dependencies between remote features, and make the model focus on extracting some key features. In our work, the PAM was used to capture global dependencies in the spatial dimensions. It was added on top of the FRR. Afterwards, the enhanced model was trained and tested on our dataset. The size of the input image and the distribution of the dataset in experiment III were the same as that in experiment II. Finally, the model with the best segmentation performance was further evaluated on empirical (real) images.

Table 1
Overview of performed main experiments I through III

Experiment	Methods/Architecture	Input size/pixel	Training set	Validation set	Test set
I-A ~ I-E	Several CNN-based models	384×384	400	100	100
II-A	FRRN	384×384	6 300	2 100	2 100
II-B	FRRN	512×512	6 300	2 100	2 100
III-A	FRRN + PAM	384×384	6 300	2 100	2 100
III-B	FRRN + PAM	512×512	6 300	2 100	2 100

3 Segmentation And Recognition Model

3.1 Semantic segmentation network architecture

Semantic segmentation method was used to segment and recognize sweet pepper in this paper. The goal of semantic segmentation is to segment and parse an image into different image regions associated with semantic categories (e.g. leaf, sweet pepper, shoot). Semantic segmentation models based on convolutional neural network have attracted extensive attention since FCNs [22] was proposed, and have made great progress. The existing depth-based semantic segmentation methods are mainly divided into two kinds: one is based on regional classification represented by Mask R-CNN [23], the other is based on pixel classification represented by FCNs. The former method is based on target detection. Firstly, the image is divided into different image patches, then each pixel in the image block is semantically classified. Finally, semantics segmentation is realized. This kind of method first appeared in R-CNN and gradually evolved into improved algorithms such as Faster R-CNN [24] and Mask R-CNN. The advantage of this method is that it can accomplish both tasks of target detection and semantic segmentation at the same time. However, due to the lack of global information of the image, this method is not effective for small-scale objects and small areas. Therefore, this method was not considered in our experiments. The semantic segmentation method based on pixel classification is favored by many researchers and has been widely used. In particular, dilation convolution, as well as atrous spatial pyramid pooling was adopted in the Deeplab series [25, 26] to embed contextual features of different scales. Moreover, the encoder-decoder structures can effectively fuse mid-level and high-level semantic features. In experiment I, since the methods based on pixel classification performed better in fine segmentation task, we mainly adopted them.

3.2 Full-Resolution Residual Networks

This section briefly introduces the FRRN, which is the backbone of the improved model in this paper. FRRN is a novel network structure similar to ResNet [27]. It does not rely on the pre-trained networks and can start training from scratch. In addition, since FRRN combines two distinct processing streams, it has strong recognition performance and precise localization capabilities. As shown in Fig. 2, one stream is pooling stream: semantic information is captured through a sequence of convolution and pooling operations for identification. The other stream is a residual stream: feature maps at the full image resolution are carried to achieve precise boundary adherence. Meanwhile, FRRN consists of a series of full resolution residual units (FRRUs). The structure of FRRU is given in Fig. 3. Each unit includes two inputs and outputs, because they are applied to both streams at the

same time. If z_{n-1} is the residual input of the *n*th FRRU and y_{n-1} is pooling input. Then the outputs are computed as:

$$z_n = z_{n-1} + H\left(y_{n-1}, z_{n-1}; W_n\right) \quad (1)$$

$$y_n = G\left(y_{n-1}, z_{n-1}; W_n\right) \quad (2)$$

where *H(·)* and *G(·)* represent residual calculation and pooling operation respectively. $W_n$is weight matrix.

3.3 Position attention module

The attention mechanism in deep learning is generated by imitating human selective visual attention mechanism. It has been widely used in image processing [28], speech recognition [29] and natural language [30] in recent years. The essence of attention mechanism lies in the automatic and efficient allocation of attention resources, which contributes to the acquisition of the most critical information to solve the current task. There are many variants of the attention model, including the soft attention model, the global attention model [31], and the key-value attention model [32]. PAM can adaptively aggregate long-range context information to improve the feature representation of semantic segmentation. The implementation details of PAM are shown in Fig. 4. A represents a local feature, where A $\in R^{C\times H\times W}$. It is first fed into a convolution layer to generate two new feature maps B and C, respectively, where {B,C}$\in R^{C\times H\times W}$. Then they are reshaped to$R^{C\times N}$, where $N = H \times W$is the number of pixels. After that, a matrix multiplication is performed between transposed B and C, and a softmax layer is applied to calculate the spatial attention map M $\in R^{N\times N}$:

$$\boldsymbol{M}_{ji} = \frac{exp\left(\boldsymbol{B}_i^T \cdot \boldsymbol{C}_j\right)}{\sum_{i=1}^{N} exp\left(\boldsymbol{B}_i^T \cdot \boldsymbol{C}_j\right)} \quad (3)$$

where M_{ji}measures i^{th}position's impact on j^{th}position. The more similar feature representations of two positions contribute to a greater correlation between them.

Meanwhile, feature A is also fed into a convolution layer to generate a new feature map D $\in R^{C\times H\times W}$and is reshaped to$R^{C\times N}$. Then a matrix multiplication is performed between D and M, and the result is reshaped to $R^{C\times H\times W}$. Finally, it is multiplied by a scale parameterα, and an element-wise sum operation is performed between it and features A to obtain the final output Q $\in R^{C\times H\times W}$as follows:

$$\boldsymbol{Q}_j = \alpha \sum_{i=1}^{N} \left(\boldsymbol{D}_i \boldsymbol{M}_{ji}\right) + \boldsymbol{A}_j \quad (4)$$

where $\alpha$is initialized as 0 and gradually assigned more weight.

3.4 Attention module embedding with full-resolution residual networks

It can be inferred from Eq. 4 that the resulting feature Q at each position is a weighted sum of the features across all positions and original features. Accordingly, the PAM provides a global contextual view and selectively aggregates contexts according to the spatial attention map. In the sweet pepper images, some ‘stem’ and ‘shoot’ are inconspicuous or incomplete objects because of the influence of lighting and view. Those dominated salient objects (e.g. leaf, sweet pepper) would harm those inconspicuous object labeling. Since the PAM can selectively aggregate the similar features of inconspicuous objects to highlight their feature representations and avoid the influence of salient objects, we appended the PAM on top of the FRRN. The proposed FRRN-based network structure for semantic segmentation of the sweet pepper image is shown in Fig. 5.

As illustrated in Fig. 5, FRRN was employed as the backbone. The whole network structure mainly refers to Encoder-Decoder [33] structure. When the network starts working, the input feature maps are down-sampled through four pooling layers to extract the deep features in the images. After that, 4 up-sample layers are used to restore the size of the feature maps. And the mapping relationship between the original image and the label is obtained. Then the features from the FRRN would be fed into the PAM. Specifically, PAM is cascaded behind the concatenate (fusion) layer. The output features of the attention module are transformed by the convolution layer and input into the last convolution layer to generate the final prediction map. Besides, the notations RU and FRRU in Fig. 5 refer to the residual unit and the full-resolution residual unit, and the numbers on the lower right corner represent the number of convolution kernels in the unit, respectively. Also, the number of convolution kernels of each convolution layer in Fig. 5 is 48, 32, and 8 (number of classes to predict). The sizes of convolution kernels are 5×5, 1×1, and 1×1, respectively. The model architecture is proposed to contribute to the field of agricultural segmentation task.

4 Model Training

4.1 Experiment platform

The experiments in this work were conducted on the Ubuntu 16.10 LTS 64bit system. All the methods were implemented based on Tensorflow, and the TF-Slim library provided main functions for building model framework. The computer is equipped with 16 GB RAM and Intel Core i7-7700K CPU. Meanwhile, NVIDIA GTX1080Ti GPU acceleration technology was applied to improve the training speed.

4.2 Implementation Details

The RMSProp optimization algorithm was adopted during the process of training models. For the experiment I, the base learning rate was set to 1e-4, and the weight decay coefficient was set to 0.995. For the remaining experiments, the base learning rate and attenuation coefficient were set to 1e-5 and 0.995, respectively. Batch size was only set to 1 for all datasets because of the limitation of GPU memory and computing power. The number of iterations was set to 300 epochs for the experiment I and 200 epochs for the other experiments. The loss function used in our study was the softmax cross-entropy function. Besides, the data augmentation technology was not adopted in the work of this paper.

The mean IoU is the most important segmentation evaluation metric, which is computed as follows:

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \quad (5)$$

where k represents different classes, p_{ii} is the number of pixels correctly predicted, $\sum_{j=0}^{k} p_{ij}$ is the number of pixels in which class *i* is predicted to be class *j*.

The precision, recall, and *F*1 score were introduced as additional evaluation indexes to evaluate the segmentation effect of sweet pepper images more objectively. In general, the higher and closer the precision and recall, the better the classification recognition effect. The *F*1 score is computed based on the mean per-class precision/recall results as follows:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (6)$$

5 Results And Discussion

5.1 Evaluation of experiment I

In experiment I, several existing semantic segmentation methods were evaluated and compared on a small sweet pepper dataset. The mean pixel accuracy (PA) and mean intersection over union (IoU) of the trained models on the test set are shown in Table 2. These model architectures achieved good segmentation results when used for sweet pepper segmentation in this paper. It is worth noting that most of these methods were training from scratch, rather than based on the pre-trained models. The approach based on pre-trained models was also adopted in our experiments. The results showed that although the training time of this method was less than other means, the segmentation effect was far inferior to those methods training from scratch. Consequently, the results of approach based on pre-trained models were not shown and compared in this paper.

As shown in Table 2, the mean PA and mean IoU of FRRN are 94.3% and 69.18%, respectively. In particular, its recognition accuracy for sweet peppers reached 93.2%. The segmentation and recognition effect of FRRN is better than that of other methods. From the perspective of pixel classification, the mean PAs of the five methods are similar, and all of them exceed 90%. It illustrates that the five networks have achieved good performance in pixel classification, especially for salient objects (e.g. leaf, sweet pepper). Meanwhile, the mean PAs of sub-experiments I-A, I-B, and I-E are slightly higher than that of the other two methods. From Table 2, it can be seen that the main difference is caused by the small objects, including peduncles, stems, shoots, and wires. The PA of wires in experiment I-C was 6.30%, while that of peduncle in experiment I-D was only 6.00%. FRRN performs better than other methods, which verifies that it is more suitable for small-scale object recognition and fine segmentation. Moreover, the effects of several methods can be visualized in Fig. 6. These visualizations show that the FRRN achieved better segmentation results, both numerically and qualitatively. However, the segmentation effect of FRRN for sweet pepper needs to be improved, especially for the pixel-level recognition of inconspicuous objects. Hence, we took experiment II and experiment III.

Table 2
Results of each category on a small sweet pepper test set

Methods	Pixel accuracy/%									Mean IoU
	Mean PA	Background	Leaf	Pepper	Peduncle	Stem	Shoot	Wire	Cut	
I-A-MobileUNet [34]	93.6	**93.8**	97.5	**93.5**	46.4	81.1	64.5	73.5	59.6	67.82
I-B-SegNet [33]	93.7	93.8	96.9	94.1	50.3	80.3	63.5	73.2	63.7	67.33
I-C-BiseNet [35]	90.7	90.9	96.2	90.8	44.1	73.2	40.9	**6.30**	54.5	53.78
I-D-ED [33]	92.0	92.3	96.7	88.2	**6.00**	73.6	55.7	55.5	50.6	55.80
I-E-FRRN	**94.3**	93.7	**97.6**	93.2	**53.1**	**81.2**	**68.3**	69.5	**67.5**	**69.18**

5.2 Evaluation of experiment II

Semantic segmentation belongs to pixel-level tasks, which need to provide pixel by pixel output. Consequently, semantic segmentation requires higher resolution and more detailed information. It is supposed that the higher resolution of the input images, the richer feature information that the convolutional neural network could capture. High-resolution images help to improve the recognition accuracy and segmentation performance of the model. Especially for fine agriculture segmentation, higher image resolution is beneficial. However, if the resolution is extremely high, maybe it causes the problem of slow training speed. In this study, the resolution of the original images in the dataset is 800×600, but the general model cannot directly perform feature analysis and extraction on such images. To balance the relationships between segmentation effect and computational complexity, the original image was cropped to 384×384 and 512×512 in experiment II, and train FRRN on images of different sizes.

With the purpose of verifying the effect of input image resolution on segmentation results, we conducted experiment II. As shown in Table 3, the higher segmentation performance was achieved by the same model after inputting high-resolution images. Compared with the input size of 384×384, employing an input image size of 512×512 yielded a result of 76.81% in mean IoU, which bringed 3.14% improvement. This shows that the increase in resolution contributes to the refinement of segmentation results. Besides, when the size of the input image is 512×512, the mean PA and the $F1$ score further improves to 95.47% and 95.57%, respectively. Meanwhile, although the improvement effect of pixel classification is not significant, the precision and recall have been kept at a high level and close to each other, which indicates that the classification effect of the model has improved. Results show that, in a certain range, the higher the resolution of the input images, the better the segmentation performance of FRRN.

Table 3
Evaluation and test results of different methods

Method(model)	Size of input image/pixel	Mean PA/%	Mean IoU/%	Precision/%	Recall/%	F1 score/%
FRRN	384×384	95.10	73.67	95.55	95.10	95.22
FRRN	512×512	95.47	76.81	95.97	95.47	95.57
FRRN + PAM	384×384	97.38	76.45	97.64	97.38	97.51
FRRN + PAM	512×512	97.94	78.88	98.20	97.94	98.07
Barth et al. [36]	300×300	-	55.00	-	-	-

5.3 Evaluation of experiment III

The PAMs were employed on top of the FRRN to capture long-distance dependencies for better feature presentation. Similarly, two different resolution images were used as input to train the improved network. Various evaluation indexes achieved by the model on the test set are shown in Table 3. The segmentation performance of PAM on two different resolution images is shown in Table 3. It is observed that the best performance was achieved at an input image resolution of 512×512. Compared with the baseline FRRN, the employment of PAM yielded a result of 78.88% in mean IoU, which brought an improvement of 2.07 percentage points. Also, mean PA increased by 2.47 percentage points. Meanwhile, the precision, recall, and *F*1 score have also improved. In conclusion, results show that attention modules bring benefits to sweet pepper segmentation.

Quantitative evaluation is not enough to verify the importance of PAM, thus the effects of PAMs were visualized in Fig. 7. As expected in Section 3.4, the PAM helps to strengthen the feature presentation of some inconspicuous objects. It can be observed from Fig. 7 that the employment of PAM makes some segmentation details and object boundaries clearer. As shown by the red circles in the first row, it is challenging to segment small-scale objects in detail, but under the action of PAM, the segmentation effect of the third column is significantly better than that of the second column. Specifically, in the second row, the FRRN model with PAM is better than the original FRRN model for the segmentation of small objects such as leaf stems. Therefore, Fig. 7 demonstrates that selective fusion over local features enhances the discrimination of details and refines the segmentation results. Besides, compared with the research of Barth et al. [36], the accuracy of pepper segmentation has been improved obviously.

To explore the generalization performance exhibited by a synthetically trained model when faced with similar data in the same domain, the best performing model (model of experiment III-B) was evaluated on 20 empirical (real) images. The segmentation results of the proposed model on empirical data are shown in Fig. 8. The model failed to discriminate and segment all the classes properly. This result confirms our hypothesis that the synthetically bootstrapped model could not generalize accurately to empirical data without fine-tuning. Note that these empirical images were not used for training or validation. In order to prove our hypothesis, 30 empirical images were used to fine-tune the proposed model, and the generalization performance of the model was tested on 20 empirical photos. As illustrated in Table 4, this practice resulted in increased performance, with a mean IoU of 59.81%. The result implies fine-tuning on a synthetically trained network can generalize to similar data in the same domain (empirical images).

Note that, for the empirical dataset, due to the influence of dark regions, some objects that were not manually annotated in the ground truth would be detected in the images and have a certain impact on the segmented images. Hence, although these parts were true positive, they were still evaluated as false positives. This annotation bias resulted in a lower segmentation performance of empirical images. In short, the results of experiment III show that the proposed optimization model of FRRN combined with PAM is feasible and effective for the recognition and segmentation of sweet pepper.

Table 4
The evaluation results of the proposed model on empirical (real) images

model	fine-tuning	Mean PA/%	Mean IoU/%
FRRN + PAM		73.42	51.63
FRRN + PAM	√	80.56	59.81

5.4 Results compared with other models

We further compared our method with existing methods on the sweet pepper testing set. The results are shown in Table 5. The FRRN + PAM model performs better than other segmentation methods. It is supposed that these networks are designed for street scene segmentation tasks, so they maybe not perform well in agricultural segmentation tasks. Among these methods, in order to achieve real-time performance, only ENet refrained from using a pre-trained network when designing the structure, so it did not get high scores. Specifically, a pre-trained network was abandoned by our network either, but it outperformed ENet by a large margin. Moreover, it also surpassed DANet, which uses a complex backbone ResNet-101.

Table 5
IoU scores on sweet pepper test set

Method	Mean IoU/%
ENet [37]	65.78
BiseNet [35]	69.75
DANet [38]	76.94
Our method (FRRN + PAM)	78.88

6 Conclusions

In this paper, CNN was applied to the semantic segmentation of the sweet pepper images. The FRRN model was improved by combining with attention mechanism. Specifically, PAMs were embedded at the top of the network to capture global dependencies in the spatial dimensions. This new architecture is clean, requires no additional post-processing, and can be trained from scratch. The proposed network (FRRN + PAM) was trained to segment sweet pepper images captured at night. Relevant experimental results show that PAMs contribute to providing more precise segmentation results. Meanwhile, our experiments indicate that proper image resolution benefits for segmentation performance. Besides, the synthetically trained model combined with fine-tuning can also be well generalized to the similar data in the same domain (empirical dataset). Our method achieves good performance on the test set of sweet pepper, with a mean IoU of 78.88%. The proposed method is generic and can be applied

for other fruits and vegetables. It is hoped that this study can provide a theoretical basis for the development of image recognition in precision agriculture.

Declarations

Acknowledgements This work is partly supported by Project of Agricultural Equipment Department of Jiangsu University (NZXB20210210), a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) and Jiangsu University undergraduate scientific research project (20AB00).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

References

1. López-Marín, J., Gálvez, A.: Selecting vegetative/generative/dwarfing rootstocks for improving fruit yield and quality in water stressed sweet peppers. Sci. Hortic. **214**, 9–17 (2017)
2. Gilanie, G., Nasir, N., Bajwa, U.I. et al.: RiceNet: convolutional neural networks-based model to classify Pakistani grown rice seed types. Multimed. Syst. **27**, 867–875 (2021)
3. Ghimire, S., Shakya, S.M., Srivastava, A.: Effects of organic manures and their combination with urea on sweet pepper production in the mid-hills. J. Agric. Environ. **14**, 23 (2018)
4. Bac, C.W., Hemming, J., Van Tuijl, J., Barth, R., Wais, E., Van Henten, E.J.: Performance evaluation of a harvesting robot for sweet pepper. J. Field Robot. **34**(6), 1123–1139 (2017)
5. Lehnert, C., English, A., Mccool, C., Tow, A.W., Perez, T.: Autonomous sweet pepper harvesting for protected cropping systems. IEEE Robot. Autom. Lett. **2**(2), 872–879 (2017)
6. Luo, Y.Q., Sun, J., Shen, J.F., Wu, X.H., Wang, L., Zhu, W.D.: Apple leaf disease recognition and sub-class categorization based on improved multi-scale feature fusion network. IEEE Access **9**, 95517–95527 (2021)
7. Ding, X.: Pepper picking robot in greenhouse. Agric. Eng. Technol. **39**(1), 78–82 (2019).
8. Sun, J., He, X.F., Ge, X., Wu, X.H., Shen, J.F., Song, Y.Y.: Detection of key organs in tomato based on deep migration learning in a complex background. Agriculture-Basel **8**, 196–210 (2018)
9. Wang, J., Zhou, Q., Yin, A.: Self-adaptive segmentation method of cotton in natural scene by combining improved Otsu with ELM algorithm. Trans. Chin. Soc. Agric. Eng. **34**(14), 173–180 (2018)
10. Dhingra, G., Kumar, V., Joshi, H.D.: A novel computer vision based neutrosophic approach for leaf disease identification and classification. Meas. **135**, 782–794 (2019)
11. Bao, G., Cai, S., Qi, L., Xun, Y., Zhang, L., Yang, Q.: Multi-template matching algorithm for cucumber recognition in natural environment. Comput. Electron. Agric. **127**, 754–762 (2016)
12. Liu, C., Cao, Y., Luo, Y., Chen, G.L., Vokkarane, V., Ma, Y.S.: DeepFood: deep learning-based food image recognition for computer-aided dietary assessment. Lect. Notes Comput. Sci. 2016, 37–48 (2016)
13. Sun, J., Tan, W., Mao, H.: Recognition of multiple plant leaf diseases based on improved convolutional neural network. Trans. Chin. Soc. Agric. Eng. **33**(19), 209–215 (2017)
14. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware CNN model. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1134–1142 (2015)

15. Sun, J., Yang, Y., He, X.F., Wu, X.H.: Northern maize leaf blight detection under complex field environment based on deep learning. IEEE Access **8**, 33679–33688 (2020)
16. He, D., Xie, C.: Semantic image segmentation algorithm in a deep learning computer network. Multimed. Syst. (2020)
17. Olimov, B., Sanjar, K., Din, S. et al.: FU-Net: fast biomedical image segmentation model based on bottleneck convolution layers. Multimed. Syst. **27**, 637–650 (2021)
18. Zhao, D., Wu, R., Liu, X.: Apple positioning based on YOLO deep convolutional neural network for picking robot in complex background. Trans. Chin. Soc. Agric. Eng. **35**(3), 172–181 (2019)
19. Fu, L., Feng, Y., Elkamil, T.: Image recognition method of multi-cluster kiwifruit in field based on convolutional neural networks. Trans. Chin. Soc. Agric. Eng. **34**(2), 205–211 (2018)
20. Yang, Y., Kailiang, Z., Li, Y.: Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Comput. Electron. Agric. **163**, 104846 (2019)
21. Barth, R., Ijsselmuiden, J., Hemming, J., Henten, E.J.V.: Data synthesis methods for semantic segmentation in agriculture: a capsicum annuum dataset. Comput. Electron. Agric. **144**, 284–296 (2018)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2014)
23. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, pp. 2980–2988 (2017)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
25. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2016)
26. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA (2017)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp. 770–778 (2016)
28. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA (2017)
29. Bahdanau, D., Chorowski, J., Serdyuk, D.: End-to-end attention-based large vocabulary speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Shanghai, China, pp. 4945–4949 (2016)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS), MIT Press, Long Beach, CA, USA, pp. 5998–6008 (2017)
31. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA (2015)
32. Michał, D., Tim, R., Johannes, W.: Frustratingly short attention spans in neural language modeling. In: 5th International Conference on Learning Representations (ICLR), Toulon, France (2017)

33. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for scene segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
34. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T.: Mobilenets: efficient convolutional neural networks for mobile vision applications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
35. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), Springer, Munich, Germany, pp. 334–349 (2018)
36. Barth, R., Ijsselmuiden, J., Hemming, J., Van Henten, E. J.: Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation. Comput. Electron. Agric. **161**, 291–304 (2017)
37. Paszke, A., Chaurasia, A., Kim, S.: ENet: A deep neural network architecture for real-time semantic segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (2016), IEEE, Las Vegas, NV, USA (2016)
38. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z.: Dual attention network for scene segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, pp. 3146–3154 (2019)

Figures

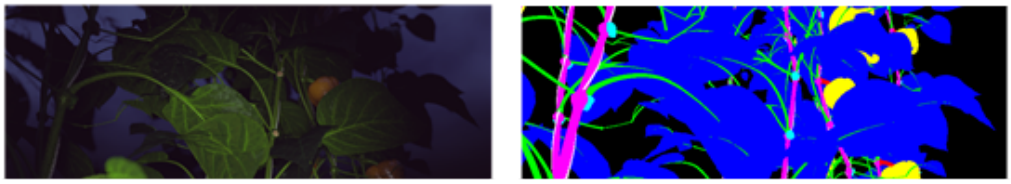

Figure 1

Examples of sweet pepper images and their corresponding ground truth labels.

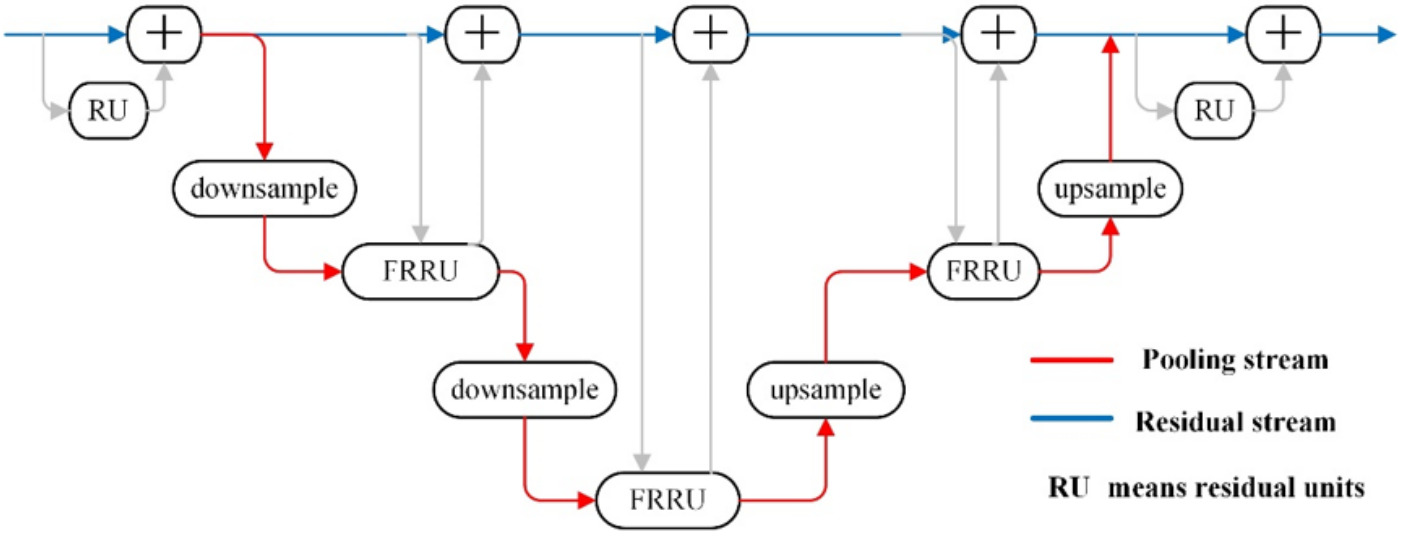

Figure 2

The structure of full-resolution residual networks (FRRN)

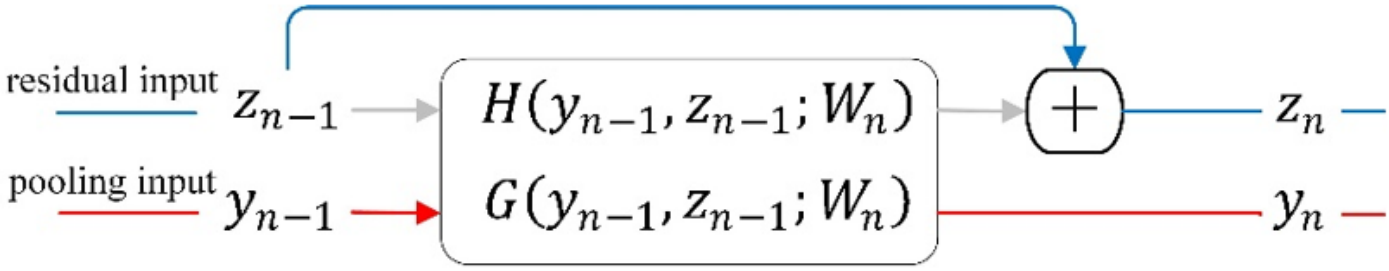

Figure 3

The structure of full-resolution residual unit (FRRU)

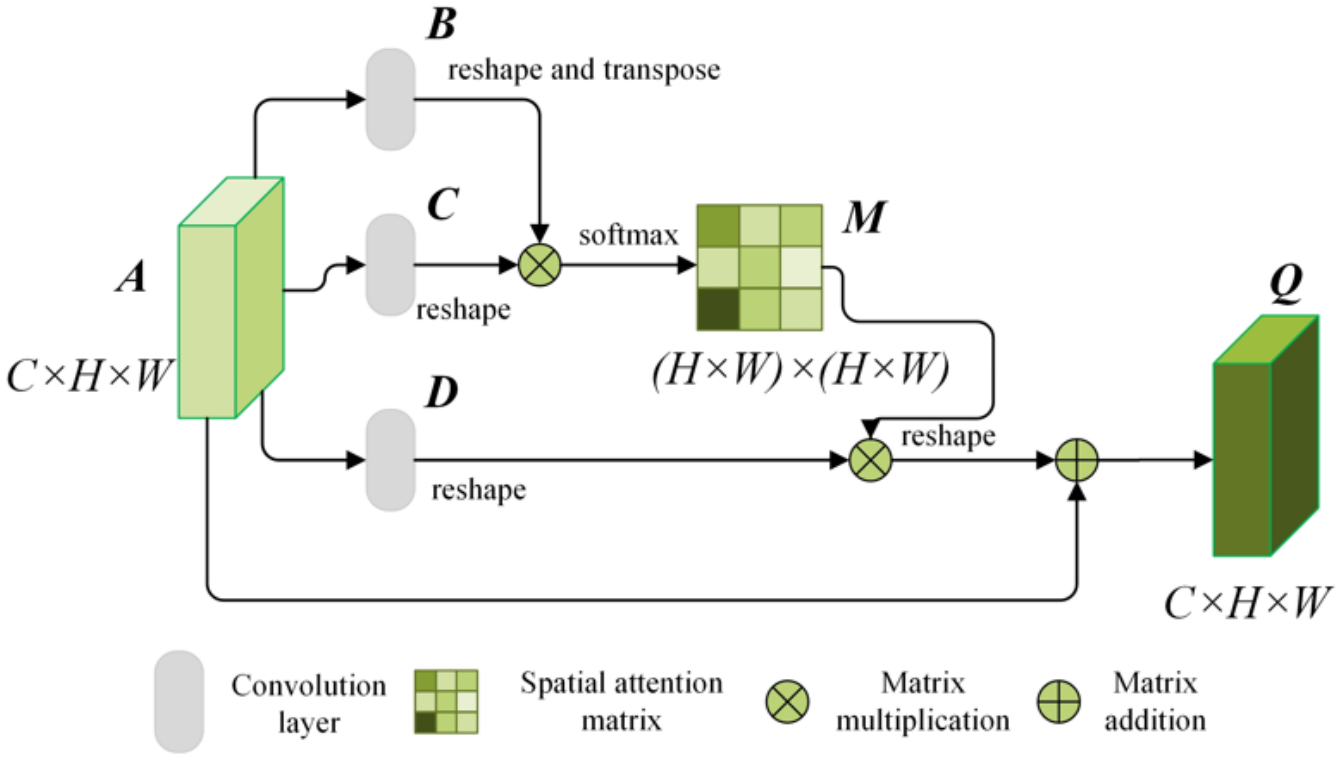

Figure 4

The details of position attention module

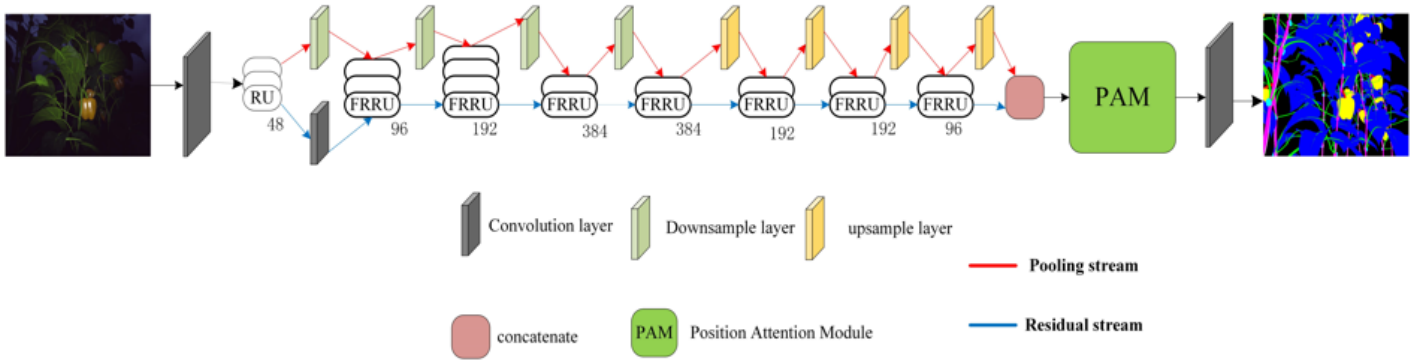

Figure 5

Overview of FRRN+PAM architecture

Figure 6

Qualitative comparison on the test set

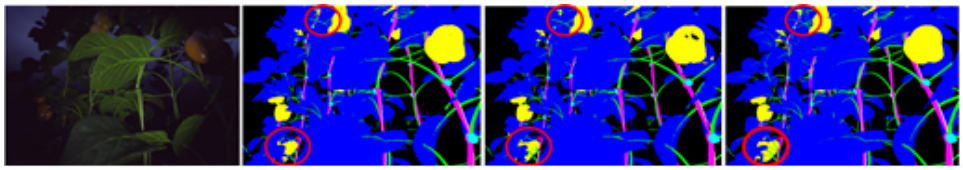

Figure 7

Visualization results of position attention module on test set

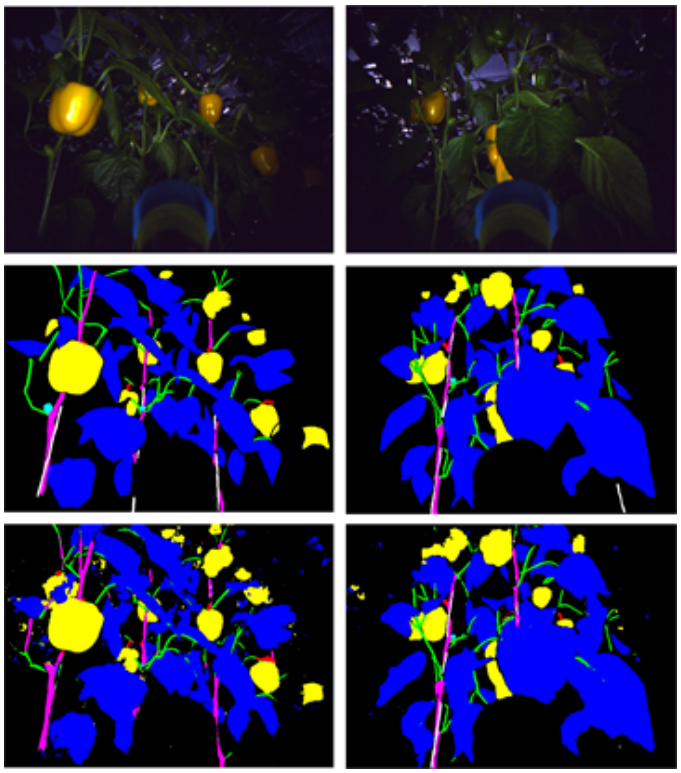

Figure 8

Example segmentation results of empirical test set without fine-tuning (left column) and results of empirical test set with fine-tuning (right column). Color images (top row), ground truth (middle row) and classification segmentation (bottom row) are shown