Spatial-channel transformer network based on mask-RCNN for efficient mushroom instance segmentation

Jiaoling Wang[1,2,4], Weidong Song[2], Wengang Zheng[3], Qingchun Feng[3], Mingfei Wang[3], Chunjiang Zhao[1,3*]

(1. *Northwest Agriculture and Forestry University, Xi'an 712199, China*;
2. *Nanjing Institute of Agricultural Mechanization, Ministry of Agriculture and Rural Affairs, Nanjing 210014, China*;
3. *Intelligent equipment Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China*;
4. *Zhejiang Provincial Key Laboratory of Agricultural Intelligent Equipment and Roboics/College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China*)

Abstract: Edible mushrooms are rich in nutrients; however, harvesting mainly relies on manual labor. Coarse localization of each mushroom is necessary to enable a robotic arm to accurately pick edible mushrooms. Previous studies used detection algorithms that did not consider mushroom pixel-level information. When these algorithms are combined with a depth map, the information is lost. Moreover, in instance segmentation algorithms, convolutional neural network (CNN)-based methods are lightweight, and the extracted features are not correlated. To guarantee real-time location detection and improve the accuracy of mushroom segmentation, this study proposed a new spatial-channel transformer network model based on Mask-CNN (SCT-Mask-RCNN). The fusion of Mask-RCNN with the self-attention mechanism extracts the global correlation outcomes of image features from the channel and spatial dimensions. Subsequently, Mask-RCNN was used to maintain a lightweight structure and extract local features using a spatial pooling pyramidal structure to achieve multiscale local feature fusion and improve detection accuracy. The results showed that the SCT-Mask-RCNN method achieved a segmentation accuracy of 0.750 on segm_Precision_mAP and detection accuracy of 0.638 on Bbox_Precision_mAP. Compared to existing methods, the proposed method improved the accuracy of the evaluation metrics Bbox_Precision_mAP and segm_Precision_mAP by over 2% and 5%, respectively.

Keywords: edible mushrooms, picking, instance segmentation, deep learning, algorithm
DOI: 10.25165/j.ijabe.20241704.8987

Citation: Wang J L, Song W D, Zheng W G, Feng Q C, Wang M F, Zhao C J. Spatial-channel transformer network based on mask-RCNN for efficient mushroom instance segmentation. Int J Agric & Biol Eng, 2024; 17(4): 227–235.

1 Introduction

Edible mushrooms are rich in nutrients and have become the fifth largest crop in China after food grains, oil crops, fruits, and vegetables; they play a crucial role in supporting national food security and ensuring a steady supply of essential agricultural products[1]. The 20th Party Congress proposed the construction of a diversified food supply system, cultivation and growth of edible mushrooms, and development of straw and forest plantations[2]. The cultivation method of edible mushrooms can be classified into factory and field planting according to different varieties; the golden needle and almond abalone mushrooms are appropriate for factory production, whereas newly introduced rare species, such as the large globe and sheep maw mushrooms, are effective incentives for farmers using industrial mining benefits. These initiatives extend the industrial chain, promote scientific development, and support comprehensive planting plans. At present, only large fields can be used for cultivation. With the rapid development of industries, such as those associated with large bulbous caps and morel mushrooms, planting areas are rapidly increasing; however, all these areas are labor-intensive and inefficient. Therefore, picking robots must be developed; however, the complex environment of large fields and objective factors, such as mutual contact and shading between individual mushrooms, considerably increase the difficulty of mushroom pose recognition. Therefore, the accurate identification and determination of the position and pose of mushrooms in natural scenes is a major technical difficulty for picking robots[3]. Under the constraints of objective factors, constructing an effective mushroom position and pose recognition model can provide an effective approach for accurate and low-damage mushroom picking.

The typical methods used in agricultural picking image recognition are conventional machine-learning and deep-learning methods[4-6]. The results obtained by these methods depend on the advantages and disadvantages of feature extraction. Their recognition effect is unstable and their degree of generalizability is low. Deep-learning methods include target detection and image segmentation. Target detection can detect a single mushroom body

Received date: 2024-04-09 **Accepted date:** 2024-05-17
Biographies: Jiaoling Wang, PhD, Associate Professor, research interest: edible fungi production and fruit and vegetable processing technology and equipment, Email: kclwjl@126.com; **Weidong Song**, Professor, research interest: edible fungi production and fruit and vegetable processing technology and equipment, Email: songwd@163.com; **Wengang Zheng**, PhD, Professor, research interest: environmental intelligent perception control and intelligent irrigation technology of water and fertilizer, Email: zhengwg@nercita.org.cn; **Qingchun Feng**, PhD, Professor, research interest: environmental intelligent perception control, water and fertilizer intelligent irrigation technology and agricultural robot, Email: fengqc@nercita.org.cn; **Mingfei Wang**, Associate Professor, research interest: environmental intelligent perception control and intelligent irrigation technology of water and fertilizer, Email: wangmf@nercita.org.cn.
***Corresponding author: Chunjiang Zhao**, PhD, Professor, CAE Academician, research interest: agricultural information technology. Beijing Research Center of Intelligent Equipment for Agriculture, Beijing 100097, China, Email: zhaocjnercita@sina.com.

in a complex environment. However, in complex scenes, such as those involving clutter occlusion and mushroom body adhesion, target detection can easily blur the target frame boundary or cause miss-detection. Image segmentation aims to divide the image into several specific regions with unique properties and propose the target of interest; this operation can effectively improve the recognition efficiency and accuracy. Segmentation was previously studied in citrus picking[7-9] and wheat pest recognition[10-12]. These methods only separate individual objects from the environment and cannot detect different objects of the same class, whereas, for mushroom picking, coarse localization of each mushroom is required. Previous studies have used detection algorithms that did not consider information at the pixel level of each mushroom, resulting in information loss when combined with depth maps.

The instance segmentation method not only enables the recognition and classification of individuals but also enables their frame-selection segmentation, which can effectively obtain individual information for aggregated overlapping objects. The training samples for image instance segmentation require the use of pixel-level mask information; hence, a large amount of trainable data is usually required for the full utilization of the feature extraction and image analysis capabilities of deep convolutional neural networks (CNNs). In general, this requirement increases the cost of manual labeling. Currently, the typical instance segmentation algorithms are Mask RCNN[13], Mask scoring RCNN[14], SOLOv2[15], and CNN-based methods. The emergence of the segment anything model (SAM[16]) has significantly improved segmentation capabilities. However, its large model parameters and computational demands prevent its usage in edge computing or embedded hardware. To address agricultural scenarios, this study focuses on lightweight networks that are easy to deploy. CNN-based methods can achieve lightweight operations. However, their extracted features do not fit the global correlation. Moreover, the existing transformer methods learn global features at the spatial scale; no global features have yet been learned in the channel dimension for segmentation tasks in agricultural scenes.

This study proposes a spatial-channel transformer network based on Mask-RCNN (SCT-Mask-RCNN). Mask-RCNN integrates a self-attention mechanism to effectively extract comprehensive similarities among image features across both channel and spatial dimensions[17,18]. Subsequently, it is used to maintain a lightweight structure and extract local features using a spatial pooling pyramid structure. The model achieves multi-scale local feature fusion, improves the detection accuracy, and reduces the number of operations required for the accurate segmentation and morphological recovery of overlapping mushroom bodies in large fields, thereby improving the picking success rate and efficiency of mushroom-picking robots.

The primary contributions of this study can be summarized as follows:

1) A new mushroom instance segmentation dataset is provided

2) A new model is constructed by combining Mask-RCNN with a spatial-channel attention module to improve the ability of the CNN-based model to learn global features

3) The experimental results reveal that the new model improves the performance compared to that of other models

2 Related work

2.1 Mushroom identification methods

Some mushroom identification methods target individual species. Chen et al.[19] designed a mushroom-picking robot with three degrees of freedom, specifically engineered for *Agaricus bisporus*. In a contrasting approach, Yang et al.[20] employed Mask-RCNN for the segmentation and localization of *A. bisporus*. Cong et al.[21] utilized YOLOv3 to design MYOLO, a specialized system for the detection and identification of mushrooms. Despite these advancements, the lack of generalization across various mushroom species and comprehensive case studies to validate their efficacy exists. Future studies should prioritize addressing these gaps, fostering a more universal and robust framework for mushroom identification technologies.

2.2 Instance segmentation algorithms

Instance segmentation, a key direction in machine vision[22-25], uses target detection to locate the box of each instance; subsequently, it segments each box to obtain the mask of instance. Another scheme involves classification on a pixel-by-pixel basis using semantic segmentation and instance identification by clustering. These approaches can be divided into the two-stage instance, one-stage instance, and query-based instance segmentation algorithms. Two-stage instance segmentation algorithms include Mask-RCNN[13], Cascade Mask-RCNN[26], and HTC[19]. One-stage instance segmentation methods include YOLACT[27], BlendMask[28], EmbedMask[29], and SOLOv2[30].

2.3 Transformers

Although current CNN-based models have satisfactorily performed in the field of agricultural image segmentation, they do not fulfill the requirements of global feature extraction; in addition, they are not easy to deploy on mobile terminals. In some scenarios, image segmentation remains challenging. Considering its architecture, CNN cannot acquire the feature information of long sequences satisfactorily. To solve this problem, some studies have added different pyramid modules[7,31] and self-attention mechanisms[32,33]. However, these methods still have limitations for long sequences of information. Recently, the transformer model[34] has yielded satisfactory results in the field of natural language processing[35]. Dosovitskiy et al.[36] applied the vision transformer (ViT) model to the visual segmentation of large datasets and achieved reasonable results. Liu et al.[37] proposed the Swin-transformer model. In contrast to the transformer model, the shift windows in the Swin-transformer model can significantly reduce the computation load and achieve more reasonable results on large datasets. Numerous studies have employed the transformer model[38,39] and Swin-transformer model[40-42] in the field of agriculture, and some researchers[43] applied a hybrid architecture model of CNN and transformer to terrace image segmentation; all approaches yielded reasonable performance.

3 Materials and methods

3.1 Datasets

Ninety-three images were collected in a laboratory setting using a 640×480 pixels depth camera. All images were manually annotated by experts. The dataset was split into a training subset containing 89 images and a test subset with 4 images. Considering the small number of samples collected, the rotation was applied, flipping, increasing brightness, reducing brightness, and mosaic data enhancement methods, as shown in Figure 1.

3.2 SCT-Mask-RCNN overall framework

Figure 2a shows the overall framework of SCT-Mask-RCNN, and the overall structure, with Mask-RCNN, aims to perform object detection, instance segmentation, and pixel-level mask prediction simultaneously. The proposed approach consists of the backbone network, region proposal network (RPN), ROI align, and two sub-

networks modules.

Figure 2a shows the overall framework of SCT-Mask-RCNN, and the overall structure, with Mask-RCNN, aims to perform object detection, instance segmentation, and pixel-level mask prediction simultaneously. The proposed approach consists of the backbone network, region proposal network (RPN), ROI align, and two sub-networks modules.

The most significant contribution of this study is the backbone network, which processes the input image and extracts layered features through a series of operations. Typical choices for previous backbone networks include ResNet and ResNeXt, which are pre-trained on large datasets to capture general visual representations. However, such CNN modules cannot learn global features; in this study, CNN is improved to a self-attention structure, utilizing a residual module that combines a channel self-attention block (ChannelAB) and a spatial self-attention block (SpatialAB) to enhance the network's global learning capability. At the end of backbone, the FPN module is embedded to improve the robustness of the network.

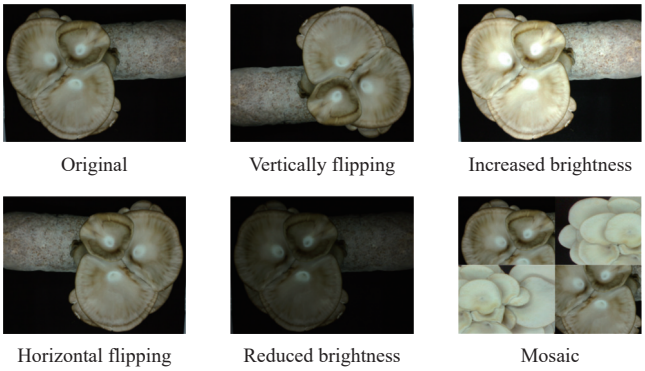


Figure 1 Data enhancement methods

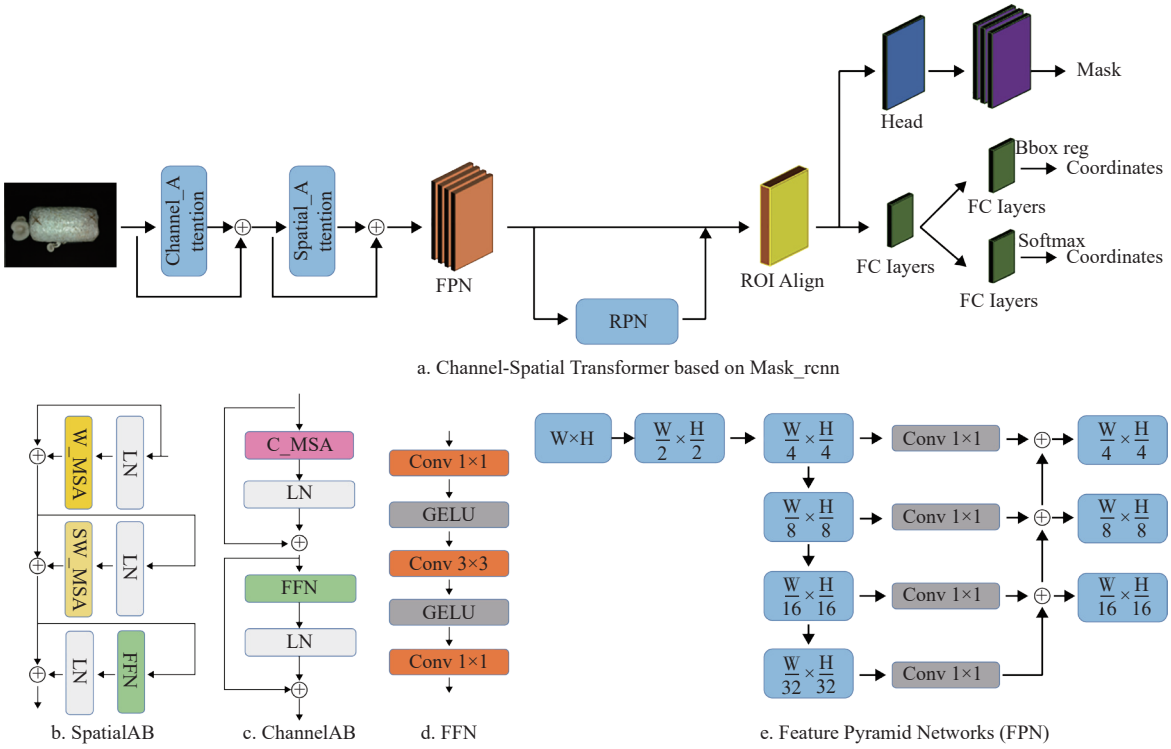


Figure 2 SCT-Mask-RCNN network framework diagram consisting of spatial self-attention block (SpatialAB), channel self-attention block (ChannelAB), feed-forward network (FFN), feature pyramid networks (FPN), and other major modules

RPN generates regional proposals for potential object instances. RPN presents candidate bounding boxes along with their objectivity scores to show the likelihood of containing objects, where the use of non-maximization suppression techniques can refine the proposals. As an example, Figure 3 shows the specific structure of the prediction layer (with a grid size of W×H). The prediction parameters of each prediction layer include the prediction frame center coordinates (*X*, *Y*), prediction frame length and width (H and W, respectively), prediction frame confidence level (C), score of fresh shiitake mushrooms in the prediction frame (Score), and number of predicted bounding boxes for which each grid is responsible (B).

ROI align and two sub-networks consist of two branches: the bounding box regression and classification sub-network and mask prediction sub-network. The bounding box regression and classification sub-network refines the bounding box of the RPN's output suggestions and classifies the objects in it. The mask prediction sub-network generates pixel-level segmentation masks for each classified object.

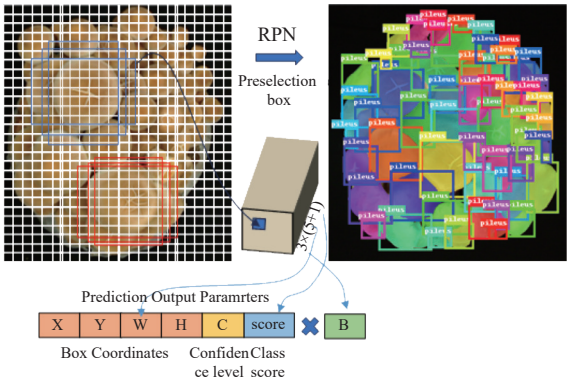
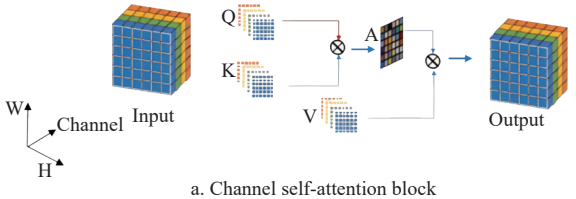


Figure 3 Principles of RPN

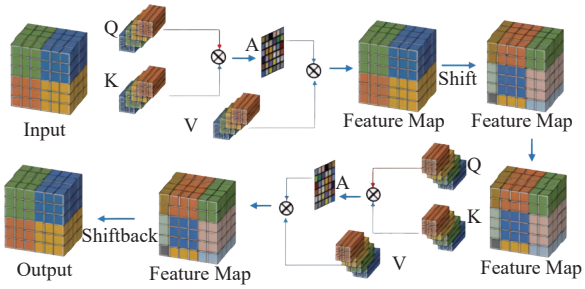
Therefore, this study focuses on the backbone network, where feature extraction determines the generation of pre-selected boxes in RPN, detection of the bounding box regression and classification sub-network and mask prediction sub-network. Segmentation. Moreover, the backbone network accounts for more than 90% of the overall network. A reasonable feature extractor can considerably improve the performance of the network.

3.3 Channel self-attention block

Figure 4a shows a channel self-attention block. The red, green, and blue (RGB) images have three channels (W×H), as shown in Figure 2a. The features in the WH dimension are associated with the spatial scene distribution, whereas those in the RGB channel dimension are associated with the spectral reflectance of the scene. Determining the global similarity of scenes based on the channel distribution is a cost-effective and high-yield approach. Because $W = H \gg C$, capturing spatial-wise interactions is less cost-effective than modelling channel-wise correlations. However, when the model reaches a certain scale, a single method cannot continue to mine image feature information.



a. Channel self-attention block



b. Spatial self-attention block

Figure 4 Channel self-attention block and spatial self-attention block in SCT-Mask-RCNN

Inspired by Cai et al.[17], the Channel AB is consistent with MSAB, as shown in Figure 2c. Channel AB treats each channel feature map as a token and calculates the self-attention along the channel dimension. The input $X_n \in R^{n_x \times n_y \times c}$ is reshaped into tokens $X \in \mathbb{R}^{n_x n_y \times c}$. Subsequently, X is linearly projected into *query*Q, *key*K, *value*$V \in \mathbb{R}^{n_x n_y \times c}$. In addition, $Q = XW^Q, K = XW^K$, $V = XW^V$, where W^Q, W^K, and $W^V \in \mathbb{R}^{c \times c}$. Subsequently, Q, K, and V are segmented into N heads along the spectral channel dimension as follows: $Q = [Q_1, \ldots, Q_N]$, $K = [K_1, \ldots, K_N]$, and $V = [V_1, \ldots, V_N]$. Therefore, each $head_j^{channel}$ and Channel AB can be expressed as

$$head_j^{channel} = SoftMax\left(\sigma_j Q_j K_j^T\right) V_j \tag{1}$$

$$ChannelAB(x) = concat_{j=1}^{N}\left(head_j\right) W + f(V) \tag{2}$$

where, K_j^T denotes the transposed matrix of $k_j, w \in \mathbb{R}^{c \times c}$, which are learnable parameters; and $f(\cdot)$ is the function that generates position embedding.

3.4 Spatial self-attention block

Figure 4b illustrates the channel self-attention block. A spatial self-attention block was introduced. The channel self-attention block primarily focuses on the global correlation of channels and determines the spatial feature distribution of RGB images. The feature map of channel self-attention is only 3×3; thus, it involves a small number of operations; however, it cannot extract detailed spatial features, resulting in the loss of high-frequency information. Our proposed spatial self-attention block (spatialB, as illustrated in Figure 2b) is based on a Swin-transformer model. The input $x^k \in \mathbb{R}^{n_x \times n_y \times c}$ is reshaped into tokens $x \in \mathbb{R}^{\frac{n_x}{n} \times \frac{n_y}{n} \times c}$, where s represents the window size of each window. Subsequently, x is linearly projected into query Q, key K, and value V, where W^Q, W^K, $W^V \in \mathbb{R}^{s \times s}$. The ability of the global attention map is enhanced by moving the window. The space attention operation can be expressed as

$$Attention = SoftMax\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{3}$$

where, Q, K, $V \in \mathbb{R}^{s^2 \times c}$ are the query, key, and value matrices, respectively; d is the *query/key* dimension; and S^2 is the number of patches in a window.

3.5 Loss

During model training, the predicted and true values have an uncertainty error. The loss function continuously reduces this error such that the predictions by model are as close as possible to the corresponding true values. The loss function of SCT-Mask-RCNN consists of three main parts: bounding box loss, mask loss, and classification loss, which is expressed as follows:

$$L = L_{cls} + L_{box} + L_{mask} \tag{4}$$

where, L_{mask} is the most important loss, whose design determines the quality of segmentation. For each ROI, the mask branch has outputs of $K \times m \times m$ dimensions which encodes K masks of size $m \times m$, and each ROI has K categories. Per-pixel sigmoid was used, and L_{msk} was defined as the average binary cross-entropy loss, which can be expressed as follows:

$$L_{mask} = \frac{1}{m^2}\sum_{i}^{k}\left(1^k\right)\sum_{1}^{m^2}\left[-y \times \log\left(sigmoid(x)\right) - (1-y) \times \log\left(1 - sigmoid(x)\right)\right] \tag{5}$$

where, 1^k denotes when the k^{th} channel corresponds to the true category of target, and 0 otherwise; y denotes the label value of the mask at the current position; x denotes the output value at the current position, and *sigmoid*(x) denotes the result of the output x transformed by the sigmoid function.

Moreover, L_{cls} is the classification loss, which can be expressed as

$$L_{cls}\left(p_i, p_i^*\right) = \frac{1}{N_{cls}}\sum_{i} -\log\left[p_i^* p_i + \left(1 - p_i^*\right)\left(1 - p_i\right)\right] \tag{6}$$

where, N_{cls} is the number of anchors and p_i is the probability that an anchor is predicted to be a target.

The regression loss box can be expressed as

$$L_{box} = \lambda \frac{1}{N_{reg}}\sum_{i} p_i^* R\left(t_i - t_i^*\right) \tag{7}$$

where, t_i is a vector denoting the offset of the anchor's predicted output in the RPN stage. Moreover, x and y, w, and h denote the center coordinates, width, and height of the anchor, respectively; t_i^* is a vector with the same dimensions as t_i, denoting the offset of the anchor's output in the RPN stage with respect to *gt*. $R(x)$ is a smooth $L1$ function, which is expressed as follows:

$$R(x) = \begin{cases} 0.5x^2 \times \frac{1}{\sigma^2}, & \text{if } |x| \le \frac{1}{\sigma^2} \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{8}$$

4 Experimental work

4.1 Experimental settings

All methods were trained from scratch and did not use pre-training weights nor were they fine-tuned based on other models to impartially evaluate the obtained results. For the other SOTA models, we used the MMSegmentation toolbox of MMLab for training.

4.1.1 Model evaluation

In this study, to verify the accuracy of the proposed method and other SOTA methods, mushroom precision (P), recall (R), average precision (AP), and mAP were used as evaluation indicators. The indicators are calculated as follows:

$$\mathrm{P}=\frac{TP}{TP+FP} \tag{9}$$

$$\mathrm{R}=\frac{TP}{TP+FN} \tag{10}$$

$$\mathrm{AP}=\int_0^1 P(R)\,\mathrm{d}R \tag{11}$$

where, TP is the number of fresh shiitake mushrooms detected correctly, FP is the number of fresh shiitake mushrooms detected incorrectly, and FN is the number of fresh shiitake mushrooms missed. For each category in the target detection, a P-R curve can be plotted based on accuracy and recall.

4.1.2 Evaluation metrics

The key evaluation metrics are bbox_Precision_mAP and segm_Precision_mAP, which are defined in an Intersection over Union (IOU) range of 0.5-0.95, and AP is calculated every 0.05×IOU and then averaged. Moreover, bbox_mAP_50 and seg_mAP_50 are the AP values when IOU = 0.5, bbox_mAP_75 and seg_mAP_75 are the AP values when IOU = 0.75, and bbox_mAP_75 and seg_mAP_75 are the AP values when IOU = 0.75. bbox_mAP_s/m/l and seg_mAP_s/m/l are the AP values of small-, medium-, and large-size objects, respectively, where small indicates that the area of the object is less than 32×32, medium indicates that the area is between 32×32 and 96×96, and large indicates that the area is larger than 96×96.

4.2 Implementation details

The proposed network was implemented using PyTorch, and the proposed SCT-Mask-RCNN was trained using a personal computer with the hardware and software specifications as presented in Table 1. We adopted an Stochastic Gradient Descent (SGD) optimizer (lr = 0.02, momentum = 0.9, and weight decay = 0.0001) for 100 epochs.

Table 1 Hardware and software configuration

Hardware or Software	Configuration
CPU	Intel i9-10700H
GPU	Nvida GeForce RTX 3090 24 G
Operating system	Ubuntu 20.04
SSD	1 T
Development environment	CUDA 11.4

5 Results and discussion

5.1 Experimental results

The comparison was made between bbox_mAP, seg_mAP, bbox_mAP_50, and seg_mAP_50, bbox_mAP_75, seg_mAP_75, bbox_mAP_s/m/l, and seg_mAP_s/m/l of our SCT-Mask-RCNN with several SOTA instance segmentation algorithms, including Mask_rcnn with resnet_50, Mask_rcnn with resnet_101, yolact, queryinst, Point_rend, and HTC. Table 2 lists the experimental results.

Table 2 Comparison of metrics outcomes obtained by the proposed method and SOTA methods

Tables metrics	Mask_rcnn_resnet_50[13]	Mask_rcnn_resnet_101[13]	Yolact[27]	Queryinst[44]	Point_rend[45]	HTC[19]	Mask_rcnn with spatial channel attention [ours]
Comparison of detection metrics outcomes obtained by the proposed method and SOTA methods							
bbox_Precision_mAP	0.374	0.588	0.242	0.250	0.443	0.470	0.638
bbox_Precision_mAP_50	0.705	0.964	0.765	0.421	0.905	0.897	0.951
bbox_Precision_mAP_75	0.240	0.551	0.046	0.333	0.419	0.403	0.631
bbox_Precision_mAP_s	0.071	0.460	0.230	0.010	0.320	0.204	0.462
bbox_Precision_mAP_m	0.179	0.643	0.147	0.022	0.429	0.396	0.618
bbox_Precision_mAP_l	0.463	0.715	0.343	0.276	0.589	0.587	0.692
bbox_Recall_mAP	0.497	0.677	0.320	0.367	0.546	0.540	0.673
bbox_Recall_mAP_s	0.200	0.552	0.372	0.056	0.476	0.272	0.544
bbox_Recall_mAP_m	0.310	0.678	0.268	0.149	0.527	0.498	0.673
bbox_Recall_mAP_l	0.630	0.790	0.397	0.460	0.653	0.647	0.750
Comparison of segmentation metrics outcomes obtained by the proposed method and SOTA methods							
segm_Precision_mAP	0.451	0.731	0.356	0.205	0.568	0.520	0.750
segm_Precision_mAP_50	0.691	0.965	0.526	0.407	0.875	0.864	0.957
segm_Precision_mAP_75	0.528	0.868	0.503	0.087	0.614	0.591	0.853
segm_Precision_mAP_s	0.067	0.512	0.008	0.006	0.282	0.155	0.512
segm_Precision_mAP_m	0.190	0.689	0.014	0.020	0.336	0.336	0.705
segm_Precision_mAP_l	0.501	0.840	0.463	0.225	0.651	0.620	0.760
segm_Recall_mAP	0.549	0.783	0.376	0.352	0.653	0.558	0.785
segm_Recall_mAP_s	0.236	0.616	0.036	0.060	0.420	0.236	0.604
segm_Recall_mAP_m	0.316	0.719	0.041	0.113	0.443	0.383	0.740
segm_Recall_mAP_l	0.590	0.873	0.480	0.400	0.683	0.633	0.803

Figure 5 shows a comparison between the segmentation results obtained using the SCT-Mask-RCNN and other SOTA methods. The results in Table 2 show that the proposed SCT method outperforms Mask_rcnn with resnet_50, Mask_rcnn with

resnet_101, yolact, queryinst, Point_rend, and HTC, improved by 26.4%, 5.0%, 39.6%, 38.8%, 19.5%, 16.8%, and 29.9%, 1.9%, 39.4%, 54.5%, 18.2%, and 23.0%, respectively.

The visual results show that the proposed method is significantly more advantageous than other methods for segmentations of small- and medium-size objects.

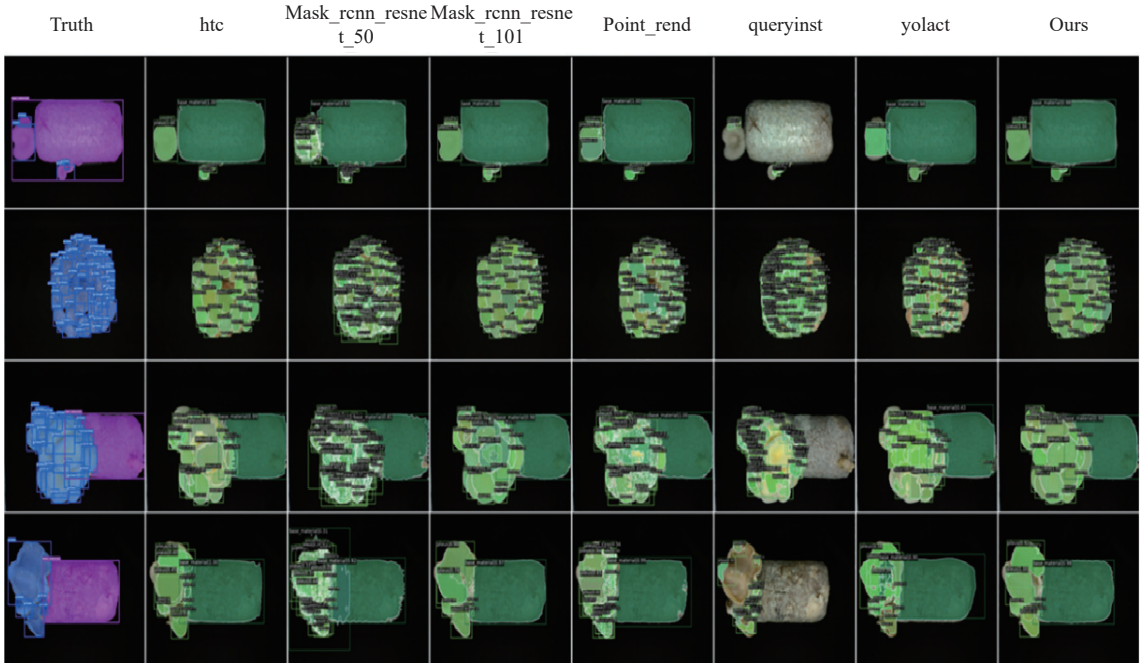


Figure 5 Visual comparisons of the proposed method with SOTA methods

5.2 Analysis

Unlike fruits, such as apples, strawberries, and kiwi fruit, mushrooms differ not only by color, size, and shape but also by growing periods for the same category. The target recognition frame and segmentation effect of the proposed method with those of the other methods are compared to further clarify the effectiveness of the proposed method; Figure 6 shows the results. The proposed method outperforms the other methods in terms of mAP metrics of Bbox (Figure 6a), mAP precision of Bbox (Figure 6c), and mAP metrics of segmentation (Figures 6d). Figure 6b shows the Recall of large Bbox metrics. Figures 6e and 6f show that the SCT-Mask-RCNN network outperforms other methods for medium-size objects, particularly mask_rcnn_resnet_101; however, the proposed method has less depth, and thus it is slightly inferior to the deeper network mask_rcnn_resnet_101 for large-size objects.

The number of parameters of the proposed and other SOTA methods was obtained to further evaluate the effectiveness of the proposed method, as presented in Table 3. The results show that SCT-Mask-RCNN outperforms the SOTA method in both Bbox_Precision_mAP and segm_Precision_mAP, and uses approximately 76.6% less parameters than Mask_rcnn_resnet_101. In addition, the Bbox_Precision_mAP and segm_Precision_mAP metrics of the proposed method compared to those of the Point_rend method were improved by 19.5% and 18.2%, respectively, even though SCT-Mask-RCNN used 2.98 M more parameters.

Table 3 Parameters of SOTA methods and the proposed method

Methods	Mask_rcnn_resnet_50[13]	Mask_rcnn_resnet_101[13]	Yolact[27]	Queryinst[44]	Point_rend[45]	HTC[19]	Mask_rcnn with spatial channel attention (This study)
Params	25.56 M	44.98 M	49.12 M	96.97 M	34.63 M	45.72 M	37.61 M

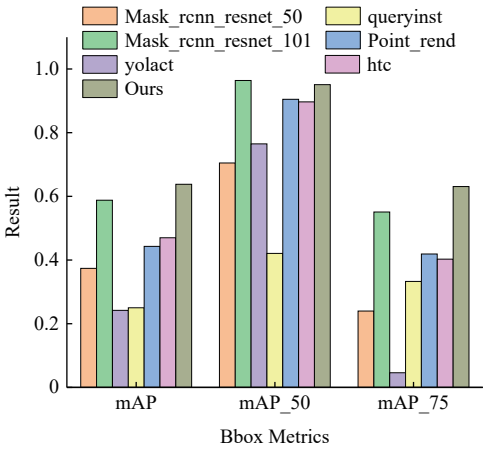
5.3 Ablation experiments

Ablation experiments are usually performed on complex neural networks to explore the effects of network-specific substructures or training strategies and parameters on the model generation; they provide key guidelines for the structural design of neural networks[46]. To evaluate the effectiveness and feasibility of the proposed SCT-Mask-RCNN lightweight model, the performance of the ChannelAB, SpatialAB, and FFN modules were verified experimentally. As the SCT-Mask-RCNN model is an improved version of Mask_rcnn, the head of the overall network remains intact; hence, Mask_rcnn_resnet_50 was used for ablation experiment comparisons.

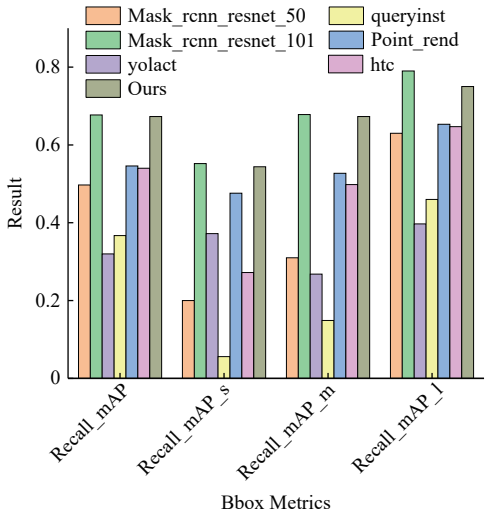
Table 4 shows the results of the ablation experiments performed using the test set on networks with different modules removed. The results show that ChannelAB, SpatialAB, and FFN positively contribute to enhance the network. The module with SpatialAB exhibits the highest contribution, where the segm_Precision_mAP metric improves by 0.288 over Mask_rcnn_resnet_50 and uses 11.57 M more parameters. ChannelAB boosts the network better, with an improvement of 0.144 in the segm_Precision_mAP metric and using only 4.02 more parameters. Finally, the FFN module enhances the network, where segm_Precision_mAP is improved by approximately 0.11.

Table 4 Ablation experiments of different modules in Mask_rcnn with spatial channel attention

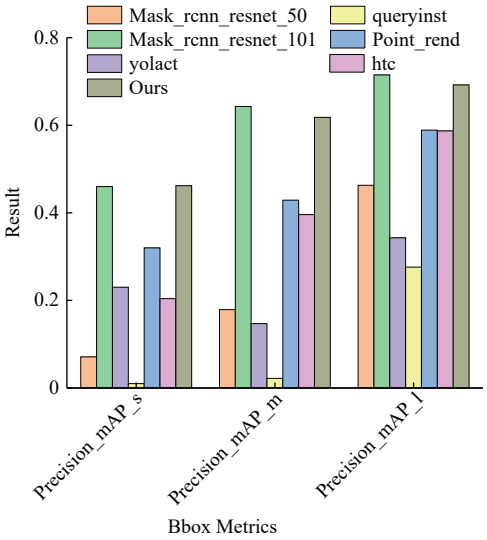
Base-line	With ChannelAB	With SpatialAB	With FFN	segm_Precision_mAP	Total parameters
Mask_rcnn with channel attention	√	√	√	0.750	37.61 M
	√	√	×	0.739	37.13 M
	√	×	×	0.595	29.58 M
Mask_rcnn_resnet_50	×	×	×	0.451	25.56 M



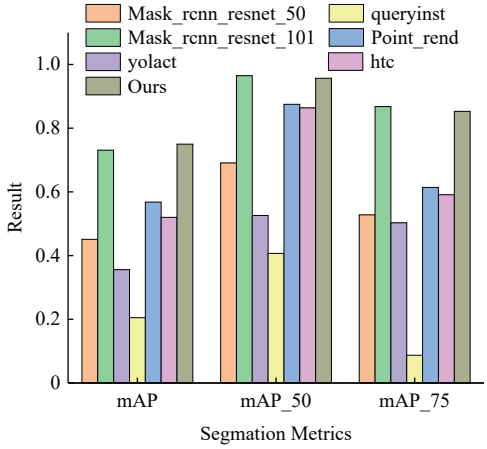
a. mAP, mAP_50 and mAP_75 on Bbox detection



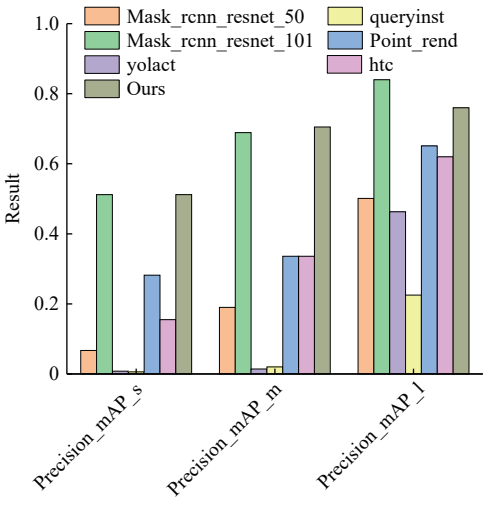
b. Recall_mAP, Recall_mAP_s, Recall_mAP_m and Recall_mAP_l on Bbox detection



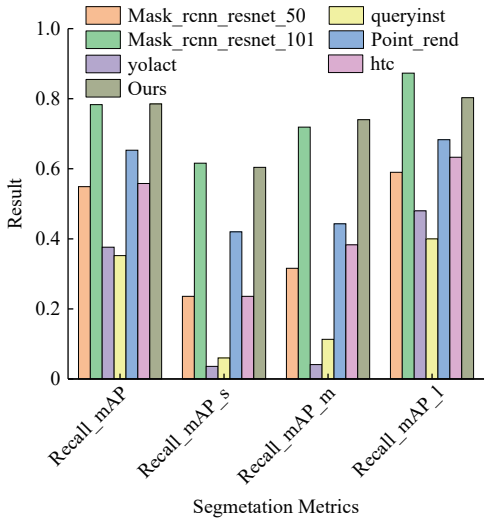
c. Precision_mAP_s, Precision_mAP_m and Precision_mAP_l on Bbox detection



d. mAP, mAP_50 and mAP_75 on segmentation



e. Precision_mAP_s, Precision_mAP_m and Precision_mAP_l on segmentation



f. Recall_mAP, Recall_mAP_s, Recall_mAP_m and Recall_mAP_l on segmentation

Figure 6 Detection results using the proposed and other SOTA methods considering different metrics

6 Conclusions

This study investigated instance segmentation techniques for coarse localization of edible mushrooms and provided a new mushroom instance segmentation dataset. To improve the accuracy of instance segmentation on a mushroom segmentation dataset, a

new spatial-channel transformer network based on a Mask-RCNN network (SCT-Mask-RCNN) was proposed. SCT-Mask-RCNN was constructed by combining Mask-RCNN with a spatial-channel attention module to improve the ability of the CNN-based model to learn global features. Experiments on the new dataset revealed that the proposed method exhibited a higher performance than that of other methods. The SCT-Mask-RCNN method yielded a segmentation accuracy of 0.750 on segm_Precision_mAP and detection accuracy of 0.638 on Bbox_Precision_mAP. Thus, the proposed method yielded improvements exceeding 2% and 5% than those achieved using other methods, respectively.

The SCT-Mask-RCNN network can efficiently coarsely localize mushrooms, improve the accuracy of previous detection algorithms, and increase the efficiency of mushroom-picking robots. In future studies, the automation of mushroom picking will be attempted by deploying a SCT-Mask-RCNN network using a two-dimensional calibration algorithm.

Acknowledgements

This work was financially supported by China Agriculture Research System of MOF and MARA (CARS-20); Zhejiang Provincial Key Laboratory of Agricultural Intelligent Equipment and Robotics Open Fund (2023ZJZD2301); Chinese Academy of Agricultural Science and Technology Innovation Project “Fruit And Vegetable Production And Processing Technical Equipment Team” (2024);Beijing Nova Program(20220484023).

[References]

[1] Wang M, Zhao R. A review on nutritional advantages of edible mushrooms and its industrialization development situation in protein meat analogues. Journal of Future Foods, 2023; 3(1): 1–7.

[2] Li C, Xu S. Edible mushroom industry in China: Current state and perspectives. Applied Microbiology and Biotechnology, 2022; 106(11): 3949–3955.

[3] Retsinas G, Efthymiou N, Anagnostopoulou D, Maragos P. Mushroom detection and three dimensional pose estimation from multi-view point clouds. Sensors, 2023; 23(7): 3576.

[4] Hua X, Li H, Zeng J, Han C, Chen T, Tang L, et al. A review of target recognition technology for fruit picking robots: from digital image processing to deep learning. Applied Sciences, 2023; 13(7): 4160.

[5] Qi X, Dong J, Lan Y, Zhu H. Method for identifying litchi picking position based on YOLOv5 and PSPNet. Remote Sensing, 2022; 14(9): 2004.

[6] Dean Z, Liu X Y, Chen Y, Jin J, Jia W K, Hu C L. Image recognition at night for apple picking robot. Transactions of the CSAM, 2015; 46(3): 15–22.

[7] Xu C, Lu Y, Jiang H, Liu S, Ma Y, Zhao T. Counting crowded soybean pods based on deformable attention recursive feature pyramid. Agronomy, 2023; 13(6): 1507.

[8] Yang C H, Xiong L Y, Wang Z, Wang Y, Shi G, Kuremot T, et al. Integrated detection of citrus fruits and branches using a convolutional neural network. Comput Electron in Agric, 2020; 174: 105469.

[9] Chen W, Lu S, Liu B, Li G, Qian T. Detecting citrus in orchard environment by using improved YOLOv4. Scientific Programming. 2020; 2020: 1–3.

[10] Chen P, Li W, Yao S, Ma C, Zhang J, Wang B, et al. Recognition and counting of wheat mites in wheat fields by a three-step deep learning method. Neurocomputing, 2021; 437: 21–30.

[11] Li R, Wang R J, Zhang J, Xie C J, Liu L, Wang F Y, et al. An effective data augmentation strategy for CNN-based pest localization and recognition in the field. IEEE Access, 2019; 7: 160274–160283.

[12] Liu T, Chen W, Wu W, Sun C M, Guo W S, Zhu X K. Detection of aphids in wheat fields using a computer vision technique. Biosystems Engineering, 2016; 141: 82–93.

[13] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, 2017; pp.2961-2969.

[14] Huang Z J, Huang L C, Gong Y C, Huang C, Wang X G. Mask scoring R-CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019; pp.6409-6418.

[15] Sun C Z, Hu X M, Yu T. Structural design of agaricus bisporus picking robot based on cartesian coordinate system. Electrical Engineering and Computer Science (EECS), 2019; 2: 103–106.

[16] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W Y, Dollár P. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023; pp.4015–4026.

[17] Cai Z Y, Jian Y, Zhang Z Y, Jin C Q, Da F P. SST-ReversibleNet: Reversible-prior-based spectral-spatial transformer for efficient hyperspectral image reconstruction. Arxiv preprint, 2023; arxiv: 2305.04054.

[18] Cai Z Y, Li C Y, Yu Y, Jin C Q, Da F P. Momentum accelerated unfolding network with spectral-spatial prior for computational spectral imaging. Applied Soft Computing, 2024; Feb 21: 111420.

[19] Chen K, Pang J M, Wang J Q, Xiong Y, Li X X, Sun S Y, et al. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019; pp.4974–4983.

[20] Yang S Z, Huang J, Yu X Y, Yu T. Research on a segmentation and location algorithm based on mask RCNN for agaricus bisporus. In 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), IEEE, 2022; pp.717–721.

[21] Cong P C, Feng H, Lv K F, Zhou J C, Li S D. MYOLO: a lightweight fresh shiitake mushroom detection model based on YOLOv3. Agriculture, 2023; 13(2): 392.

[22] Hafiz A M, Bhat G M. A survey on instance segmentation: state of the art. International Journal of Multimedia Information Retrieval, 2020; 9(3): 171–89.

[23] Romera-Paredes B, Torr P H. Recurrent instance segmentation. In Proceedings of 14th European Conference on Computer Vision–ECCV 2016, Amsterdam, The Netherlands, 2016; pp.312–329.

[24] Arnab A, Torr PH. Pixelwise instance segmentation with a dynamically instantiated network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; pp.441–450.

[25] Lee Y, Park J. Centermask: Real-time anchor-free instance segmentation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp.13906–13915.

[26] Cai Z W, Vasconcelos N. Cascade R-CNN: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019; 43(5): 1483–1498.

[27] Bolya D, Zhou C, Xiao F, Lee Y J. Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019; pp.9157–9166.

[28] Chen H, Sun K Y, Tian Z, Shen C H, Huang Y M, Yan Y L. Blendmask: Top-down meets bottom-up for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp.8573–8581.

[29] Ying H, Huang Z, Liu S, Shao T J, Zhou K. Embedmask: Embedding coupling for one-stage instance segmentation. Arxiv preprint, 2019; arxiv: 1912.01954.

[30] Wang X L, Zhang R F, Kong T, Li L, Shen C H. Solov2: Dynamic and fast instance segmentation. Advances in Neural information Processing Systems, 2020; 33: 17721–17732.

[31] Shojaiee F, Baleghi Y. EFASPP U-Net for semantic segmentation of night traffic scenes using fusion of visible and thermal images. Engineering Applications of Artificial Intelligence, 2023; 117: 105627.

[32] Kaur A, Goyal P, Rajhans R, Agarwal L, Goyal N. Fusion of multivariate time series meteorological and static soil data for multistage crop yield prediction using multi-head self-attention network. Expert Systems with Applications, 2023; 226: 120098.

[33] Yang Q L, Ye Y, Gu L C, Wu Y T. MSFCA-net: A multi-scale feature convolutional attention network for segmenting crops and weeds in the field. Agriculture, 2023; 13(6): 1176.

[34] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017; 30: 1–11.

[35] Gillioz A, Casas J, Mugellini E, Abou Khaled O. Overview of the Transformer-based Models for NLP Tasks. In 15th Conference on

Computer Science and Information Systems (FedCSIS), IEEE, 2020; pp.179–183.

[36] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv preprint arxiv: 2010.11929. 2020 Oct 22.

[37] Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp.10012–10022.

[38] Bao W X, Xie W J, Hu G S, Yang X J, Su B B. Wheat ear counting method in UAV images based on TPH-YOLO. Transactions of the CSAE, 2023; 39(1): 155–161. (in Chinese)

[39] Xu Y L, Kong S L, Chen Q Y, Gao Z Y, Li C X. Model for identifying strong generalization apple leaf disease using transformer. Transactions of the CSAE, 2022; 38(16): 198–206. (in Chinese)

[40] Wang C, Wu X H, Zhang Y Q, Wang W J. Recognizing weeds in maize fields using shifted window Transformer network. Transactions of the CSAE, 2022; 38(15): 133–42. (in Chinese)

[41] Fu L L, Huang H, Wang H, Huang S C, Chen D. Classification of maize growth stages using the Swin transformer model. Transactions of the CSAE, 2022; 38(14): 191–200.

[42] Zhu D L, Yu M S, Liang M F. Real-time instance segmentation of maize ears using SwinT-YOLACT. Transactions of the CSAE, 2023; 39(14): 164–172. (in Chinese)

[43] Liu X, Yi S, Li L, Cheng X H, Wang C. Semantic segmentation of terrace image regions based on lightweight CNN-transformer hybrid networks. Transactions of the CSAE, 2023; 39(13): 171–181. (in Chinese)

[44] Fang Y X, Yang S S, Wang X G, Li Y, Fang C, Shan Y, et al. Instances as queries. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp.6910–6919.

[45] Kirillov A, Wu Y, He K, Girshick R. Pointrend: Image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp.9799–9808.

[46] Cai Z Y, Jin C, Da F. DMDC: Dynamic-mask-based dual camera design for snapshot Hyperspectral Imaging. arxiv preprint, 2023; arxiv: 2308.01541.