IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

KM-Mask RCNN: A lightweight instance segmentation algorithm for strawberries with multiple growth cycles

Peichao Cong[1], Yutao Xu[1], Tianheng Li[1], Shanda Li[1], Hao Feng[1], and Xin Zhang[1]

[1]School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China; cpclzx2022@gxust.edu.cn (P.C.); 221076898@stdmail.gxust.edu.cn (T.L.); 221068137@stdmail.gxust.edu.cn (S.L.); 221068136@stdmail.gxust.edu.cn (H.F.); zhang-xin6688@gxust.edu.cn(X.Z.).
Correspondence:

Corresponding author: Yutao Xu (e-mail: 221076938@stdmail.gxust.edu.cn).

This research was funded by the Central Government Guides Local Science and Technology Development Foundation Projects (grant no.ZY19183003), Guangxi Key Research and Development Project (grant no.AB20058001).”

ABSTRACT Accurate and efficient detection of multi-growth-cycle strawberry fruits can improve automated harvesting. However, the small size and unbalanced distribution of strawberry fruits make accurate identification of multi-growth-cycle strawberries difficult using the existing detection models. Herein, local enhancement technology is adopted for preprocessing when compiling the dataset to ensure the balance of the number of samples in each category to solve the above problems. Second, a new instance segmentation algorithm called KM-Mask RCNN is developed, which optimally adjusts the size of the anchor frame and the anchor ratio based on the K-Means clustering algorithm to improve the recognition accuracy of the algorithm on small targets and uses MobileNet V3 to replace the Resnet50 structure in the Mask RCNN backbone network to reduce the complexity of the algorithm and realize lightweight operation. Finally, the experimental results reveal five strawberry growth stages in the homemade dataset(based on StrawDI_Db1 database): ‘Green ripe stage’, ‘White ripe stage’, ‘Turning stage’, ‘Mature’, and ‘Deformed’, for which the KM-Mask RCNN yields mAPs of 91.19%, 88.09%, 93.70%, 93.19%, and 87.13%, respectively. The P, R, and F1-score values of this algorithm are 93.9%, 94.2%, and 94.05%, respectively. Additionally, the number of parameters, FLOPs, and fps of this algorithm are 27M, 12G, and 22.32, respectively, satisfying the real-time requirements for strawberry detection. The findings provide important theoretical support for the automated harvesting of strawberries with multiple growth cycles.

INDEX TERMS K-means clustering algorithm, lightweighting, mask RCNN, strawberry

I. INTRODUCTION

The strawberry, known as the ‘Queen of Fruits’, is a berry fruit consumed by people every day and is loved for its delicious taste and rich nutritional value. Consequently, the planting area and production of strawberries are increasing worldwide. Strawberry fruits are non-uniform (individuals have deformities) and have small, fragile, aggregated achenes [1]. Moreover, strawberries on the same plant have different growth stages (green ripening stage, white ripening stage, turning stage, and mature stage). Therefore, strawberry fruit picking needs to be conducted in batches. The traditional method of hand picking is difficult and inefficient, resulting in high waste, imposing a significant operational burden on strawberry growers.

In recent years, with the rapid development and application of artificial intelligence technology, new methods (picking robots) that incorporate visual detection technology to realize accurate picking of strawberries have emerged. Although picking robots are inefficient compared with the traditional methods, they are based on the premise of accurate perception of the operational target, which improves the accuracy of picking, reduces the difficulty of picking, and effectively ensures the quality and yield of the crop. The primary problem with strawberry fruit and vegetable picking robots based on visual detection technology is the realization of accurate

recognition of the operation target [2]. To solve this problem, a considerable amount of research work has been conducted, and the methods employed can be roughly divided into two categories: traditional methods and deep learning methods.

Traditional visual detection methods include Support Vector Machines (SVM), K Nearest Neighbor Algorithm (KNN), algorithms based on edge detection, and algorithms based on texture features. For example, Vitzrabin et al. [3] proposed an adaptive thresholding algorithm combined with sensor fusion, which can detect bell peppers with a high detection rate in environments with large variations in light. Azarmdel et al. [4] used feature dimensionality reduction to extract the geometric properties, color and texture features of segmented mulberries and applied an artificial neural network(ANN) and SVM to classify of mulberry fruits. Wang et al. [5] proposed a geometric center-based matching method to detect lychee fruits and then utilized pixel thresholding method to classify them. Arefi et al.[6] performed thresholding analysis to extract the combined features from the RGB, HIS, and YIQ spaces to locate ripe tomatoes, but the ability to detect small fruits was poor. Although traditional visual detection methods have many practical applications and have achieved certain results, they still have several insurmountable drawbacks: (1) the feature information required by such methods relies on manual design, which makes automatic extraction and discrimination difficult and relies more on scene information, and the algorithm has low recognition accuracy; (2) when conditions such as the illumination and shooting angle change, the image feature information of the target is affected as well, a large amount of manual analysis is needed, and the generalization ability of this type of method is poor. Therefore, the traditional visual detection methods have difficulty meeting the needs of real-time picking of strawberries with multiple growth cycles.

In recent years, deep learning theory has undergone rapid development [7]. Visual detection methods based on deep learning theory have been increasingly applied in the field of fruit picking [8]. Compared to traditional visual detection algorithms, these approaches have obvious advantages in terms of accuracy and generalization. Detection algorithms based on deep learning theory in the field of fruit detection are listed in Table 1, which are primarily divided into two-stage and one-stage object detection algorithms. For example, Yu et al. [9] proposed an instance segmentation algorithm based on Mask RCNN, which uses a residual network and feature pyramid network to fuse multiscale feature maps to segment instances and localize strawberries visually, resulting in an average accuracy of 95.78%. Peng et al. [10] proposed an instance segmentation method, ResDense-Focal-DeepLabV3+, model based on DeepLabV3 for accurate segmentation of lychee fruits in an orchard environment, achieving an mIoU of 79.7%, which is 0.9%, 1.8%, and 12.9% higher than those obtained using ResNet-CE, DenseNet, and Xception, respectively. Chu et al. [11] developed a detection model (Suppression Mask RCNN) for apple detection, adding suppression branches to the original network structure to suppress the non-apple features generated by the original network and achieving good detection results. Xu et al. [12] improved Mask RCNN by introducing a multi-class prediction sub-network to decouple the pixel-level category prediction of fruits and stems; simultaneously, they used multi-tasking loss balancing and adaptive feature pooling to overcome the limitation caused by the differences in fruit stem size. Liu et al. [13] added a depth filter to the Mask RCNN model to help the normalized exponential function (SoftMax) for anchor point classification to improve the detection accuracy of green asparagus during the autonomous harvesting process. Cong et al. [14] proposed an instance segmentation algorithm for bell pepper detection, which integrates the Swin Transformer attention mechanism into the backbone network of Mask RCNN to enhance the feature extraction capability of the algorithm; additionally, UNet3+ is utilized to improve the mask head and its segmentation quality.

The one-stage detection algorithms, represented by the YOLO series, have also been widely used [15]. Li et al. [16] proposed the YOLOv3-tiny-Litchi model based on YOLO v3 and a 3D clustering algorithm (K-Means) that detects lychee fruits and divides their harvesting area in 3D, obtaining 87.43% accuracy and 91.15% average harvesting rate in field trials. Jiang et al. [17] proposed a young apple fruit detection algorithm based on the improved YOLOv4, which employs a nonlocal attentional matching module (NLAM) and a convolutional attentional module (CBAM) to enhance the image deep feature information extraction and achieved 85.8% detection accuracy. Chen et al. [18] developed a lightweight multi-class occlusion target detection algorithm (YOLO-COF) for oleander fruit detection. The K-Means++ clustering algorithm was introduced under the YOLOv5s framework to filter the target dataset automatically and was combined with the coordinate attention mechanism to improve the feature extraction capability for occluded targets, yielding a detection accuracy of 94.1%. Li et al. [19] proposed a multi-task perceptual network (MTA-YOLACT) to build a classification and regression tree (CART) model to improve the generalization of the model and obtained a high average detection rate of 98.9% for the homemade maidenhair fruit image dataset. Wang et al. [20] proposed the 'DSE-YOLO' model for multi-stage strawberry fruit detection. The model focuses on small fruits and integrates YOLO, DSE module, exponentially enhanced binary cross entropy (EBCE) and doubly enhanced mean square error (DEMSE) loss functions to solve the problem

of resolving the foreground-prospect class imbalance. Zhang et al. [21] tuned the parameters of Yolov5s and reduced the depth and width of the network to make the model lighter. Grape fruit clusters were detected quickly and accurately, and a mAP of 99.40% and F1-score of 99.40% were achieved.

The above visual detection methods based on deep learning theory have further improved in terms of accuracy compared with traditional visual detection methods. In addition, the feature extraction ability is improved significantly, and the generalization ability of algorithms is more powerful. However, most of the existing models are only applicable to medium- and large-sized ripe fruit targets. When the detection targets are small, unevenly distributed, and in different growth cycles, it is difficult to ensure the accuracy of detection, and leakage and misdetection can easily occur. Additionally, the above algorithms increase the network depth by adding different modules to improve the learning ability of the algorithms, which tends to increase the number of calculations, increasing the number of model parameters, which in turn consumes considerably more computational resources and time, directly affecting the algorithm detection speed and making it difficult to fulfill the actual picking needs of strawberry fruits in terms of economy and effectiveness. Therefore, the development of a visual detection algorithm for small, multi-cycle strawberries is a popular topic in the field of fruit and vegetable picking robots.

TABLE I
RESEARCH ON VISUAL DETECTION METHODS BASED ON DEEP LEARNING THEORY

Author	Proposed algorithm	Backbone	Object	mAP (%)
Yu et al. [9]	New Mask RCNN	Resnet 50	Strawberry	95.78
Peng et al. [10]	ResDense-Focal-DeepLabV3+	Resnet 50	Litchi	79.70
Chu et al. [11]	Suppression Mask RCNN	Resnet 101	Apple	93.90
Xu et al. [12]	Mask RCNN with attention	Resnet 50	Cherry tomato	93.76
Liu et al. [13]	Mask RCNN with SoftMax	Resnet 50	Asparagus	99.3
Cong et al. [14]	Mask RCNN with Swin transformer	Swin transformer	Sweet pepper	98.1
Li et al. [16]	YOLOv3-tiny-Litch	Darknet-53	Litchi	87.43
Jiang et al. [17]	YOLOv 4-NLAM -CBAM	Darknet-53	Apple	85.80
Chen et al. [18]	YOLOv5 With K-means++ and Attention	Darknet-53	Camellia oleifera fruit	94.10
Li et al. [19]	MTA-YOLACT	Resnet101	Cherry tomato	98.90
Wang et al. [20]	DSE-YOLO	Darknet-53	Strawberry	86.58
Zhang et al. [21]	Improved Yolov5s	Darknet-53	Grape	99.4

The accuracy of the above algorithms for small target detection cannot be guaranteed. Moreover, the growth cycle of the detection target is relatively single, and the economy and effectiveness of the algorithms are poor. To address these problems, this study considers Mask RCNN as a basic model and a lightweight instance segmentation algorithm called KM-Mask RCNN is developed based on K-Means clustering algorithm and MobileNetV3 to conduct multi-growth-cycle strawberry fruit image detection.The main contributions are as follows:

(1) A multi-growth cycle strawberry fruit dataset is constructed by local data augmentation technique. For the collected dataset of deformed strawberry fruit images, scale, rotation, and other methods were used for combined data enhancement, and the remaining images were expanded using the pretzel noise method to ensure a relative balance of the number of samples in each category. This provides a high quality dataset for the research of multi-growth cycle strawberry segmentation

(2) Considering the characteristics of the actual situation, such as the small size of strawberry fruits and large individual differences in different growth cycles, the K-Means clustering algorithm is introduced to optimize and adjust the size of the anchor frame and the anchoring ratio, which realizes the customization of the anchor frames settings, so that the generated candidate frames are more accurate in enhancing the accuracy of the algorithm for the segmentation of small targets in the samples.

(3) The original backbone network Resnet 50 of Mask RCNN, which is based on the residual structure, is replaced with the more lightweight MobileNet V3, making the network model more lightweight and facilitating the practical application of mobile picking robots with guaranteed detection accuracy.

The research framework of this study is shown in Fig. 1 and involves three main parts. In Section 2, dataset acquisition and preprocessing are presented. Section 3 is methods, this part provides a detailed description of the proposed model and algorithm. In Sectition 4, data enhancement scheme comparison experiments, ablation experiments, and mainstream instance segmentation algorithm comparison experimental results are presented to verify the reasonableness of the work in this study, and a performance evaluation is conducted based on the relevant indices. Finally, Section 5 summarizes the essential features and areas for further improvement of the research.

II. Data acquisition and preprocessing

A. DATA ACQUISITION

1) IMAGE ACQUISITION

The dataset used in this study comprises two main parts. The first one was the StrawDI dataset [22], which contains 8000 strawberry images taken from approximately 150 hectares of plantations in the province of Huelva, Spain. The experimental plantations were not altered, and images were

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3485490

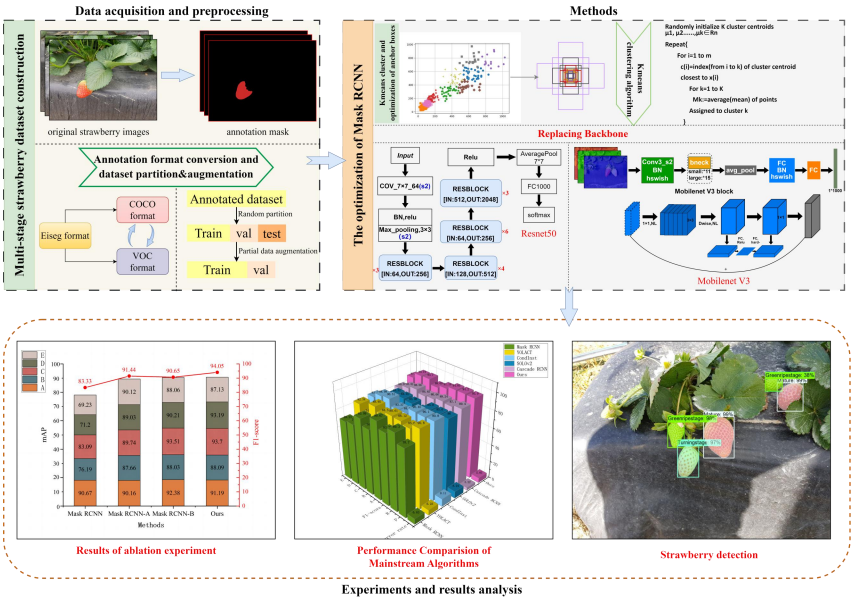

FIGURE 1. Schematic of the steps performed in this study.

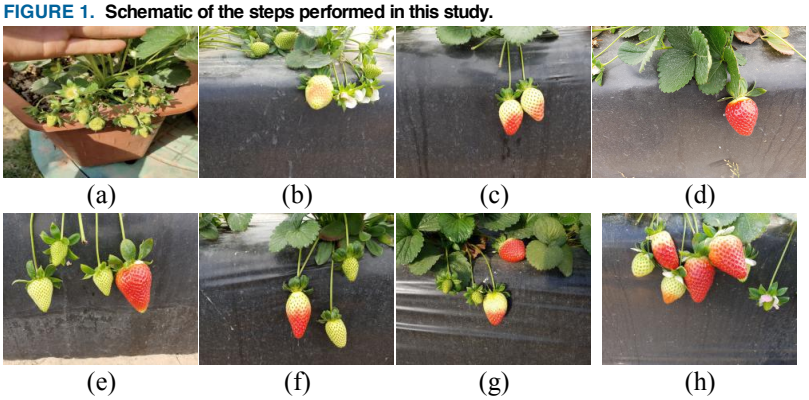

FIGURE 2. Strawberry fruits in different growth stages. (a) Green ripening; (b) White ripening; (c) Turning stage; (d) Mature; (e) Green ripening+Mature; (f) Turning+green ripening; (g) Mature+Turning; (h) All stages.

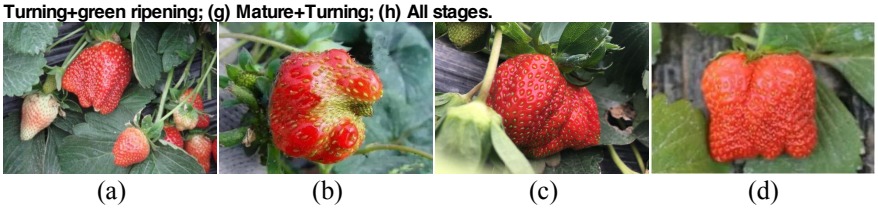

FIGURE 3. Deformed strawberries. (a), (b), (c), (d) represents different types of deformed strawberries.

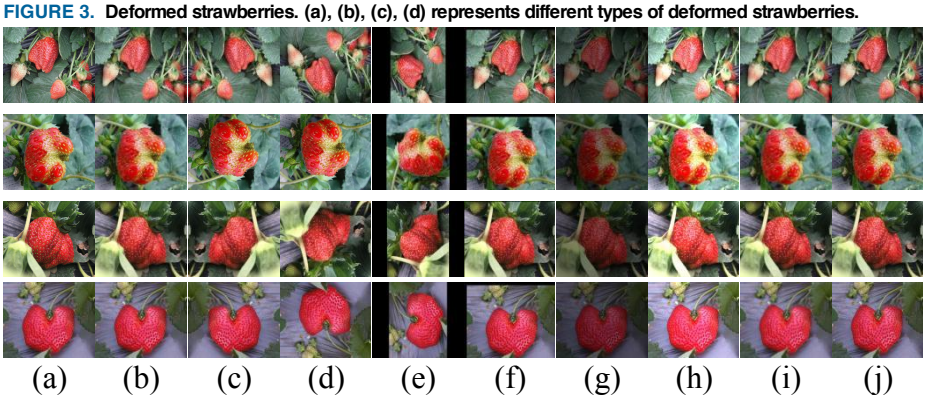

FIGURE 4. Effects of different data enhancements. (a) initial image; (b) zoomed in or zoomed out image; (c) horizontally flipped image; (d) vertically flipped image; (e) image rotated by 90°; (f) depicts image panning; (g) darkened image; (h) lightened image; (i) depicts the image with Gaussian noise added; (j) image with pretzel noise added.

captured from real production scenarios during the complete harvesting campaign (from mid-December 2018 to early May 2019). The StrawDI dataset contains images of strawberry fruits in different growth cycles, as shown in Fig. 2, which consist of four main periods: green ripening, white ripening, color change, and ripening. Strawberries are inevitably deformed during their growth cycles. Although StrawDI is rich in images of strawberries during different growth cycles, it lacks images of deformed strawberries. To solve this problem, deformed strawberry pictures available on the web are also utilized in this study and formed the second part of the dataset. When using the Python script for web image crawling, to ensure the quality of the images in the dataset, it was firstly stipulated that the pixels of the crawled images were larger than 500×500. Data cleaning was conducted to screen the crawled images manually and remove blurred, duplicated, inconsistent, and watermarks, as shown in Fig. 3. Upon obtaining high-quality images, they were proportionally extracted to form the training and validation sets.

2) DATASET ANALYSIS

The strawberry images in the dataset used in this study has the following characteristics.

(1) The quality of strawberry images collected under different lighting conditions and shooting angles in an actual environment varies significantly.

(2) When certain types of interference, such as fruit flies, flowers, various diseases, and fertilizers left behind during fertilization exist, it is easy to produce a certain degree of false detection.

(3) The strawberry fruit feature pixel areas in the images are small, making them typical small targets. Convolutional neural networks (CNN) usually perform multiple down-sampling operations, which may cause the advanced feature map to lose the feature information of small fruits, thus reducing recognition accuracy.

(4) The differences in the length and width ratios of different classes of strawberry images are small; therefore, the size of the anchor frames and anchor ratios has to be considered carefully to ensure high masking accuracy.

(5) The color of immature strawberries is close to the pixel value of strawberry plant leaves, which could cause interference.

B. DATA PREPROCESSING

1) DATA ENHANCEMENT

Upon analyzing and cleaning the images in the original dataset, a Python script was used to determine the numbers of strawberry fruit images and GT frames in each category of the dataset. The 'Deformed' category was found to be much smaller than the other categories, and the data distribution was not balanced. These characteristics exist because when constructing the original dataset, the initial images of the deformed category were crawled from the network and then manually screened, and the number of images was much smaller than those in the other two categories. This ultimately resulted in the proportion of the deformed category being much lower than those of the other categories, resulting in an imbalance in the data distribution. To solve this problem and balance the number of strawberry fruit images in each category with the recognition accuracy of each category, we divided the initial dataset into two parts and performed segmented data enhancement. The first part mainly includes ' Green ripe stage ', ' White ripe stage ', 'Turning stage', and ''Mature'. The total initial number of images in this section is 2800, and the number of images after enhancement reached 5600 by using the single-class data enhancement method. The second part contained only deformed strawberry fruit images, and the images in this part were enhanced with a combination of scale, horizontal, vertical, rotate, move, darker, brighter, blur, salt, etc. The enhancement effect is shown in Fig. 4, in which (a) is the initial image in the dataset, (b) is a zoomed in or zoomed out image, (c) is a horizontally flipped image, (d) is a vertically flipped image, (e) is an image rotated by 90°, (f) depicts image panning, (g) shows a darkened image, (h) presents a lightened image, (i) depicts the image with Gaussian noise added, and (j) shows the image with pretzel noise added.

The statistical results of the enhanced dataset, as listed in Table 2, are divided into five categories: ' Green ripe stage', 'White ripe stage', 'Turning ripe stage', 'Mature', and 'Deformed', which contains 16,054 images and 25,816 labels. The ' Deformed ' category accounts for approximately 8%, which can effectively solve the data imbalance problem that appeared before, and basically satisfies the experimental requirements.

TABLE II NUMBER OF SAMPLES OF EACH TYPE OF STRAWBERRY FRUIT

Classes	Train		Validation		Total	
	Images	GT	Images	GT	Images	GT
Green ripe stage	4041	8264	252	472	4293	8736
White ripe stage	3348	5403	159	228	3507	5631
Turning stage	3790	5543	191	265	3981	5808
Mature	2750	3680	156	209	2906	3949
Deformed	1269	1530	98	162	1367	1692
Total	15,198	24,420	856	1336	16,054	25,816

2) DATA ANNOTATION

In this study, we used EISeg [23] for data annotation, which is an efficient and intelligent interactive segmentation annotation software developed based on Flying Paddle, which covers high quality interactive segmentation models with different objectives such as high precision or light weight. Using this approach, developers can realize the annotation of semantics and instance labels quickly and reduce the annotation cost. The annotation effect using EiSeg is shown in Fig. 5, in which there are five types of labels are ' Green ripe stage ', ' White ripe stage ', ' Turning stage ', 'Mature' and 'Deformed'.

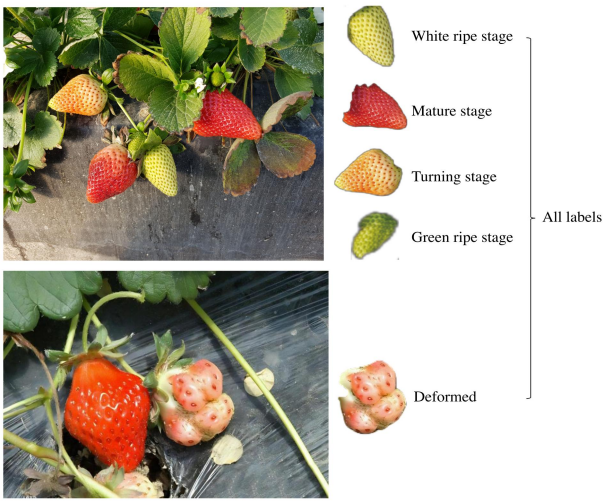

FIGURE 5. **Data Labeling.**

III. Instance segmentation of strawberry fruit with multiple growth cycles based on optimized Mask RCNN

A. KM-MASK RCNN ARCHITECTURE

The detection accuracy of two-stage detection models is generally higher than that of single-stage detection models, so we choose two-stage detection model(Mask RCNN) as our baseline [24]. Mask RCNN [25] is a popular two-stage instance segmentation method in the field of deep learning, mainly consisting of a feature extraction network (Resnet50) [26], multi-scale feature fusion network [27], region generation network (RPN) [28], feature pooling layer, and head-end network (FCN) [29]. It adds a new mask branch to the Faster R-CNN [30], which is responsible for outputting the mask information of each detected object. However, owing to the large differences in the sizes and shapes of different targets, this region-suggestion network-based approach may result in the poor detection and segmentation of small targets. The presence of smaller targets in the homemade dataset of strawberry fruits with multiple growth cycles used in this study has a large gap from the official range of anchor frames, which can significantly affect the detection accuracy of the model. Additionally, the initial Mask RCNN model uses Resnet50 as the feature extraction network along with a deep residual network to solve the problem that with the deepening of the number of layers of the network, the performance of the model tends to saturate, or even rapidly decline, but the number of model parameters will grow dramatically as the number of convolutional layers increases, resulting in higher computational complexity and greater memory consumption, which seriously affect the detection efficiency.

To solve the problems of Mask RCNN such as poor detection and segmentation of small targets, slow computation speed, and low detection efficiency and to improve the detection and segmentation accuracy of Mask RCNN further for strawberry fruits with multiple growth cycles, this study proposes a lightweight instance segmentation algorithm for strawberries with multiple growth cycles called KM-Mask RCNN. The specific network structure is shown in Fig. 6. The main innovations of this algorithm are (1) RPN improvements and (2) model backbone network improvements. Regarding RPN improvements, to enhance the adaptability of the algorithm to complex scenes, such as strawberry fruit images with multiple growth cycles, and to improve the recognition accuracy of the algorithm for small targets, the K-Means clustering algorithm was used to optimize the size of the original anchor frames and anchoring ratios of the Mask RCNN, which matches them with experimental datasets and improve the ability of the ascending model to detect small targets. Regarding the improvement of the model backbone network, the MobileNet V3 Large lightweight

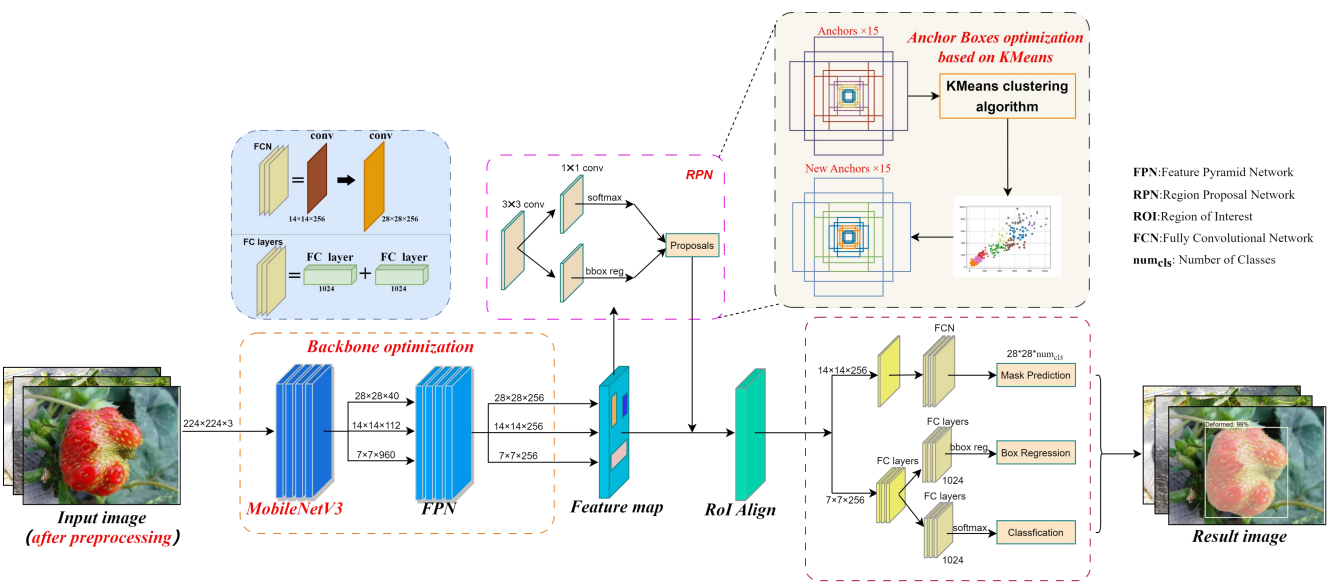

FIGURE 6. **KM-Mask RCNN structure diagram.**

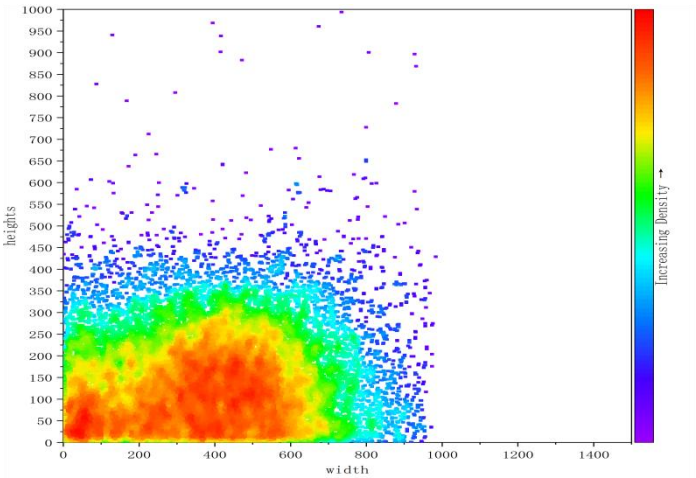

FIGURE 7. **Size distribution statistics of the dataset.**

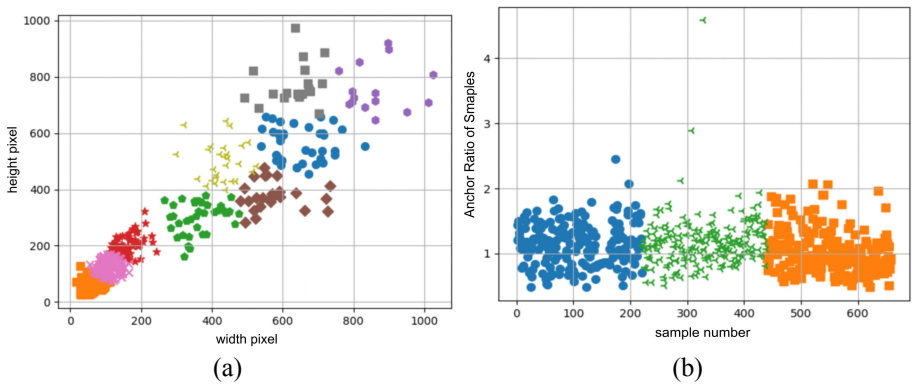

FIGURE 8. **Graphs of clustering results. (a) Anchor-frame height and width clustering; (b) Anchor-ratio clustering**

network was used to replace the original backbone network of Mask RCNN, Resnet 50, to reduce the number of algorithm parameters, reduce the model complexity, and improve the detection speed and with the help of depth-separable convolution to replace the original ordinary convolution, reduce the number of parameters and computation of the backbone network of Mask RCNN, reduce the computational memory required by the model, improve the detection speed, and realize a lightweight network model.

B. RPN Improvement

Mask RCNN, a two-stage algorithm based on RPN, employs RPN to extract multi-scale anchor frames with varying width-to-height ratios during training, facilitating model stage interconnection. Anchor frames, fundamental in target detection, predict target locations and scales by correlating generated frames on the image with real targets, enabling precise localization and classification. Anchor frame dimensions dictate the target scale range, with too small frames missing large targets and too large frames encompassing multiple or partial targets, causing redundant/false detections. Additionally, the initial anchor frame's width-to-height ratio critically impacts results, as mismatches with target ratios can distort target shapes or result in missed detections due to target aspect variability. Therefore, selecting a suitable anchor frame width-to-height ratio for targets with different shapes can effectively improve the detection accuracy.

All label scales in this study's dataset were aligned within the same coordinate system and visualized in Fig. 7. Based on the definition of targets in the MS COCO [31], the dataset of this study contains a significant number of small targets and middle targets. The large target has a relatively small proportion in the dataset used in this article. The original anchor box and anchor ratio of Mask RCNN take into account all scales of targets, Based on the above analysis, it can be concluded that strawberry fruits are primarily small and medium targets, while original anchor box and anchor ratio size of Mask RCNN consider targets of all scales. This will result in the model being unable to better extract strawberry fruit features, which is not conducive to high-precision target localization.

Therefore, new anchor frames and anchor ratio sizes suitable for the target of this task need to identified. To solve the above problems, this article proposes a method based on K-Means clustering algorithm to re-optimize the sizes of Mask RCNN anchor frames and anchoring ratios to make them suitable for a multi-growth cycle strawberry fruit dataset. The method clusters strawberry fruits with similar sizes and shapes by analyzing the clustering of the varied morphological features of strawberry fruits in the dataset.

TABLE III. COMPARISON OF PERFORMANCE OF NETWORK MODELS AT DIFFERENT SCALES.

scale	mAP(%)					P(%)	R(%)	F1-score(%)
	A	B	C	D	E			
1	85.15	85.80	93.06	81.97	84.19	86.10	86.00	86.05
2	86.96	89.31	90.66	88.66	87.02	88.50	91.10	89.78
3	**91.19**	**88.09**	**93.70**	**93.19**	**87.13**	**93.90**	**94.20**	**94.05**
4	89.77	88.31	92.73	84.77	83.52	88.70	90.60	89.64
5	90.67	76.19	83.09	71.20	69.23	78.10	89.30	83.33
6	92.94	74.31	89.46	90.84	84.57	86.40	90.70	88.50

TABLE IV. COMPARISON OF LOSS OF NETWORK MODELS AT DIFFERENT SCALE.

scale	train loss	classifier loss	box_reg loss	mask loss	objectness loss	rpn_box_reg loss
1	0.1976	0.0516	0.7680	0.0661	0.0120	0.0019
2	0.1575	0.0293	0.0586	0.0594	0.0066	0.0035
3	**0.1315**	**0.0220**	**0.0259**	**0.0645**	**0.0056**	**0.0136**
4	0.2291	0.0797	0.1063	0.0904	0.0122	0.0106
5	0.2402	0.1249	0.0281	0.0787	0.0066	0.0020
6	0.1649	0.0180	0.0339	0.0794	0.0184	0.0152

Based on the clustering results, the average size of the strawberry fruit morphology of each category is determined, and suitable anchor frames and anchoring ratios are generated. Accordingly, the customization of anchor frame settings for different types of strawberries can be achieved, which effectively improves the detection ability (accuracy and robustness) of the Mask RCNN model and makes the Mask RCNN better adapted to the task of instance segmentation of strawberry fruit images with multiple growth cycles.

The K-Means clustering results for all ground-truth widths, heights, and ratios are shown in Fig. 8.

It is known that the initial anchor frame scale combination used by Mask RCNN is (64, 128, 256, 512), and the ratio combination takes the value of (0.5, 1, 2). The clustering results in Fig. 8(a) show that this anchor frame scale combination cannot completely cover the data samples in the multi-growth-cycle strawberry fruit dataset. The mismatch between the anchor frame and labelled real frame size directly reduces the detection effect of the Mask RCNN on strawberry fruits with multiple growth cycles. We improved the anchor frame design to solve this problem. Specifically, the anchor frame size combination was improved to (128, 256, 512, and 1024), which could cover 100% of the height and 99.87% of the width. As the anchor ratio used in the Mask RCNN was (0.5, 1, 2), as shown in Fig. 8(b), most of the anchor ratios of the real frames in the dataset in this study were not more than 2, and and the ratio of the Mask RCNN anchor frames was modified to (0.5, 1, 1.5). To verify the validity of the modified anchor frame size and anchoring ratio size based on the clustering results, we considered different anchor frames and anchoring ratio sizes and combined them to compare their detection performance. The experimental results are presented in Table 3, in which scale5 is the initial value of the Mask RCNN anchor frames and the anchoring ratio size, and scale3 is the value selected in this study.

Scales1, 2, 3, 4, 5, and 6 in Table 4 represent respectively: (32, 64, 128, 256)+(0.5, 1, 1.5), (64, 128, 256, 512)+(0.5, 1, 1.5), (128, 256, 512, 1024)+(0.5, 1, 1.5), (32, 64, 128, 512)+(0.5, 1, 2), (64, 128, 256, 512)+(0.5, 1, 2), (128, 256, 512, 1024)+(0.5, 1, 2). In addition, A, B, C, D, and E represent ‘green ripening’, ‘white ripening’, ‘turning ripening’, ‘mature’ and ‘deformed’. The results in Table 3 indicate that when scale3 is selected, the mAP values reach 91.19%, 88.09%, 93.70%, 93.19%, and 87.13% for models A, B, C, D, and E, respectively, which are improved by 0.52%, 11.90%, 10.61%, 21.99%, and 17.83% compared to those when scale5 was selected, respectively. In addition, the F1-score reached the highest value when scale3 was selected, which is 8.00%, 4.27%, 4.41%, 10.72%, and 5.55% higher than when the other scales were selected. The above results indicate that the reselected anchor frame and anchor ratio size are more suitable for the strawberries with multiple growth cycles in the dataset.

Meanwhile, to verify the influence of different scale combinations on the convergence of the model, we conducted 30 epochs of training for the above six anchor frame scale combinations and compared the total loss of the model and the loss of each part of the model under each scale combination. The statistical results are shown in Table 4. Scales1, 2, 3, 4, 5, and 6 in Table 4 represent respectively: (32, 64, 128, 256)+(0.5, 1, 1.5), (64, 128, 256, 512)+(0.5, 1, 1.5), (128, 256, 512, 1024)+(0.5, 1, 1.5), (32, 64, 128, 512)+(0.5, 1, 2), (64, 128, 256, 512)+(0.5, 1, 2), (128, 256,

512, 1024)+(0.5, 1, 2). The results in Table 4 reveal that the network has the lowest convergence values for all types of losses for the combination of scale3. This result verifies the effects of the anchor frame size and anchoring ratio on the convergence of the network. Therefore, it is reasonable and valid to select scale3, i.e.: (128, 256, 512, 1024)+(0.5, 1, 1.5) as the final value.

Table 4 is analyzed to assess the impacts of different anchor frame ratio combinations on the convergence speed and accuracy of the model more intuitively, and the results are shown in Fig. 9, where Scale1-6 represent the six combinations of anchor ratios and anchor box sizes described earlier, and different colors represent different components of the loss function. Although Scale3 did not achieve the lowest loss for each branch, it achieved the lowest total convergence loss, which once again verifies the excellence of the scale selected in this study.

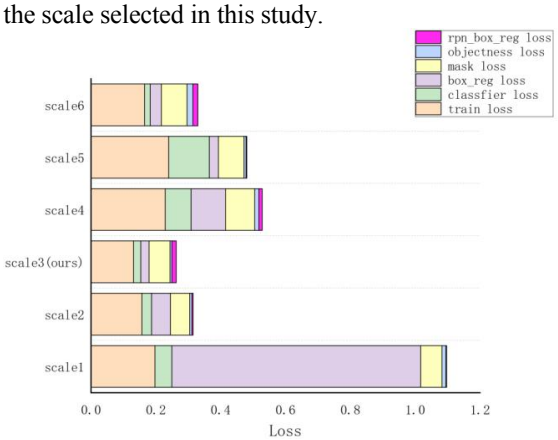

FIGURE 9. Comparison of convergence losses of network models at different scales.

C. Model backbone network improvements

The backbone network needs to extract the image features fully to realize the accurate detection of strawberries with multiple growth cycles. Although modifying the anchor frame and anchor ratio size of the Mask RCNN model effectively solves the problem of the poor ability of the model to extract feature information from small targets, the original backbone network of the Mask RCNN, ResNet50, is larger and requires large amounts of storage space and computational resources, and the training time is longer, which fails to satisfy the practical deployment requirements of the KM-Mask RCNN in the future. Therefore, the backbone network of Mask RCNN must be improved. Compared with ResNet50, the lightweight network MobileNet V3 is characterized by fewer parameters, shorter computation time, and shorter inference time; therefore, MobileNet V3 is generally preferred by many researchers [32]-[34]. Inspired by the above advantages, MobileNet V3 was used in this study to replace the original ResNet50 structure and achieve a higher inference speed.

The network structure of MobileNet V3 [35] is illustrated in Fig. 10. Compared with MobileNet V1 and MobileNet V2, three main points should be noted: (1) the time-consuming layer is redesigned, (2) the block (bneck) is updated and SE-Net is added, and (3) a new activation function, h-swish, is used. First, considering the high latency of the original time-consuming layer, to reduce the latency and to retain the high-dimensional characteristics, MobileNet V3 reduces the number of convolutional kernels in the first convolutional layer of the network from 32 to 16; then, the final stage is streamlined by moving the original last 1×1 convolutional layer behind the final average pooling layer, and the features in the last set are now computed with 1×1 spatial resolution instead of 7×7 spatial resolution. The final stage of the original and optimization is shown in Fig. 11. The above operation ensures that the network can extract deep features of strawberry fruit images in the dataset to recognize strawberry fruits with multiple growth cycles, thereby realizing lightweight operation of the network.

SE-Net [36] is a neural network structure based on the attention mechanism, which is mainly composed of Squeeze , Excitation, and scale; its specific structure is shown in Fig. 12, where X is the input, U is the output of each convolutional layer of the backbone network, Fex is an excitation operation, Fscale is a scale operation, and X' is the final output after combining the weights. Since U is the feature information obtained from the conventional

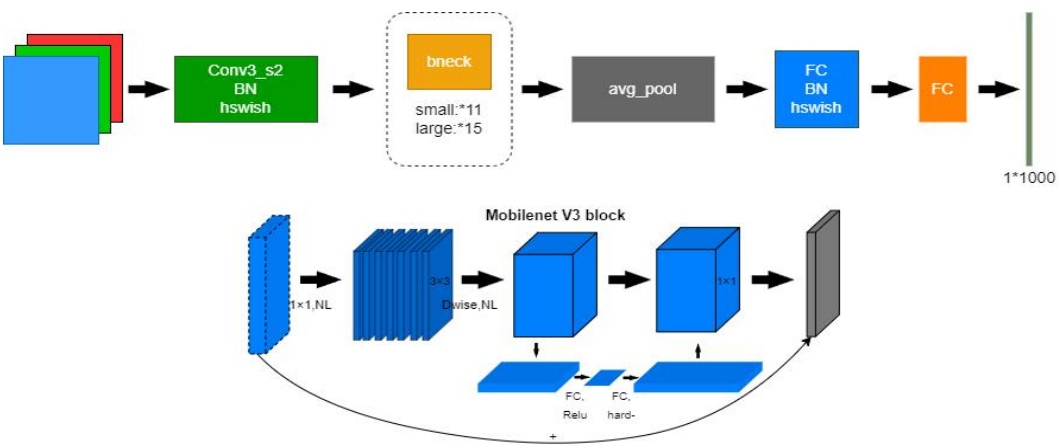

FIGURE 10. MobileNet V3 network structure diagram.

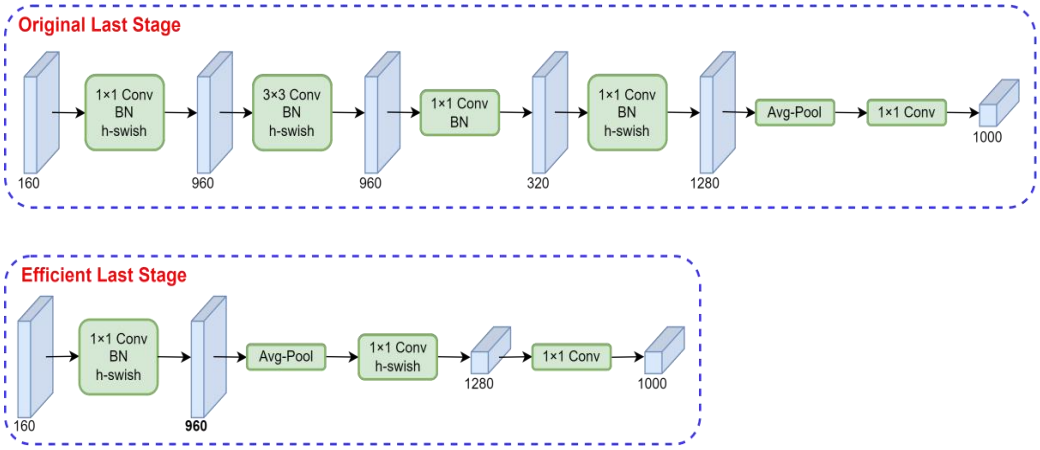

FIGURE 11. **Comparison of convergence losses of network models at different scales.**

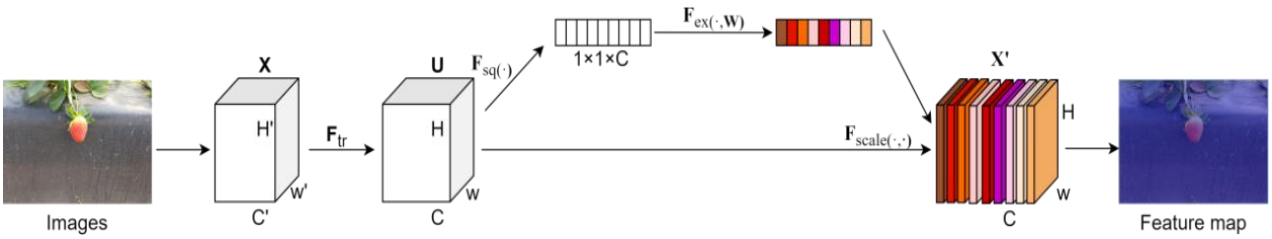

FIGURE 12. **SE-Net structure.**

convolutional Ftr that operates in the local space, providing sufficient information to extract the relationship between the features of different channels is difficult, which will result in the misdetection of strawberry fruits with multiple growth cycles. Therefore, this study introduces the squeeze, excitation, and scale operations into the KM-Mask RCNN.

Firstly, the squeeze operation compresses spatial features of each feature channel in the strawberry fruit image into one global feature (with dimensions $1 \times 1 \times C$) through global average pooling, fusing information across channels. Its formula is shown in (1). Subsequently, the Excitation operation utilizes FC layers to predict the importance of each channel, generating weights that are then applied (excitation) to the compressed feature map(with the dimension of $1 \times 1 \times C$), as formulated in Equation (2). Finally, these weights are multiplied with the original feature map's corresponding channels, enhancing relevant strawberry features and suppressing irrelevant ones. This weighting mechanism strengthens inter-channel connections for strawberries across growth stages, enabling the model to focus on the fruit region, enhance perception, and improve detection accuracy.

$$z_c = F_{sq}(u_c) \frac{1}{H \times W} \sum_i^H \sum_j^W u_c(i,j) \tag{1}$$

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1) z) \tag{2}$$

where zc represents the fused global features; Fsq represents the squeeze operation; Fex represents the excitation operation; δ represents ReLU function, which guarantees that the output is positive; and $W \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{\frac{C}{r} \times C}$, where r is a scaling parameter, that is primarily used to reduce the computational complexity and number of parameters in the network.

In the original ReLU function for calculation, the target image features are overly shielded, resulting in the model failing to effectively learn the strawberry fruit image features, reducing the detection accuracy. Therefore, in this study, the Swish function was used to replace the ReLU function, and its specific computational formula is shown in (3). Although introducing the Swish function can improve the accuracy of the neural network, it contains a sigmoid function that is computationally intensive, unsuitable for deployment in mobile terminals, and cannot satisfy the actual detection requirements of strawberry fruits with multiple growth cycles.

To solve these problems, this article proposes a solution method. First, the h-swish function was constructed and the sigmoid function in the swish function was replaced using ReLU6, as shown in (4). Using the ReLU6 nonlinear function not only ensures that the KM-Mask RCNN fuses different levels of features of the strawberry fruit image, but also enables it to be deployed in many software and hardware frameworks, avoiding the loss of numerical accuracy when quantizing and operating at a faster speed.

$$swish[x] = x \cdot \sigma(x) \tag{3}$$

$$\mathrm{h-swish}[x] = x \frac{RELU6(x+3)}{6} \tag{4}$$

Subsequently, the h-swish function was implemented as a segmented function. Since the memory occupied by the activation layer is halved as the resolution decreases, the cost of applying the nonlinear function decreases as the network gets deeper. Therefore, only the h-swish function is used in the second half of the MobileNet V3 model, and this

operation reduces the number of memory accesses and drastically lowers the latency cost. The KM-Mask RCNN also effectively utilizes the advantages of the h-swish function itself under the premise of successfully extracting the deep feature information from the strawberry fruit images.

In summary, we used MobileNet V3 to replace the original Resnet50 structure, under the premise of guaranteeing detection accuracy. The algorithm proposed in this paper, KM-Mask RCNN, can be lightened; thus, the number of parameters is greatly reduced, the computational complexity is simplified, and the processing speed is increased. Consequently, this approach can be deployed effectively on top of edge computing devices.

IV. Data acquisition and preprocessing

A. Implementation details

1) EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTINGS

The experimental hyperparameter settings applied in this study are listed in Table 5. Stochastic gradient descent (SGD)was used, and all hyperparameters employed weights pretrained on the COCO dataset for migration learning. In this study, 50 epochs were trained, and the learning rate was reduced by a factor of 10 when the epochs were trained to 8 and 11. The software environment used in the experiments is described in Table 5. CUDA10.1 and cuDNN10.2 were used for the experiments.

TABLE V EXPERIMENTAL HARDWARE AND PARAMETER SETTINGS.

Hardware or Software	Configuration
System	Windows 10
CPU	Intel Core i9-10900K
GPU	RTX-3090Ti/Jetson Xavier NX
Network framework	PyTorch 1.10.0
Batch size	4
Learning rate	0.01
Weight decay	1e–4
Momentum	0.9

2) INTRODUCTION TO EVALUATION INDICATORS

Detection evaluation metrics are used to measure the performance of the model in different domains and categories. The intersection over union (IoU) is calculated as the area of overlap between predicted and real regions within the same category divided by their total area. The IoU threshold determines the correctness of the detection evaluation of the algorithm, as shown in the following equation:

$$IoU = \frac{area(A_p \cap A_{gt})}{area(A_p \cup A_{gt})} \tag{5}$$

In (5), A_p and A_{gt} are the predicted and real regions, respectively. Upon determining the IoU threshold, all other categories of evaluation metrics were measured. The evaluation metrics used in this study contained two parts: model detection performance and model complexity. The model detection performance included the following metrics: precision (P), recall (R), average precision (AP), mAP, and F1-score. The first four metrics were used to assess the accuracy of strawberry fruit detection. The F1-score was the reconciled average of P and R, which was utilized to assess the quality of the model comprehensively, as shown in (6)–(10):

$$P = \frac{TP}{TP + FP} \times 100\% \tag{6}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{7}$$

$$AP = \int_0^1 P(R)dR \tag{8}$$

$$mAP = \sum_{k=1}^{m} AP_m \tag{9}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{10}$$

where TP represents the number of correctly identified samples; FP represents the number of missing negative samples; FN represents the number of missing positive samples; mAP is the overall detection accuracy of the model for single or multiple classes of identified objects in the dataset; m represents the number of classes in the dataset; and AP_m is the detection accuracy of the model for a class of identified objects.

The model complexity is measured using three metrics: parameters (params), floating-point operations (FLOPs), and inference speed (fps), as shown in (11) and (12):

$$params = C_o \times (k_w \times k_i \times C_i + 1) \tag{11}$$

$$FLOPs = [(C_i \times k_w \times k_h) + (C_i \times k_w \times k_h - 1)] \times C_o \times W \times H \tag{12}$$

where C_o is the number of output channels; C_i is the number of input channels; k_w is the convolution kernel width; k_h denotes the convolution kernel height; and W and H denote the length and width of the feature map, respectively.

3) EXPERIMENTAL PROGRAM

All experimental environment configurations in this study are listed in Table 5. The experiments were conducted on a self-made strawberry fruit dataset with multiple growth cycles. The experiments primarily included comparisons of different data enhancement schemes, ablation experiments, and comparisons of different instance segmentation algorithms, as follows.

Experiment 1 (comparison of different data enhancement schemes): The F1-score and mAP were compared after training different data enhancement schemes, verifying the reasonableness of the combined use of multiple data enhancement techniques.

Experiment 2 (ablation experiment): The ablation experiment was used to compare the F1-score and mAP after training the anchor frame and anchor ratio optimization, MobileNet V3, and KM-Mask RCNN individually and to evaluate the effectiveness of various improvement schemes proposed in this paper.

Experiment 3 (comparison of different instance segmentation algorithms): The FLOPs, params, detection

accuracy, and speed were compared after training Mask RCNN, YOLACT, CondInst, SOLOv2, and KM-Mask

Table VII. Ablation experiments.

Method	+Anchor Box optimization	+MobileNet V3	+Resnet18	+ConvNext Tiny	+RepViT	+GhostNet	P (%)	R (%)	F1-score (%)
Mask RCNN	×	×	×	×	×	×	78.70	89.30	83.33
Mask RCNN-A	√	×	×	×	×	×	89.10	93.90	91.44
Mask RCNN-B	×	×	√	×	×	×	86.80	91.60	89.14
Mask RCNN-C	×	×	×	√	×	×	87.10	87.80	87.45
Mask RCNN-D	×	×	×	×	√	×	89.70	90.80	90.25
Mask RCNN-E	×	×	×	×	×	√	86.40	88.10	87.24
Mask RCNN-F	×	√	×	×	×	×	90.40	90.90	90.65
Mask RCNN-G	√	×	√	×	×	×	88.20	91.30	89.72
Mask RCNN-H	√	×	×	√	×	×	88.00	90.50	89.23
Mask RCNN-I	√	×	×	×	√	×	89.20	90.90	90.04
Mask RCNN-J	√	×	×	×	×	√	87.50	88.00	87.75
Our method	√	√	×	×	×	×	93.90	94.20	94.05

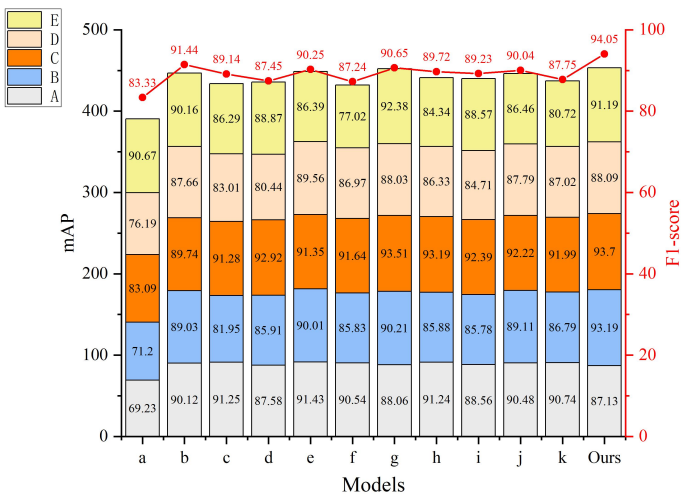

FIGURE 13. **Comparison of ablation experiment results.**

RCNN, quantitative and qualitative analysis were conducted to assess the detection performance of KM-Mask RCNN.

B. Comparison test of different data enhancement schemes

To verify the effect of the combined data enhancement techniques on the local enhancement and data detection results of each category, KM-Mask RCNN was used to train the dataset with different data enhancement methods for 50 epochs, and the experimental results shown in Table 6 were obtained. In the Table 6, Method 1 and Method 2 represent not using data augmentation, using the same data augmentation method for all categories and using partial data augmentation, respectively.

TABLE VI EXPERIMENTAL HARDWARE AND PARAMETER SETTINGS.

Methods	F1-score (%)	mAP (%)
Method 1(without augmentation)	61.89	77.30
Method 2(with normal augmentation)	90.71	91.88
Ours(partial data augmentation)	94.05	97.83

As shown in the results of the above table, the scheme of this study using a combination of multiple data enhancement methods reaches 94.05% and 97.83% for the F1-score and mAP, respectively, which are much higher than other combination schemes. Thus, the scheme selected in this study is effective.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3485490

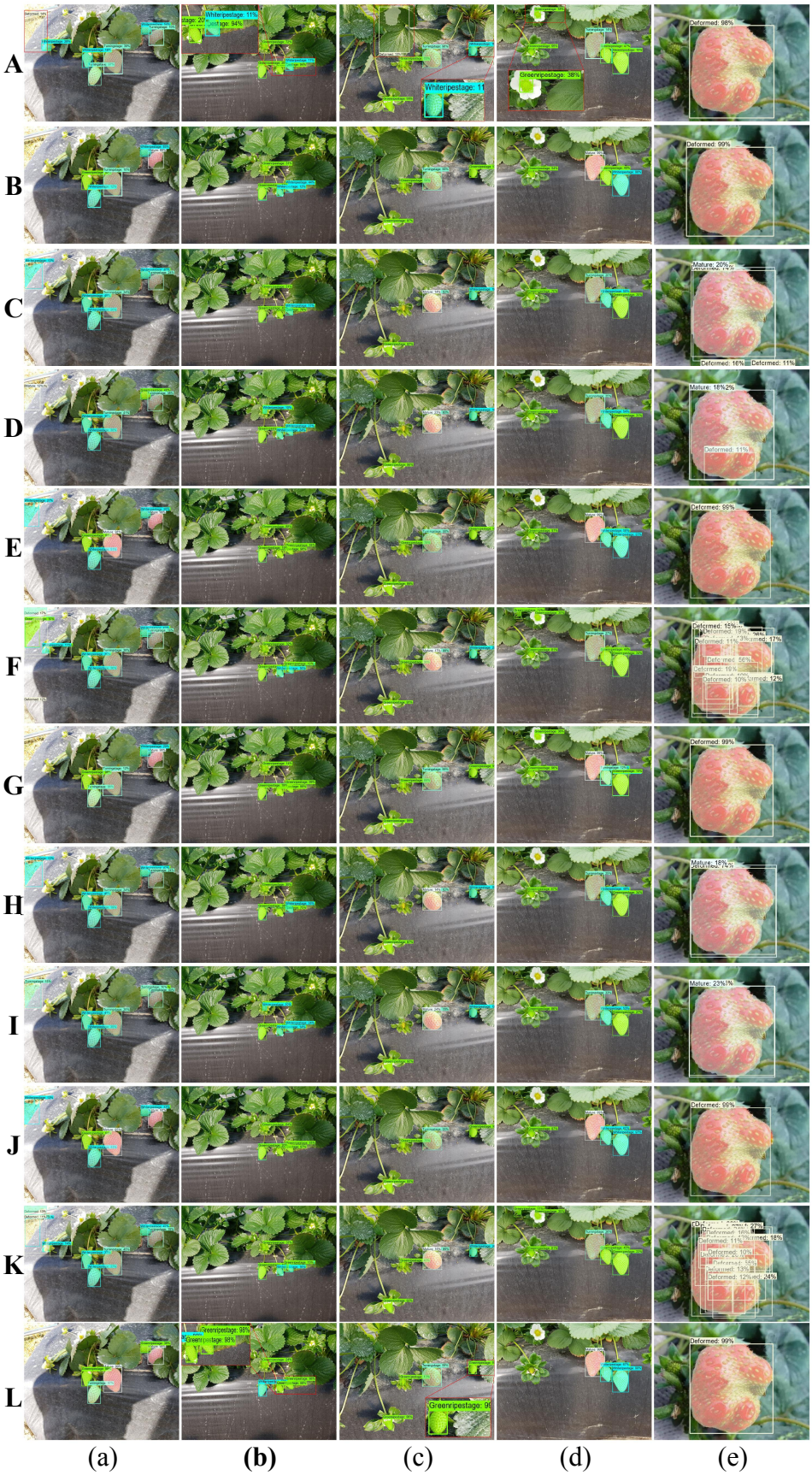

FIGURE 14. **Qualitative comparison of ablation experiment results.**

TABLE VIII. PERFORMANCE COMPARISON OF MAINSTREAM INSTANCE SEGMENTATION ALGORITHMS

Model	P (%)	R (%)	F1-score(%)	mAP(%)	error rate(%)				
				A	B	C	D	E	
Mask RCNN	78.7	89.3	83.3	90.7	76.2	83.1	71.2	69.2	6.63
YOLACT	90.8	85.6	88.1	88.6	84.5	88.9	74.5	82.3	6.23
YOLACT++	88.8	88.7	88.7	91.2	83.1	82.1	85.4	88.3	6.25
CenterMask	89.9	92.5	91.2	91.4	84.6	93.1	88.7	91.4	5.73
BlendMask	88.1	79.3	83.5	91.9	82.1	90.1	86.2	90.1	4.85
CondInst	90.8	90.1	90.5	84.1	83.4	90.5	88.9	83.5	9.11
SOLOv2	86.0	85.6	85.8	78.5	83.1	84.4	84.2	55.2	8.24
Cascade RCNN	89.4	87.1	88.2	86.1	84.8	88.1	85.0	80.6	3.22
YOLOv8-seg	79.9	89.0	84.2	92.1	86.7	90.5	91.1	87.0	7.41
YOLOv9-seg	79.6	89.0	84.0	92.1	84.2	91.2	91.9	87.6	8.56
Ours	93.9	94.2	94.1	91.2	88.1	93.7	93.2	87.1	5.85

C. Ablation experiments

We performed ablation experiments to verify the effectiveness of the various improvement schemes proposed in this study [37]. Mask RCNN-A was optimized based on Mask RCNN for the anchor frame size and anchor frame ratio, and Mask RCNN-B to Mask RCNN-F is an improvement of the backbone network of Mask RCNN by using ResNet 18, ConvNext Tiny, RepViT, GhostNet, MobileNet V3, respectively. The experimental results are shown in Table 7, revealing that when MobileNetV3 is used for improvement, the F1-score reaches 90.65%, which is higher than that of other lightweight backbone network models. Moreover, P and R of KM-Mask RCNN reach 93.90% and 94.20%, which are 15.2% and 4.9%, 5.7% and 2.9%, 5.9% and 3.7%, 4.7% and 3.3% and 6.4% and 6.2% higher than those of Mask RCNN, Mask RCNN-G, Mask RCNN-H, Mask RCNN-I and Mask RCNN-J, respectively. Additionally, P and R of Mask RCNN-A and Mask RCNN-F improved by 10.4% and 4.6% and by 11.7%, and 1.6%, respectively, compared to those of Mask RCNN. In addition, as shown in Fig. 13, the KM-Mask RCNN F1-score reaches 94.05%, which is 10.72%, 2.61%, 4.91%, 6.60%, 3.80%, 6.81%, 3.40%, 4.33%, 4.82%, 4.01%, and 6.30% better than the other models, respectively.The KM-Mask RCNN maintains high accuracy while capturing positive example samples better and minimising misclassification. In addition, the comprehensive performance of KM-Mask RCNN for each category recognition due to other models.

To demonstrate the segmentation effect of each improved method proposed in this study more intuitively, these models were qualitatively compared to predict the images, and the results are shown in Fig. 14. Here, A, B, C, D, E, F, G, H, I, J, K and L represent Mask RCNN, Mask RCNN+Anchor box optimization, Mask RCNN+ResNet18, Mask RCNN+ConvNext Tiny, Mask RCNN+RepViT, Mask RCNN+GhostNet, Mask RCNN+Mobilenet V3, Mask RCNN+ResNet18+Anchor box optimization, Mask RCNN+ConvNext Tiny+Anchor box optimization, Mask RCNN+RepViT+Anchor box optimization, Mask RCNN+GhostNet+Anchor box optimization, and KM-Mask RCNN, respectively. The locally enlarged part of the figure represents the details of the different models when segmenting the strawberry fruit image. According to Fig. 14(a), (c), and (d), Model A (Mask R-CNN) produces misdetections when segmenting the strawberry fruit image and recognizes irrelevant objects as targets. Additionally, its multitarget recognition ability is poor, and its detection accuracy is low, as shown in Fig. 14(b). For small targetssuch as strawberry fruits at the green ripening stage, Model A has the problem of low recognition accuracy. Model B (Mask RCNN+Anchor box optimization) has the problem of poor multitarget recognition ability, although it solves the problem of false detection to some extent and improves the detection accuracy for small targets. As shown in Fig. 14(a) and (e), there are phenomena of missed detection, false detection, and severe repeated detection from Model C to Model K. As shown in Fig. 14(b) and (c), Model L (KM-Mask RCNN) shows large improvements in both multitarget and small-target detection accuracies. Moreover, Model L exhibits a greater improvement in mask accuracy, and the mask matches the edges of the actual object more closely. The above analysis shows that the segmentation performance of the KM-Mask RCNN is greatly improved by optimizing the size of the Mask RCNN anchor frame and anchor ratio and replacing the original ResNet50 model with MobileNet V3.

D. Comparisons of different instance segmentation algorithms

To validate the superior performance of the KM-Mask RCNN instance segmentation further and explore its computational complexity, we selected several current mainstream instance segmentation algorithms (Mask RCNN, YOLACT [38], YOLACT++, BlendMask [39], CnterMask [40], CondInst [41], SOLOv2 [42], Cascade RCNN [43], YOLOv8-seg and YOLOv9-seg) and compared them with the proposed KM-Mask RCNN. The results are shown in Table 8, which reveals that KM-Mask RCNN improves the mAP of each category compared with those of Mask RCNN

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3485490

FIGURE 15. **Comparison of ablation experiment results.**

TABLE IX. COMPARISON EXPERIMENTS

Model	Params (M)	FLOPs (G)	fps		error rate(%)
			RTX3090	Jetson	
Mask RCNN	43.80	38.00	16.43	14.28	6.63
YOLACT	25.56	25.13	14.18	10.34	7.41
YOLACT++	42.02	22.35	30.68	22.32	6.25
CenterMask	25.12	21.94	124.8	24.25	4.85
BlendMask	32.15	25.76	27.78	16.29	5.63
CondInst	33.94	192.51	18.02	14.66	5.12
SOLOv2	27.32	81.08	15.60	12.52	8.24
Cascade RCNN	71.73	206.85	20.08	17.73	6.14
YOLOv8n-seg	3.26	12.10	116.00	25.91	7.41
YOLOv9-seg	5.07	236.70	100.00	23.98	8.56
Ours	**27.00**	**12.00**	**22.32**	**19.72**	**5.85**

and the above ten mainstream segmentation algorithms, with values of 91.19%, 88.09%, 93.70%, 93.19%, 87.13%, 89.40%, 89.0% respectively. In addition, KM-Mask RCNN achieved the lowest detection error rate.

To demonstrate the segmentation effect of the different algorithms more intuitively, the predicted images of the above models were compared qualitatively, and the results are shown in Fig. 15. Here, A, B, C, D, E, F, G, H, I, J and K represent Mask RCNN, YOLACT, YOLACT++, BlendMask, CenterMask, CondInst, SOLOv2, Cascade RCNN, YOLOv8-seg, YOLOv9-seg and KM-Mask RCNN, respectively. The locally enlarged red boxes and regions in the figure indicate the defects in the segmentation results. As shown in Fig. 15(a), Model A (Mask RCNN) suffers from the problem of duplicate detection, which produces different detection results for the same strawberry fruit, and simultaneously, its ability to detect small targets is poor. From Figs. 15(b) and (c), it can be observed that Model B (YOLACT) has poor mask quality when segmenting small targets that have a large gap with the edge of the real object, and the segmentation effect is poor in the presence of stems and leaf occlusion situation. Model C(YOLACT++), Model D(BlendMask) and Model E(CenterMask) are all misdetected, as shown in Fig. 15(a). As shown in Fig. 15(d), model D also suffers from duplicate testing. Model F (CondInst) is prone to missed detections when performing the segmentation task, as shown in Fig. 15(c). Additionally, the generated mask is often mutilated and fails to cover the fruit edges completely, as depicted in Fig. 16(d). As shown in Figs. 15(c) and (d), model G (SOLOv2) is prone to misdetection, producing different detection results for the same strawberry. Model H (Cascade RCNN) is prone to leakage detection during the detection process, thus ignoring some of the strawberry fruit targets. As shown in Fig. 15(a) and (b), the segmentation mask quality of Model F(YOLOv8-seg) is poor and there are missed detections. As shown in Fig. 15(d), Model J(YOLOV9-seg) suffers from severe false detections and segmentation. Model K (KM-Mask RCNN), in addition to maintaining a high accuracy rate, obtains the most complete edge profile of strawberries with the best mask quality. In summary, KM-Mask RCNN maintains a very high detection accuracy for small target strawberry fruits and stems and leaf masks affected fruits, in addition to solving the problems of misdetection and omission. KM-Mask RCNN also has a high mask quality; therefore, it maintains a high level of sophistication in the task of multi-growth cycle strawberry fruit image segmentation.

To explore the computational complexity and error rate of KM-Mask RCNN, we also compared the above models. The results are listed in Table 9 and it is evident that KM-Mask RCNN outperforms the other networks in terms of model complexity and inference speed. The results reveal that KM-Mask RCNN is a lightweight model with high inference speed that simultaneously achieves the requirement of high accuracy, which is a great advantage for the subsequent deployment of edge computing devices.

V. CONCLUSION

In this study, we developed a lightweight convolutional neural network model called KM-Mask RCNN for the efficient and accurate detection of strawberry fruits with multiple growth cycles. The main contributions of this study are as follows: (1) based on the K-Means clustering algorithm, we optimized the size of the Mask RCNN anchor frame and anchoring ratio, thereby improving the detection accuracy of the algorithm for small strawberry fruits and (2) MobileNet V3 was used to replace ResNet50, the original backbone network of the Mask RCNN, which effectively reduced the model complexity and enhanced the detection efficiency of KM-Mask RCNN. KM-Mask RCNN was

tested experimentally by constructing a dataset of strawberries with multiple growth cycles. The experimental results show that for the 'green ripe stage', 'white ripe stage', 'turning stage', 'mature', and 'deformed', the mAP values of KM-Mask RCNN were increased by 0.52%, 11.9%, 10.61%, 21.99% and 17.90% compared with that of Mask RCNN. In addition, the params, FLOPs, and fps of KM-Mask RCNN are reduced by 16.8M, reduced by 26G, and improved by 5.89, respectively, compared with those of Mask RCNN. Compared to YOLACT, YOLACT++, CenterMask, BlendMask, CondInst, SOLOv2, Cascade RCNN, YOLOv8-seg and YOLOv9-seg, the F1-score was improved by 5.93%, 5.32%, 14.33%, 2.87%, 3.60%, 8.25%, 5.81%, -3.89% and -3.75% , respectively; the FLOPs were reduced by 52.25%, 46.31%, 45.31%, 53.42%, 93.77%, 85.20%, 94.20%, 0.83% and 94.93%, respectively; and the fps on Jetson Xavier NX was improved by 90.72%, -11.64%, -18.68%, 21.06%, 34.52%, 57.51%, 11.22%, -23.89% and -17.76%, respectively. The above results demonstrate that KM-Mask RCNN is advantageous in terms of both detection accuracy and model complexity and can be better adapted to the need for picking robots with accurate sensing in the harvesting of strawberries with multiple growth cycles.

The algorithm proposed in this paper, as one of the typical representatives of two-stage algorithms, also has some disadvantages and still has room for improvement in future work. Firstly, there is still room for improvement in the robustness and generalization ability of the model. The existing multi-growth-cycle strawberry fruit dataset will be substantially expanded to incorporate a wider variety of strawberry images, aiming to continuously improve the model's robustness and generalization ability. Secondly, the real-time performance and mask segmentation accuracy of the model need to be further improved. We will focus on further optimizing the backbone and mask head of the KM-Mask RCNN model. Moreover, we plan to explore state-of-the-art network architectures to enhance the representational power and computational efficiency of the backbone. Additionally, advanced techniques, such as attention mechanisms and multi-scale feature fusion, will be investigated to improve the segmentation accuracy and detection efficiency of the mask head. Lastly, the adaptability of our approach still needs to be improved. We will attempt to explore the adaptability of our proposed approach to various agricultural applications. By studying the unique characteristics of different crops, we believe that the KM-Mask RCNN model can be tailored to fulfill the specific requirements of various agricultural scenarios, thereby promoting the rapid development of automated harvesting systems.

REFERENCES

[1] I. Pérez-Borrero, D. Marín-Santos, M.E. Gegúndez-Arias, and E. Cortés-Ancos, "A fast and accurate deep learning method for strawberry instance segmentation," *Comput. Electron. Agric.*, vol. 178, no. 105736, Nov. 2020.

[2] S. Hayashi, S. Yamamoto, S. Saito, Y. Ochiai, J. Kamata, M. Kurita, and K. Yamamoto, "Field operation of a movable strawberry-harvesting robot using a travel platform," *Jpn Agric Res Q JARQ*, vol. 48, pp. 307–316, Jul. 2018.

[3] E. Vitzrabin and Y. Edan, "Adaptive thresholding with fusion using a RGBD sensor for red sweet-pepper detection," *Biosyst. Eng.*, vol. 146, pp. 45–56, Jan. 2016.

[4] H. Azarmdel, A. Jahanbakhshi, S. S. Mohtasebi, and A.R. Muñoz, "Evaluation of image processing technique as an expert system in mulberry fruit grading based on ripeness level using artificial neural networks (ANNs) and support vector machine (SVM)," *Postharvest Biol. Technol.*, vol. 166, no. 111201, Aug. 2020.

[5] C. Wang, Y. Tang, X. Zou, L. Luo, and X. Chen, "Recognition and matching of clustered mature litchi fruits using binocular charge-coupled device (CCD) color cameras," *Sensors*, vol. 17, no. 11, pp. 2564, Nov. 2017.

[6] A. Arefi, A. M. Motlagh, K. Mollazade, and R. F. Teimourlou, " Recognition and localization of ripen tomato based on machine vision," *Aust. J. Crop Sci*., vol. 5, no. 10, pp. 1144-1149, Sep. 2011.

[7] W. Lin, J. Chu, L. Leng, J. Miao, and L. Wang, "Feature disentanglement in one-stage object detection," Pattern Recognition, vol. 145, no. 109878, Aug. 2024.

[8] X. Wang, H. Kang, H. Zhou, W. Au, and C. Chen, "Geometry-aware fruit grasping estimation for robotic harvesting in apple orchards," *Comput. Elecron Agric.*, vol. 193, no. 106716, Jan. 2022.

[9] Y. Yu, K. Zhang, L. Yang, and D. Zhang, " Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN," *Comput. Elecron Agric.*, vol. 163, no. 104846, Aug. 2019.

[10] H. Peng, J. Zhong, H. Liu, J. Li, M. Yao, and X. Zhang, "ResDense-focal-DeepLabV3+ enabled litchi branch semantic segmentation for robotic harvesting," *Comput. Elecron Agric.*, vol. 206, no. 107691, Mar. 2023.

[11] P. Chu, Z. Li, K. Lammers, R. Lu, and X. Liu, "Deep learning-based apple detection using a suppression Mask R-CNN," *Pattern Recognit. Lett*., vol. 147, pp. 206-211, May. 2021.

[12] P. Xu, N. Fang, N. Liu, F. Lin, S. Yang, and J. Ning, " Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation," *Comput. Electron. Agric.*, vol. 197, no. 106991, Jun. 2022.

[13] X. Liu, D. Wang, Y. Li, X. Guan, and C. Qin, "Detection of Green Asparagus Using Improved Mask R-CNN for Automatic Harvesting," *Sensors*, vol. 22, no. 23, pp. 9270, Nov. 2023.

[14] P. Cong, S. Li, J. Zhou, K. Lv, and H. Feng, "Research on Instance Segmentation Algorithm of Greenhouse Sweet Pepper Detection Based on Improved Mask RCNN," *Agronomy*, vol. 13, no. 1, pp. 196, Jan. 2023.

[15] X. Xu, X. Zhang, and T. Zhang, " Lite-yolov5: A lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 sar images". *Remote Sensing*, vol. 14, no. 4, pp. 1018, Feb. 2022.

[16] C. Li, J. Lin, B. Li, S. Zhang, and J. Li, "Partition harvesting of a column-comb litchi harvester based on 3D clustering," *Comput. Electron. Agric.*, vol. 197, no. 106975, Jun. 2022.

[17] M. Jiang, L. Song, Y. Wang, Z. Li, and H. Song, "Fusion of the YOLOv4 network model and visual attention mechanism to detect low-quality young apples in a complex environment," *Precis. Agric.*, pp. 1-19, Apr. 2021.

[18] S. Chen, X. Zou, X. Zhou, Y. Xiang, and M. Wu, "Study on fusion clustering and improved YOLOv5 algorithm based on multiple occlusion of Camellia oleifera fruit," *Comput. Electron. Agric.*, vol. 206, no. 107706, Mar. 2023.

[19] Y. Li, Q. Feng, C. Liu, Z. Xiong, Y. Sun, F. Xie, T. Li, and C. Zhao, "MTA-YOLACT: Multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting," *Eur. J. Agron.*, vol. 146, no. 126812, May. 2023.

[20] Y. Wang, G. Yan, Q. Meng, T. Yao, J. Han, and B. Zhang, "DSE-YOLO: Detail semantics enhancement YOLO for multi-stage strawberry detection," *Comput. Electron. Agric.*, vol. 198, no. 107057, Jul. 2022.

[21] C. Zhang, H. Ding, Q. Shi, and Y. Wang, "Grape cluster real-time detection in complex natural scenes based on YOLOv5s deep learning network," *Agriculture*, vol. 12, no. 8, pp. 1242, Aug. 2022.

[22] Pérez-Borrero, I., Marín-Santos, D., Gegúndez-Arias, M.E., Cortés-Ancos, E., 2020. Strawberry digital images (StrawDI), available at https://strawdi.github.io/.
[23] X. Min, X. Fei, H. D. Cheng, Y. Zhang, and J. Ding, "EISeg: Effective Interactive Segmentation," In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, Dec. 2016, pp. 1982–1987.
[24] X. Xu, X. Zhang, Z. Shao, J. Shi, S. Wei, T. Zhang, and T. Zeng, "A group-wise feature enhancement-and-fusion network with dual-polarization feature enrichment for SAR ship detection," Remote Sensing, vol. 14, no. 20, pp. 5276, Oct.2022.
[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," In Proceedings of the IEEE International Conference on Computer Vision,Venice, Italy, Oct. 2017, pp. 2961–2969.
[26] R. Padilla, W. L. Passos, T. L. Dias, S. L. Netto, and E. A. Da Silva, "A Comparative Analysis of Object Detection Metrics with a Companion Open-source Toolkit," *Electronics*, vol. 10, no. 3, pp. 279, Jan. 2021.
[27] W. Y. Hsu and W. Y. Lin, "Adaptive fusion of multi-scale YOLO for pedestrian detection," *IEEE Access.*, vol. 9, pp. 110063-110073, Aug. 2021.
[28] G. Liu and Q. Zhang, "Mask wearing detection algorithm based on improved tiny YOLOv3," Int. J. Patt. Recogn. Artif. Intell., vol. 35, no. 2155007, Feb. 2021.
[29] Z. Zhai, L. Yao, X. Sun, and X. Zhang, "Multi-objective salient detection combining FCN and ESP modules," *Multimed. Tools Appl.*, vol. 82, pp. 4405–4417, Jul. 2022.
[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans Pattern Anal Mach Intell.*, vol. 6, pp. 1137–1149, Jun. 2015.
[31] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft COCO: common objects in context," In Comput Vis–ECCV. Proceedings, Part V 13: 13th European Conference, Zurich, Switzerland, Sep. 2014, pp. 740–755.
[32] T. Zeng, S. Li, Q. Song, F. Zhong, and X. Wei, "Lightweight tomato real-time detection method based on improved YOLO and mobile deployment," *Comput. Electron. Agric.*, vol. 205, no. 107625, Feb. 2023.
[33] C. Wen, J. Wen, J. Li, Y. Luo, M. Chen, Z. Xiao, Q. Xu, and X. Liang, An. H, "Lightweight silkworm recognition based on Multi-scale feature fusion," *Comput. Electron. Agric.*, vol. 200, no. 107234, Sep. 2022.
[34] Z. Wu, F. Xia, S. Zhou, and D. Xu, "A method for identifying grape stems using keypoints," *Comput. Electron. Agric.*, vol. 209, no. 107825, Jun. 2023.
[35] A. Howard, M. Sandler, G. Chu, L.C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. Le, and H. Adam, "Searching for MobileNetV3. In Proceedings of the IEEE/CVF international conference on computer vision, Seoul, Korea, Nov. 2019, pp. 1314-1324.
[36] J. Hu, L. Shen, and G. Sun, " Squeeze-and-excitation networks," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, Jun. 2018, pp. 7132-7141.
[37] X. He, R. Cheng, Z. Zheng, and Z. Wang, "Small object detection in traffic scenes based on YOLO-MXANet," *Sensors*, vol. 21, no. 7422, Nov. 2021.
[38] D. Bolya, C. Zhou, F. Xiao, Y. Lee, "YOLACT: Real-time Instance Segmentation. In Proceedings of the IEEE/CVF international conference on computer vision, Seoul, Korea, Nov. 2019," pp.9157-9166.
[39] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, Jun. 2020, pp. 8573-8581.
[40] Y. Lee, and J. Park, "Centermask: Real-time anchor-free instance segmentation". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, Jun. 2020, pp. 13906-13915.
[41] Z. Tian, C. Shen, and H. Chen, "CondInst: Conditional convolutions for instance segmentation." In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Aug. 2020; pp. 282-298.
[42] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," A*dvances in Neural information processing systems*, vol. 33, pp. 17721-17732, Dec. 2020.
[43] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, 1483–1498. Nov. 2019.

Peichao Cong received the M.S. degree in solid mechanics from Harbin Institute of Technology, China, in 2005, and the Ph.D. degree in aircraft design from Harbin Institute of Technology, China, in 2009. From 2009 to 2010, he was a postdoctoral researcher at Shenyang Institute of Automation, Chinese Academy of Sciences. From 2016 to 2019, he was a postdoctoral researcher at Shanxi Coking Coal Group Co. He is currently a Distinguished Associate Professor at the School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology. His research interests include self-driving cars, dynamics modeling and simulation, and autonomous navigation and control problems of intelligent mobile robots.

Yutao Xu received the B.S. degree in Mechanical and Electronic Engineering from Jiangsu University of Science and Technology, China, in 2022. He is currently pursuing the M.S. degree with the School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology, China. His research interests include smart agriculture.

Tianheng Li received the B.S. degree in Process Equipment and Control Engineering from Jianghan University, China, in 2021. He is currently pursuing the M.S. degree with the School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology, China. His research interests include smart agriculture.

Shanda Li received the B.S. degree Mechanical Engineering from Guangxi University of Science and Technology, China, in 2021. He is currently pursuing the M.S. degree with the School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology, China. His research interests include Machine Vision and Precision Agriculture.

Hao Feng is currently pursuing the M.S. degree with the School of Mechanical Engineering, Guangxi University of Science and Technology, China. His research interests include Machine Vision and Precision Agriculture.

Xin Zhang received the M.S. degree in Mechanical Engineering from Northeast Electric Power University, China, in 2014. She is currently a Research Associate at the School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology. Her research interests .