

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/173878/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Jia, Weikuan, Wang, Zhifen, Zhao, Ruina, Ji, Ze , Yin, Xiang and Liu, Guoliang 2025. FBSM: Foveabox-based boundary-aware segmentation method for green apples in natural orchards. Expert Systems with Applications 260 , 125426. 10.1016/j.eswa.2024.125426

Publishers page: <http://dx.doi.org/10.1016/j.eswa.2024.125426>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



FBSM: Foveabox-based Boundary-Aware Segmentation Method for Green Apples in Natural Orchards

Weikuan Jia^{1,2*}, Zhifen Wang¹, Ruina Zhao¹, Ze Ji³, Xiang Yin⁴, Guoliang Liu^{5*}

1 School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

2 Key Laboratory of Facility Agriculture Measurement and Control Technology and Equipment of Machinery Industry, Zhenjiang 212013, China

3 School of Engineering, Cardiff University, Cardiff, CF24 3AA, United Kingdom

4 School of Agricultural Engineering and Food Science, Shandong University of Technology, Zibo 255000, Shandong, China

5 School of Control Science and Engineering, Shandong University, Jinan 250061, China

Abstract: Image segmentation of target fruits is an essential part of machine vision systems, aiming to facilitate more accurate early fruit measurement and machine harvesting in natural orchard environments. Given that instance-level pooling and down-sampling operations in conventional segmentation models often lead to the loss of detailed information, resulting in coarse partitioning masks, we endeavor to restore the boundary information of the masks. To achieve high-quality fruit segmentation with clear boundaries and refined masks, a Foveabox-based boundary-aware segmentation model (FBSM) is constructed by adding a multi-stage mask prediction head incorporating fine-grained features to the anchor-free detection model FoveaBox to realize accurate segmentation of green apple fruits. At each stage, a bi-layer fusion structure (BFS) is employed to fully fuse the fine-grained features in a double-fusion manner to guide the subsequent instance mask prediction. Finally, the boundary recovery module (BRM) is leveraged to recover the lost boundary detail information and obtain more accurate boundaries for the continuous optimization process of the fruit instance mask. Experimental results demonstrate that the proposed FBSM model achieves a mean average precision (mAP) of 60.7% for green apple instance segmentation in unstructured natural orchard environments, surpassing traditional instance segmentation models. Moreover, it enhances base model detection accuracy by 1.3% while striking a better balance between detection and segmentation performance.

Keywords: Instance segmentation, Green apple, FBSM, Fine-grained features

*Corresponding author: Weikuan Jia, Guoliang Liu



Weikuan Jia

wkjia@sdu.edu.cn (W Jia), 2020317046@stu.sdu.edu.cn (Z Wang),
2021317063@stu.sdu.edu.cn (R Zhao), jiz1@cardiff.ac.uk (Z. Ji), yinxiang@sdu.edu.cn (X Yin),
liuguoliang@sdu.edu.cn (G Liu)

1 Introduction

The task of image segmentation holds paramount importance in the domain of computer vision. Current research in this field has garnered unprecedented attention from scholars and has found widespread applications across various domains such as medical imaging (Asgari et al., 2021; Zheng et al., 2022; Chen et al., 2023), autonomous vehicles (Chen et al., 2023), virtual reality (Che et al., 2021; Klingenberg et al., 2023), industrial settings (Lian et al., 2020; Huang et al., 2021), and intelligent agricultural sciences (Maheswari et al., 2021; Luo et al., 2023). Additionally, practical applications like pesticide spraying (Hafeez et al., 2022; Meshram et al., 2023), fruit harvesting (Jia et al., 2020; Zhou et al., 2023), and early yield prediction (He et al., 2022; Palacios et al., 2023) have witnessed exponential growth. The gradual adoption of robots equipped with machine vision systems in orchards has increased automation and intelligence in the agricultural fields, facilitating efficient orchard work while reducing the potential cost of manual labor (Jha et al., 2019; Tian et al., 2020; Tang et al., 2023). The accuracy and speed of image segmentation are related to the performance of robotic vision systems, thus, achieving more accurate and efficient fruit image segmentation remains a critical scientific research task that helps to improve and enhance the accuracy and comprehensive capability of agricultural robots for fruit recognition, and greatly promotes real-time and accurate fieldwork in orchards.

Traditional machine learning methods have been widely employed in the early research on fruit recognition and segmentation (Jia et al., 2015; Zhang et al., 2020; Saleem et al., 2021), primarily leveraging color, shape and texture features. Arivazhagan proposed an efficient method for fruit recognition based on color and texture features, utilizing a minimum distance classifier (Arivazhagan et al., 2010). Song presented a novel two-step method to identify and locate pepper fruits within a complex greenhouse environment, resulting in an improved detection rate of the fruits (Song et al., 2014). Yu introduced an RGB-D (red-green-blue-depth) image-based recognition method for ripe litchi utilizing color and texture features to train a random forest binary classification model for litchi fruit recognition, which was used to estimate litchi fruit yield (Yu et al., 2021). To perform real-time and accurate yield prediction as well as automatic harvesting of green targets in natural orchards, Wang designed a new KDC (kernel density clustering) method that used a double-ranking algorithm to automatically select cluster centers and efficiently segment target fruits (Wang et al., 2021). Additionally, Choi utilized depth images to acquire 3D information for fruit region identification, coupled with CNN networks for precise identification, highlighting NIR spectral images as optimal for object detection and classification (Choi et al., 2017). While traditional machine learning methods can generally achieve accurate fruit recognition in scenarios with simple backgrounds, they often falter when confronted with images captured in real orchard environments. Here, fruits are frequently obscured by foliage and irregularly outlined, rendering them challenging to detect. Moreover, conventional target recognition algorithms exhibit limited stability when tasked with fruit detection in unstructured environments. They are susceptible to various environmental complexities, lack robustness, and struggle to meet the practical operational demands of smart devices deployed in natural orchard settings.

With the continuous maturity of artificial intelligence, automated agricultural harvesting and early yield prediction are gradually being effectively implemented by deep learning methods (Kamilaris et al., 2018; Zhu et al., 2018; Koirala et al., 2019; Farjon et al., 2023). Currently, a wide range of researchers have proposed many deep learning-based object detection and image

segmentation methods applied to green fruits (Liu et al., 2022; Zhang et al., 2022; Wang et al., 2022; Sun et al., 2023). The proposal of the classical two-stage instance segmentation model Mask R-CNN can be regarded as a milestone in the deep learning process, which has been applied widely in different research areas. Mask R-CNN (He et al., 2017) was constructed by adding a mask branch with a fully convolutional neural network (FCN) behind the two-stage anchor-based detection model Faster R-CNN (Girshick et al., 2015). Subsequently, it has been continuously optimized and applied to target detection and segmentation research in different fields, including fruit target detection research in agriculture (Jia et al., 2020). Yu put forward a Mask R-CNN based two-stage detection algorithm for strawberry fruit (MRSD), which provided a strong foundation for the localization of strawberry fruit picking points and facilitates the accurate operation of harvesting robots (Yu et al., 2019). Jia also optimized the Mask R-CNN to propose a stable segmentation network RS-Net for green fruits, which improved the segmentation accuracy for a single class of targets and experimentally proved that the recognition performance is better than other classical models (Jia et al., 2022). The above deep learning-based detection and segmentation models indeed improved the precision and robustness over traditional fruit recognition methods, and the performance can remain stable in complex outdoor environments, effectively distinguishing between fruit targets and background regions.

However, like Mask R-CNN, instance segmentation models (Cheng et al., 2019; Chen et al., 2019; Kirillov et al., 2020) tend to lose image detail information during image processing, which affects the high-quality requirements of the segmentation task. Thus, despite the encouraging achievements of such methods in fruit recognition research, they still inevitably lose some of the boundary information, resulting in a less complete and accurate boundary mask in practical natural applications, especially for heavily obscured homochromatic fruits. With the aforementioned advances, accurate and efficient harvesting of fruits in a real orchard environment is still challenging. Based on the two-stage instance segmentation model Mask R-CNN, refinemask (Zhang et al., 2021) is presented to achieve fine-grained instance segmentation focusing on border regions by replacing the mask branch with a newly built multi-stage optimization header incorporating fine-grained semantic features, which is validated to outperform the baseline model on public datasets. For the unstructured and complex environment of natural orchards, due to the different acquisition angles of data images and the diversity of fruit growth postures, there are cases of occlusion by same-colored branches and leaves and the mutual overlapping fruits in the acquired images, resulting in the surface contours of exposed fruits are not regular. Therefore, we are expected to clearly and accurately outline the boundary areas in the image segmentation process to achieve more accurate fruit segmentation. The main idea of this paper is to construct an instance segmentation model with fine boundaries, improve the feature fusion ability of the model for green fruits, and meet the requirements for high-quality segmentation of visual systems in natural orchards.

Therefore, to achieve high-quality instance segmentation with clear boundaries and fine masks, we add a multi-stage mask prediction head combining fine-grained features to the anchor-free detection model FoveaBox (Kong et al., 2020) and construct a FoveaBox-based boundary-aware segmentation model to achieve accurate instance segmentation of target fruits with different scales in the natural orchard environment. Specifically, the multi-stage mask prediction process begins with the RoI-Align operation being executed and the prediction mask size being expanded in stages. Meanwhile, the fine-grained features are iteratively merged by the formed bi-layer fusion structure to reduce the loss of details. In addition, a boundary recovery module focusing on border

regions is applied to obtain more accurate boundaries for optimizing the fruit instance mask. Experiments show that the detection and segmentation accuracy of the model on the green apple dataset is 62.1% and 60.7%, which exceeds the benchmark model Foveabox by 1.3 points in detection accuracy and the classical instance model Mask R-CNN by 2.1 points in segmentation accuracy, meeting the requirements for accurate work of picking robots under complex orchard conditions.

Briefly, the contributions of this paper can be summarized as follows:

(1) A Foveabox-based boundary-aware segmentation model (FBSM) is constructed by using a multi-stage mask prediction head combined with fine-grained features, which meets accurate instance segmentation requirements for green apples in the orchard environment.

(2) Bi-layer fusion structure (BFS) is employed to fully integrate fine-grained features in a twice-fused fashion at each stage to guide subsequent mask prediction.

(3) Boundary recovery module (BRM) is utilized to recover lost boundary detail information and obtain more accurate boundaries for continuous instance mask optimization process.

The remaining sections of this paper are arranged as follows. Section 2 presents the details of the produced green apple fruit dataset. Section 3 proposes a boundary-aware fruit segmentation model consisting of a multi-stage masked prediction head network. Section 4 describes the experimental setting and shows the comparison results with other popular models from both detection and segmentation perspectives. Finally, section 5 outlines the segmentation approach proposed in this work and presents the next steps for future work.

2 Dataset

2.1 Image acquisition

As publicly available green fruit datasets are rare and the images are often of the whole fruit tree in a distant state, which are not eligible for the various fruit images in this study. Therefore, in order to fully test the effectiveness of the method proposed in this paper to be applied in the real apple orchard, we collected close and distant fruit images at different scales from natural orchards and performed various data augmentation processes to generate a dataset in a certain format to support the favorable implementation of this study.

The images were acquired using Sony Alpha 7 II cameras, and abundant green apple images were collected at the apple production base in Longwang Mountain of Yantai City in Shandong Province. The resolution of all images was 6000×4000, and the captured contents were saved as 24-bit color images in JPG format.

To fully reproduce the actual conditions in the orchard in the experiment, the dataset was collected by taking into account the variation of lighting conditions throughout the day, including green apple fruits under natural light during the day and LED lights at night. Besides, the dataset also includes apple images with branch and leaf shading, fruit overlapping and small targets at a distance to simulate as much as possible the visual angle of the harvesting robot while working in the orchard.

2.2 Dataset production

A total of 1361 apple images with valid features are used to label and create the dataset. The dataset is made by dividing it into training and validation sets at a ratio of 7:3, with 953 and 408

images, respectively. Also, the image pixels are compressed to 600×400 to enhance the real-time performance of the algorithm. The contour information of the apple fruit in the images is labeled using the LabelMe tool to generate labels, with the green target fruit inside the labeled points and the rest of the images as background. Then, the labels and coordinates of each image are saved to the corresponding JSON file of the image. Finally, all the images in the dataset and the JSON files generated by the labels are integrated and converted into a dataset in COCO format. The details of images and instances on the divided apple dataset are below.

Table 1. The numbers of images and instances on the divided apple dataset.

Dataset	Images	Total instances
Training	953	4943
Validation	408	2194
Total	1361	7137

Moreover, an insufficient amount of data or incomplete samples during model training often leads to poor model generalization performance. To address this problem, we perform data augmentation on the collected images. Data augmentation is the basis for preventing overfitting of large convolutional neural networks, especially when the number of training samples is limited, and is mainly used to increase the training dataset to make it as diverse as possible so that the trained model has a stronger generalization ability and also minimizes the imbalance of training samples. The methods of data augmentation include geometric and photometric transformations. For geometric transformations, random shifts, flips and rotations are performed. For photometric transformations, the brightness, contrast, and saturation of the image are adjusted. The images are randomly enhanced and strengthened according to a certain probability.

3 Green apple segmentation model FBSM

This method combines the powerful recognition capability of existing two-stage segmentation methods with fine-grained semantic feature representation in the fruit segmentation process. Boundary-aware fruit segmentation is achieved by recognizing and segmenting fruit targets while using fine-grained features to supplement the lost detail information. To achieve this goal, we propose a new green fruit segmentation method FBSM with refined boundaries that can segment obscured green apples in a more fine-grained manner. FBSM builds on the one-stage anchor-free detector Foveabox by appending two simple yet efficient head networks, the mask prediction head and the semantic assisted head, to perform accurate instance segmentation of fruits. The overall framework flowchart is shown in Fig 1.

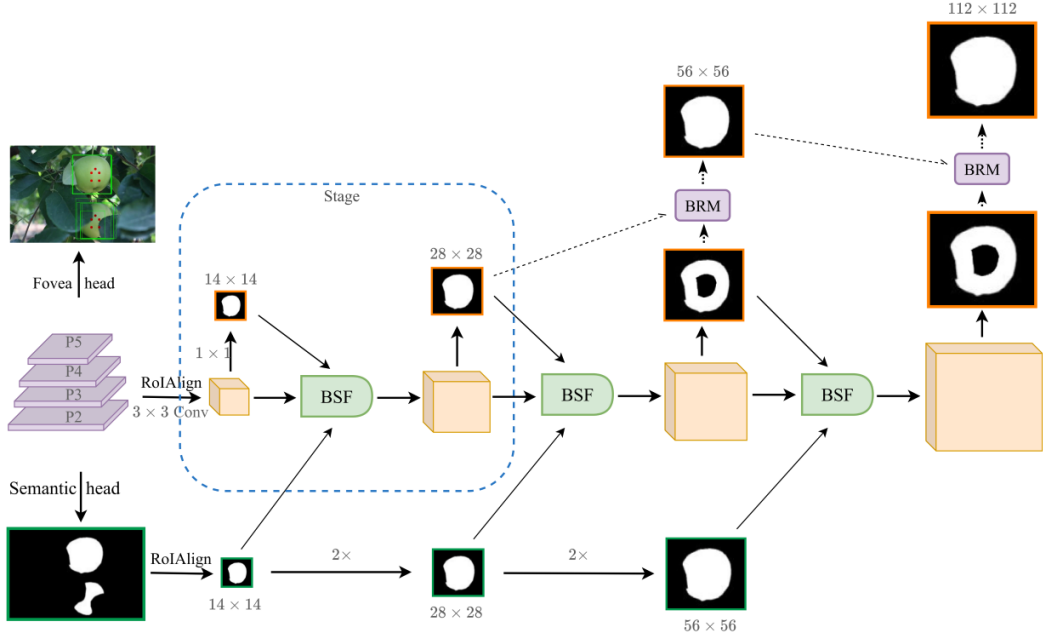


Fig. 1. Overall flowchart of the proposed model FBSM.

Specifically, the multi-stage mask prediction process starts with the execution of the RoI-Align operation to generate a coarse initial mask and optimizes the instance mask in three stages. Meanwhile, the semantic head followed after the P2 layer of the FPN extracts fine-grained features which are iteratively merged by the bi-layer fusion structure with instance masks and instance features. In addition, a boundary recovery module focusing on border regions is applied to reduce the loss of details and obtain more accurate boundaries to continuously optimize the instance mask. The modules are described as follows.

3.1 Foveabox detection model

The green fruit segmentation model introduced in this paper is based on the anchor-free FoveaBox detection model. In the benchmark detection model FoveaBox, the training targets are generated directly from the ground truth boxes without the need for anchors or region proposals. Considering that FoveaBox does not require any additional techniques in the post-processing stage and does not bring additional computational costs, so Foveabox is chosen as the detection part of this model to directly predict the confidence and the corresponding boundary of the target class.

As shown in Fig 2, the structure of FoveaBox contains a backbone network and a fovea detection head network. The backbone network uses a classical convolutional network to extract the effective features of the image, producing a convolutional feature map of the entire input image. The fovea detection head follows the feature extraction network and consists of two subbranches: the classification branch performs per-pixel category prediction on the output of the backbone network; and the box regression branch performs box regression on each location covered by the object. The two branches jointly predict the position of the central region of the object, as well as the bounding box for each valid position. FoveaBox succeeds in locating the target fruit region and generating the initial detection boxes.

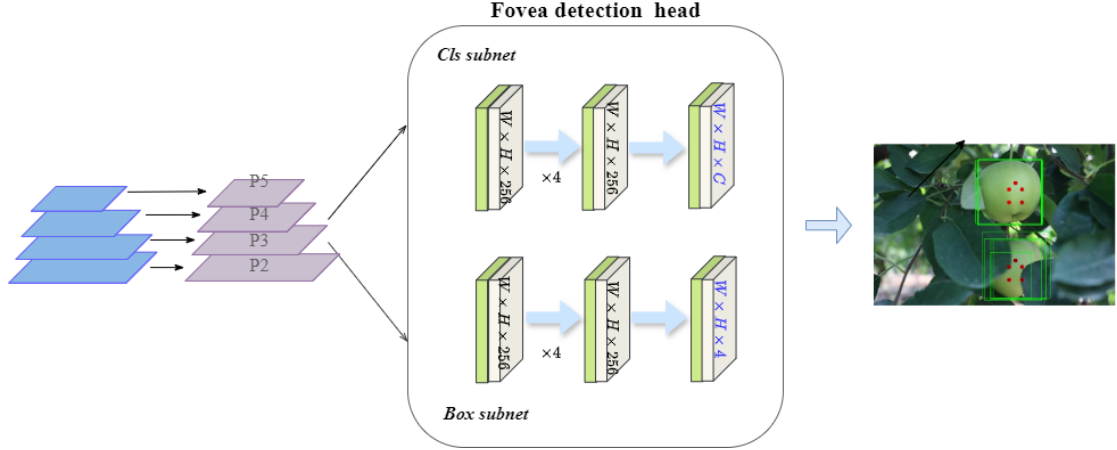


Fig. 2. The prediction stage of FoveaBox.

3.2 Semantic assisted head

The semantic assisted head performs semantic segmentation on the highest resolution P2 feature map from the FPN. The output shares the same resolution as the input image and the whole process does not cause semantic loss because of no spatial compression operation. The semantic head generates fine-grained features to fuse with an instance mask for high-quality instance segmentation.

The semantic assisted head has a simple and clear structure, mainly consisting of convolutional blocks, a binary classifier, and loss supervision. The pipeline is shown in Fig 3. First, four convolutional layers are used to extract the semantic features of the whole image, and then the classes of the target regions are predicted at the pixel level by a binary classifier. Finally, high-resolution semantic masks are generated and optimized under the supervision of a binary cross-entropy loss function. The two components, semantic features and semantic masks constitute the fine-grained features, which are sufficient to fully compensate for the detailed information lost during the instance-level operations.

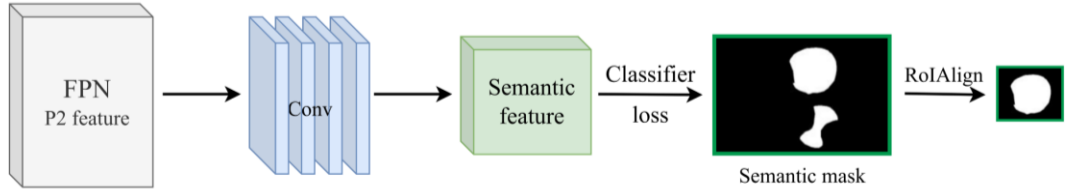


Fig. 3. Pipeline of the semantic assisted head.

3.3 Mask prediction head

Overall, the mask prediction head works by progressively optimizing the instance mask step-by-step in multiple stages to achieve fruit segmentation. The main block of each stage in the mask head is shown in the figure below. At each stage, a bi-layer fusion structure (BFS) is constructed to fully fuse the semantic features and semantic masks extracted by the semantic assisted head, maintaining high resolution while scaling up the features to attain high-resolution instance mask prediction. In addition, the boundary recovery module (BRM) is employed to focus on the border regions and generate clearer boundary predictions by interfusing the coarse instance masks of adjacent stages. The structure of each module is described in detail as follows.

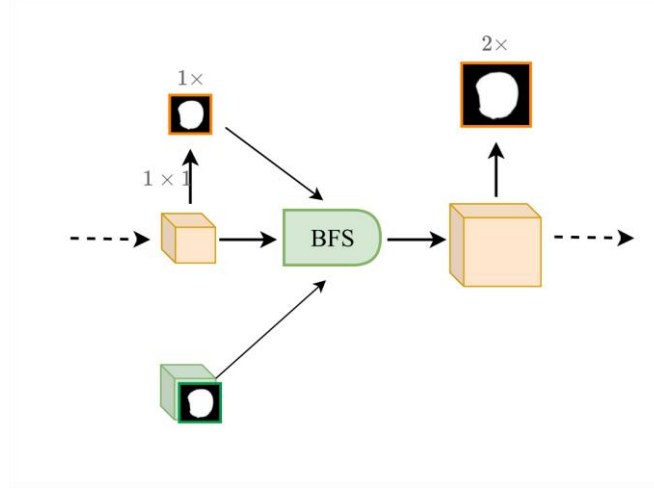


Fig. 4. Main block for each stage in the mask head.

3.3.1 Multi-stage mask optimization

In the mask, the features extracted by the 14×14 RoI-Align operation are first fed into two 3×3 convolutional layers to generate the instance features. Later, a 1×1 convolutional layer is used to predict the instance mask, and the spatial size of the mask is only 14×14 . This coarse mask is used as the initial mask for the later refinement stage.

The initial mask is then fed into the multi-stage mask optimization process that progressively refines the mask in an iterative manner. At each stage, features and masks are fed into the BFS for adequate fusion, and the output of the previous stage is adopted as the input of its next stage, thus achieving cross-fusion of features in adjacent stages. The mask head iteratively performs this mask optimization process and outputs a high-quality instance mask with a resolution of 112×112 .

3.3.2 Bi-layer fusion structure (BFS)

To efficiently bond fine-grained features to generate high-resolution instance masks, BFS is constructed to ensure that each region in the mask head perceives the information around it. BFS is responsible for fusing four inputs in each stage, which are the instance mask with the instance features from the previous stage, with the semantic features and semantic mask derived from the semantic assisted head.

Designed as a key module at each stage, the structure of the BFS is divided into two layers, as illustrated in Fig 5. In the first layer, a 1×1 convolution is applied to fuse input features and compress them, reducing the number of channels and avoiding too much computation. Following, three 3×3 atrous convolutions with different dilation coefficients are used to fuse information around individual positions while maintaining detail in small regions. In the second layer, the instance mask and the semantic mask are again concatenated with the initial fusion features generated by the first layer to provide semantic guidance for mask prediction in the later stages. Finally, the cascaded features are scaled up to a higher spatial size. Before the fused features in BFS are scaled up, the channels are halved by 1×1 convolution. Thus, although the feature size increases, the resulting computational cost is not high.

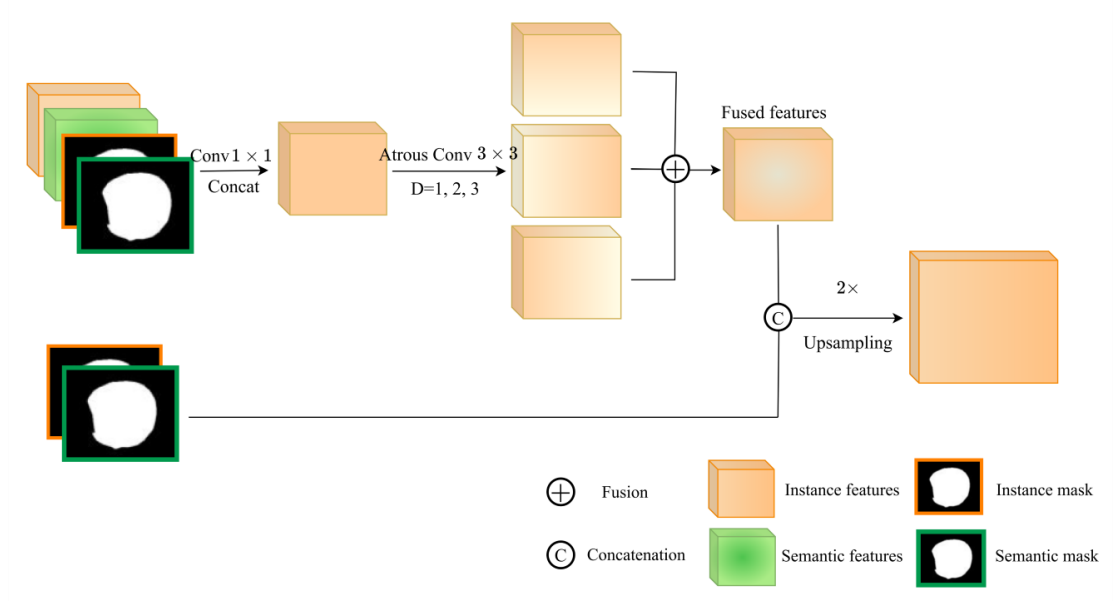


Fig. 5. The structure of the BFS at each stage.

3.3.3 Boundary recovery module (BRM)

Generating clear boundaries is crucial for accurate image segmentation. In order to refine the border regions and recover accurate boundary masks, we employed a boundary recovery module (BRM) that focuses on the border regions by a process in which the instance masks of the before and after stages are fused in a specific way. More accurate fruit contours are obtained by focusing on the border regions and recovering as much as possible of the detailed information lost in the mask prediction stage.

Identification of border region. As shown in Fig 6, the white line sketches the contour of the fruit, and the gray area is the border region of the fruit instance. Different instance masks M correspond to different border regions. M_k denotes the fruit instance mask of the k_{th} stage, and the spatial size of the mask can be expressed as $14 \cdot 2^k \times 14 \cdot 2^k$, where $k=1, 2, 3$. The pixels within a distance less than $\hat{D}$ from the outline of the instance mask constitute the border region. Predicted boundary mask B_k , which is used to represent the border region of the instance mask M_k , is defined as follows.

$$B_k(i, j) = \begin{cases} 1, & \text{if } D_{ij} \leq \hat{D} \\ 0, & \text{else} \end{cases} \quad (1)$$

where, (i, j) is the pixel location in the instance mask M_k . And D_{ij} is the euclidean distance from the pixel in the M_k to the nearest pixel on the instance contour.

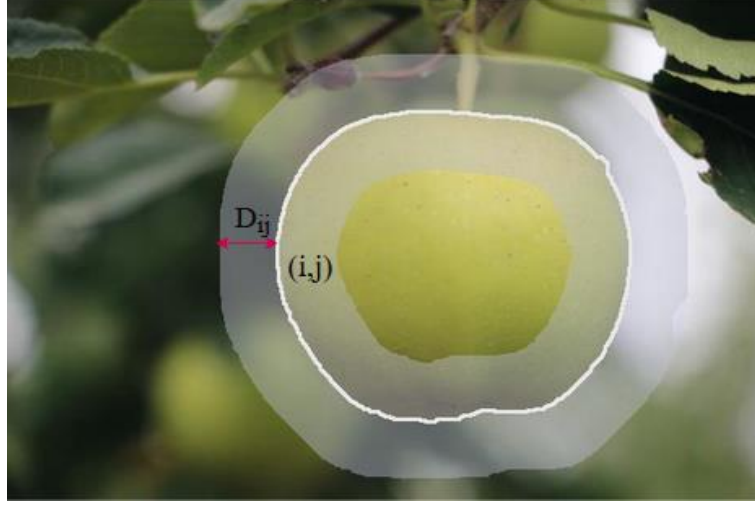


Fig. 6. Graphical representation of the object border region.

Recovery of boundary information. In order to optimize the instance mask effectively, we take the second stage as an example to illustrate the inference process of the boundary information recovery module, as shown in Fig 7.

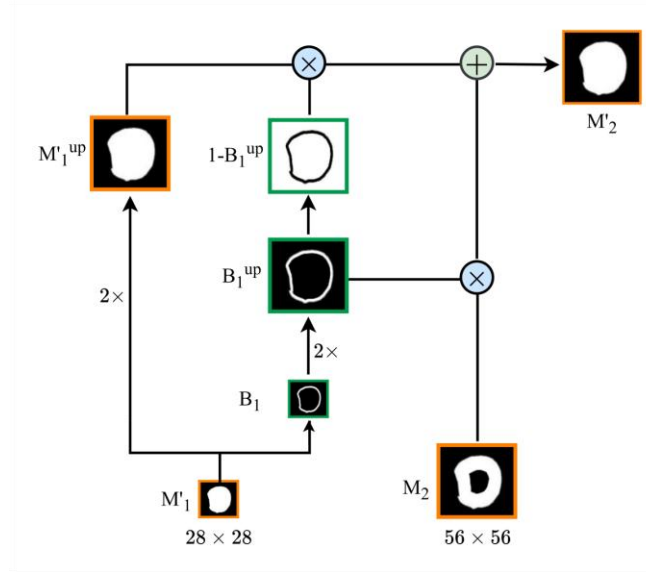


Fig. 7. The inference process of boundary information recovery. (M'_1 and M_2 are inputs of the BRM, B_l is the generated border region of M'_1 , and M'_2 is the output instance mask of the second stage.)

For each fruit instance, a rough but complete mask M'_1 of size 28×28 is output in the first optimization stage, while the boundary mask B_1 is produced. In the BRM, the generation process of a more refined and complete instance mask M'_k (the final output of stage k) can be expressed as:

$$M'_1 = M_1 \quad (2)$$

$$M'_k = (B_{k-1})_{up} \otimes M_k + (1 - (B_{k-1})_{up}) \otimes (M'_{k-1})_{up} \quad (3)$$

where $\otimes$ and $+$ indicate pixel-level multiplication and addition respectively. This convolutional process is repeated until the optimized mask is obtained. The up denotes the bilinear up-sampling operation.

Realization details. Considering that target fruits have different scales, the instance mask is first adjusted to a fixed size, and then a boundary mask is obtained. The first stage outputs an initial instance mask with a 28×28 size. The two succeeding stages yield masks of sizes 56×56 and 112×112 , respectively, with certain border areas $Area^k$ trained using supervised signals. These areas are defined by the ground truth mask and the prediction mask of the previous stage, as shown in Equation (4).

$$Area^k = (B_{k-1}^g \cup B_{k-1})_{up} \quad (4)$$

B_{k-1}^g indicates the border region of the ground truth mask at the $k-1$ th stage. B_{k-1} represents the predicted border region at the $k-1$ th stage. $\cup$ means the union of the above two border regions. For the last two stages with output size of $S_k \times S_k$ ($k = 2, 3$), the training loss L^k can be specified as:

$$L^k = \frac{1}{\delta_n} \sum_{n=0}^{N-1} \sum_{i=0}^{S_k-1} \sum_{j=0}^{S_k-1} Area_{nij}^k \cdot l_{nij}, k=2,3 \quad (5)$$

$$\delta_n = \sum_{n=0}^{N-1} \sum_{i=0}^{S_k-1} \sum_{j=0}^{S_k-1} Area_{nij} \quad (6)$$

where N is the number of fruit instances in the image. l_{nij} is the binary cross entropy loss at position (i, j) of instance N . $Area_{nij}^k$ represents the area of the grid cell (i, j) in the k -th stage for the N -th sample.

The mask prediction head has three fine-tuning stages. In the initial stage ($k=1$), the average binary cross-entropy loss function is utilized. The subsequent two optimization stages use the loss function of Equation (5) above. In addition, the semantic-assisted head uses the average binary cross-entropy loss function. The loss weights for the initial mask prediction stage and the three refinement stages are 0.25, 0.5, 0.75 and 1.0, respectively.

4 Results and Analysis

4.1 Experimental Configurations

This study is conducted on a server with Ubuntu 16.04 OS and V10.0 CUDA, using the GPU with a single block of Tesla V100-PCIE-32GB configured in the server. The experiments are based on MMDetection and Pytorch framework, and the training and testing of the algorithm on the green apple dataset are performed using Python programming language.

4.2 Evaluation Metrics

The common algorithm evaluation metric is the Precision-Recall (P-R) curve, which consists of precision and recall rates. Precision indicates the proportion of actual positive samples among the instances predicted as positive samples, and Recall is a measure of positive sample coverage, which indicates the proportion of positive samples correctly predicted among all actual positive samples, as shown in Equation (7)(8).

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

For the green apple dataset, we employ the common Average Precision (AP) metric as an evaluation metric for the algorithm results to evaluate the performance of the model in segmenting

target apples. The average precision AP can be obtained from P-R results and is calculated as shown in Equation (9). AP is the average accuracy for only one target class, while mAP is the average accuracy of AP for all target classes, as shown in Equation (10). The AP₅₀ in the evaluation metric is the AP value at an Intersection-over-Union (IoU) threshold of 0.5.

$$AP = \int_0^1 P(r)dr \quad (9)$$

$$mAP = \frac{\sum_{C=1}^C AP_C}{C} \quad (10)$$

Where the dataset category used in this experiment has only one type of target fruit, thus the category number C is 1, and AP is equal to mAP. Also, the mAP_s, mAP_m, and mAP_L metrics are used to indicate the segmentation accuracy of the model for target fruits in the small, medium, and large scale ranges.

4.3 Model Training

Green apple images of uniform size are fed into the training process of the model with N samplable regions of interest (RoI) per image. Pre-trained weights on the COCO dataset are used as the initialization parameters of the model. Iterative training with a batch size of 8 and a training period of 12 is performed. The initial learning rate is set to 0.0025 and decreases ten times in the first ten training cycles, and the corresponding parameters are retained at the end of each training cycle. During the training period the weight decay is set to 0.0001 and the momentum is 0.9 to prevent overfitting phenomena. The model is trained using the above configuration, and the loss variation during training is visualized to make the results more intuitive and clearer, as shown in Fig 8.

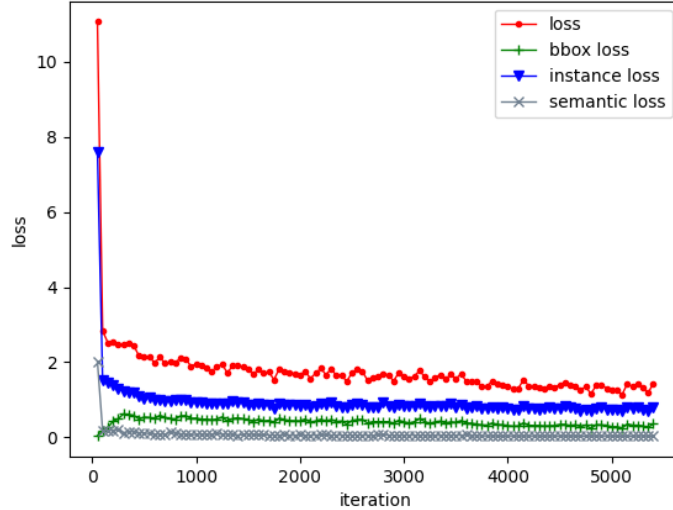


Fig. 8. Different loss evolution curves during model training.

As shown in the figure, the overall loss value of the model decreases rapidly at the beginning of the training and soon plateaus. There is a small fluctuation with the increase in the number of training iterations, showing a decreasing trend. Finally, all the losses remain stable, reflecting that the model has strong robustness and stability. In addition, accuracy (acc) is used to measure the recognition ability of the algorithm for positive samples, indicating the ratio of the number of correct

predictions to the number of all predictions, as shown in Equation (11). The higher acc means that the algorithm is more accurate and the algorithm is better at classifying positive and negative samples.

$$acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

where True Positives (TP) indicates the number of times a positive sample is correctly predicted as a positive sample. True Negatives (TN) denotes the number of times a negative sample is correctly predicted as a negative sample. False Positives (FP) indicates the number of times a negative sample is incorrectly predicted as a positive sample. False Negatives (FN) indicates the number of times a positive sample is incorrectly predicted as a negative sample.

The change curve of model accuracy obtained on the validation set is depicted in Fig 9. The model always maintains a high correct rate as the number of iterations increases during the training process. Although there are fluctuations, the overall accuracy is higher than 90% and shows an upward trend overall. This effectively shows that the model can efficiently cope with the classification of positive and negative samples, maintaining a high correct detection and segmentation rate, demonstrating the accurate performance of the model.

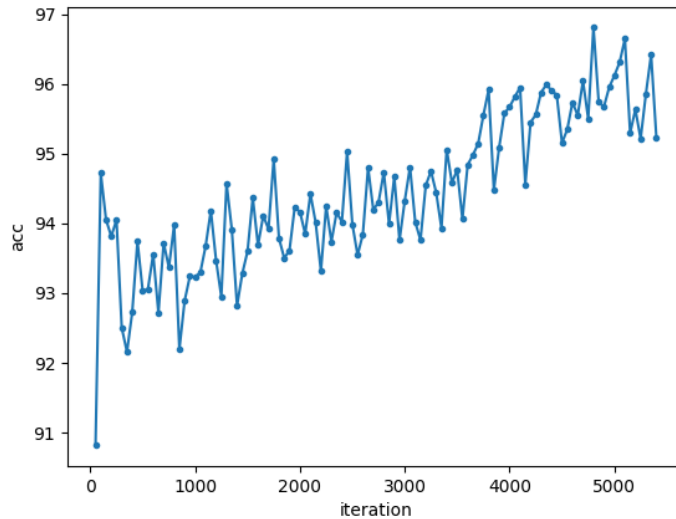


Fig. 9. Acc change curve during model training.

4.4 Detection and Segmentation Results

The main goal of this section is the systematic validation of the proposed new model on the created apple dataset. The orchard presents unstructured and complex characteristics in the natural environment. To fully consider the actual operating environment of the apple orchard, the experiment focused on a variety of interference conditions such as branch and leaf shading, overlapping cover, low light at night, after rain, and small fruit, etc. The segmentation effect is shown in Fig 10.



a Original apple images

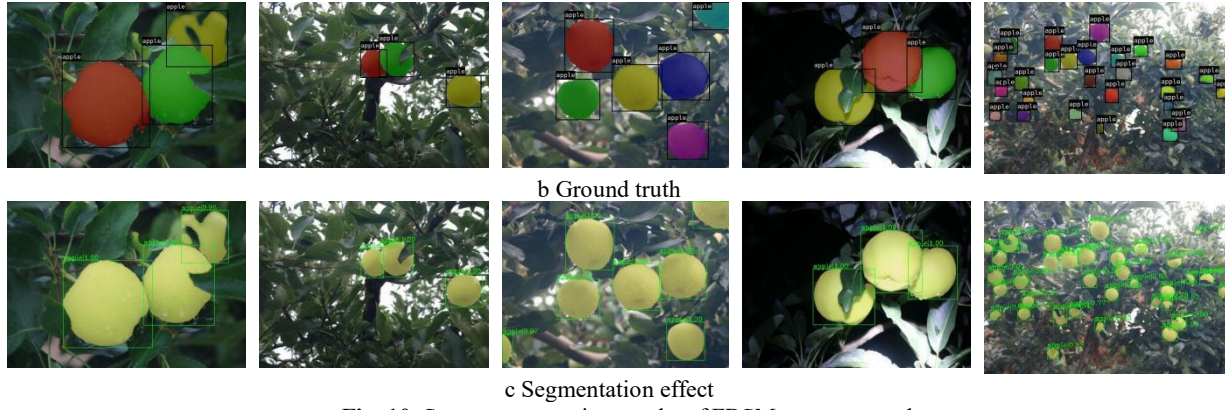


Fig. 10. Some segmentation results of FBSM on green apples

To further objectively show the overall performance of the new model, the detection and segmentation effects of the model were verified separately on the constructed apple dataset, and the accuracy, recall, and other objective evaluation metrics are listed in Table 2 to comprehensively assess the overall performance of the model.

Table 2. Detection and Segmentation results on the validation set of the *apple* dataset.

<i>apple</i>	Detection result						Segmentation result					
Metric	mAP ^b	AP ₅₀ ^b	mAP _S ^b	mAP _M ^b	mAP _L ^b	AR ^b	mAP ^s	AP ₅₀ ^s	mAP _S ^s	mAP _M ^s	mAP _L ^s	AR ^s
Value	62.1	86.8	46.4	69.0	86.2	68.1	60.7	85.3	43.3	68.0	89.5	66.1

According to Fig 10 and Table 2, from the detection effect, the FBSM model can accurately locate the target fruit and give the precise rectangular bounding boxes, without any false detection and omission of small target fruit. From the segmentation effect, the FBSM model can clearly delineate the target fruit boundary contour. From the objective evaluation indexes, the average accuracy of the detection and segmentation of the FBSM segmentation model proposed in this paper for green apples is 62.1% and 60.7%, respectively, and the detection and segmentation performance of the model is considered satisfactory. Thus, for the blurring of target fruit characteristics caused by interference conditions such as illumination, branch and leaf shading, overlapping targets, and small distant targets, the FBSM model can efficiently achieve the detection and segmentation of target fruit, demonstrating the generalization ability and robustness of the model, which is suitable for operation in complex orchard environments.

4.5 Comparison Experiments

To better validate the effectiveness of the FBSM model, current mainstream models and more advanced models are selected for comparison, and the detection and segmentation results on the constructed green apple dataset are listed in Tables 3 and 4 respectively to achieve a comparison of the advantages of the various models.

Table 3. Comparisons in the detection effect.

Model	Backbone	mAP ^b /%	AP ₅₀ ^b /%	AP ₇₅ ^b /%	AR ^b /%	mAP _S ^b /%	mAP _M ^b /%	mAP _L ^b /%
-------	----------	---------------------	----------------------------------	----------------------------------	--------------------	----------------------------------	----------------------------------	----------------------------------

<i>Evaluation metric = bbox</i>								
YOLOF (Chen et al., 2021)	ResNet50+C5	50.6	79.7	52.1	59.7	27.8	60.4	83.4
FCOS (Tian et al., 2019)	ResNet50+FPN	59.4	87.3	64.3	66.9	44.6	64.3	85.1
FSAF (Zhu et al., 2019)	ResNet50+FPN	60.7	87.8	66.1	67.5	44.6	67.1	85.2
BoxSnake (Yang et al., 2023)	ResNet50+FPN	59.6	79.1	66.6	-	40.5	67.2	89.2
Faster R-CNN	ResNet50+FPN	61.0	86.2	67.5	66.8	45.3	68.0	85.5
FoveaBox	ResNet50+FPN	60.8	87.8	67.0	67.8	44.7	66.5	85.4
FBSM(Ours)	ResNet50+FPN	62.1	86.8	69.1	68.1	46.4	69.0	86.2

Table 4. Comparisons in the segmentation effect.

Model	Backbone	mAP^s/%	AP₅₀^s/%	AP₇₅^s/%	AR^s/%	mAP_s^s/%	mAP_M^s/%	mAP_L^s/%
<i>Evaluation metric = segm</i>								
MS R-CNN (Huang et al., 2019)	ResNet50+FPN	58.4	84.7	63.5	64.1	40.7	64.0	86.9
PointRend	ResNet50+FPN	59.8	86.3	64.6	65.4	41.7	66.6	87.3
YOLACT (Bolya et al., 2019)	ResNet50	51.9	80.9	54.4	59.4	31.6	59.6	84.8
BoxSnake (Yang et al., 2023)	ResNet50+FPN	50.6	77.9	54.2	-	31.7	56.8	82.4
Mask R-CNN	ResNet50+FPN	58.6	84.7	63.3	63.5	40.3	65.2	87.3
FBSM(Ours)	ResNet50+FPN	60.7	85.3	65.0	66.1	43.3	68.0	89.5

As demonstrated in Tables 3 and 4, the FBSM segmentation model proposed in this paper outperforms other segmentation methods in terms of both detection and segmentation average precision for green apples, and the performance improvement is relatively obvious. The FBSM outperforms other models in almost all metrics significantly better than the YOLO family of models, exceeding YOLOF by 11.5 percentage points in detection precision and YOLACT by 8.8 percentage points in segmentation accuracy. Although the FBSM shows slightly lower AP_{50b} at IoU=0.5 compared to the FoveaBox detection model and PointRend segmentation model, and its mAP_{Lb} is inferior to the BoxSnake segmentation model, FBSM demonstrates strong competitiveness in terms of mAP. Additionally, FBSM topped other advanced models in terms of mAP, AP₇₅, AR, and the average precision for fruit with different scales. Based on these metrics, the FBSM model exhibits a significant overall advantage in the detection and segmentation of green target fruit.

For a more intuitive comparison of the detection and segmentation effects of the FBSM model with other mainstream or advanced models, representative images were selected to observe the visualization results graph. The green apple images under multiple environments such as obscured by branches and leaves, overlapping fruits, small target fruits in the distance and small target fruits at night are chosen for the experiments. These disturbed target fruits are more difficult to process and better illustrate the detection and segmentation performance of the model, and the visualization

results of the above images are shown in Fig 11.

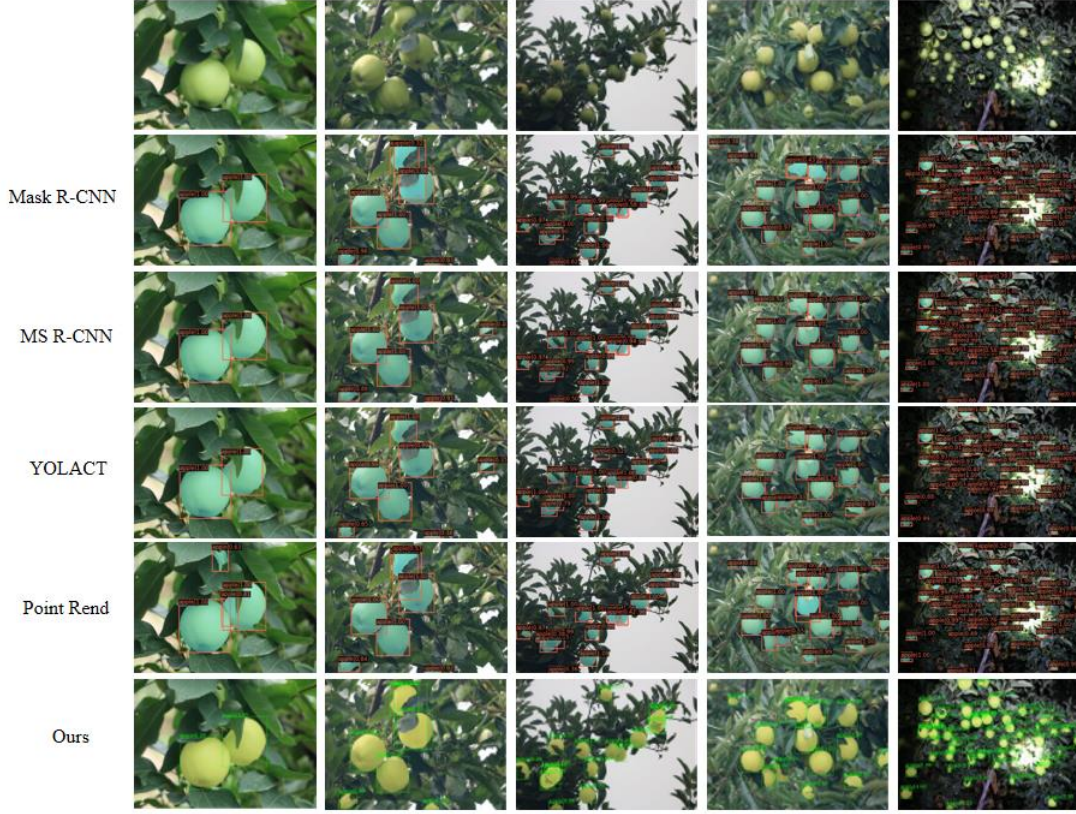


Fig. 11. Comparison of segmentation effects of different models.

As can be seen from Fig 11, compared with the classical R-CNN series models, the results show that Mask R-CNN and MS R-CNN both suffer from misdetection, especially for the heavily occluded target fruits. Mask R-CNN generates redundant detection frames in images with severe fruit occlusion, while MS R-CNN identifies non-target fruit regions as fruits and is not accurate enough to segment fruit edges that are occluded by branches and leaves. And then, observed the single-stage real-time instance segmentation model YOLACT, although YOLACT can achieve real-time segmentation, the detection of target fruits is rougher, which is easy to miss small target fruits in the distant view and is not sensitive to small targets. PointRend model achieves clear target mask edges at relatively low arithmetic power, but in high illumination environments, non-target fruit areas are easily misidentified as targets and produce detection frames, and overlapping target fruits tend to produce redundant detection frames, especially in nighttime environments.

The proposed FBSM achieves target fruit detection while completing accurate segmentation of target fruits. Based on the visualization results, FBSM demonstrates robust detection and segmentation capabilities even in the presence of various interfering factors, including branch and leaf shading, target overlap, small fruit targets, and low-light conditions during nighttime, thereby fulfilling the application requirements in natural orchard environments.

5 Conclusion

In this work, recognizing that the two-stage instance segmentation models often lose detailed boundary information during the feature extraction stage, we propose an innovative two-stage fruit segmentation model FBSM for green apples in the unstructured orchard environment. Specifically, a multi-stage mask prediction head is appended to the FoveaBox detection model to continuously

optimize the instance mask by combining the fine-grained features from the semantic head. Additionally, a bi-layer fusion structure is used to integrate the fine-grained features adequately, while the boundary recovery module is employed to recover the lost boundary detail information to obtain more accurate fruit instance masks. Compared with other advanced segmentation models, the performance of the proposed model has improved by 2.3, 0.9, 8.8, 10.1 and 2.1 points over MS R-CNN, PointRend, YOLACT, BoxSnake and Mask R-CNN, respectively. Given the great recognition accuracy of FBSM, the adopted multi-stage mask prediction head can be extended to other detection models to achieve two-stage instance segmentation for contour region refinement.

Additionally, despite the demonstrated segmentation performance of the model on the apple dataset, further improvements should focus on simplifying the algorithm to facilitate its integration into the vision system of fruit-picking robots for real-time work in the complex orchards.

Acknowledgments

This work is supported by National Nature Science Foundation of China (No.: 62372278); Young Innovation Team Program" of Shandong Provincial University (No.: 2022KJ250); New Twentieth Items of Universities in Jinan (2021GXRC049).

Declarations

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Arivazhagan S, Shebiah R N, Nidhyandhan S S, et al. Fruit recognition using color and texture features. *Journal of Emerging Trends in Computing and Information Sciences*, 2010, 1(2): 90-94.
- Asgari Taghanaki S, Abhishek K, Cohen J P, et al. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 2021, 54: 137-178.
- Bolya D, Zhou C, Xiao F, et al. Yolact: Real-time instance segmentation. *Proceedings of the IEEE/CVF international conference on computer vision*. 2019: 9157-9166.
- Che T, Zheng Y, Yang Y, et al. SDOF-GAN: Symmetric dense optical flow estimation with generative adversarial networks. *IEEE Transactions on Image Processing*, 2021, 30: 6036-6049.
- Chen C, Wang C, Liu B, et al. Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- Chen K, Pang J, Wang J, et al. Hybrid task cascade for instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019: 4974-4983.
- Chen Q, Wang Y, Yang T, et al. You only look one-level feature. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021: 13039-13048.
- Chen Z, Zheng Y, Gee J C. TransMatch: A transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. *IEEE Transactions on Medical Imaging*, 2024,43(1): 15-27.

- Cheng T, Wang X, Huang L, et al. Boundary-preserving mask r-cnn. European conference on computer vision. Springer, Cham, 2020: 660-676.
- Choi D, Lee W S, Schueller J K, et al. A performance comparison of RGB, NIR, and depth images in immature citrus detection using deep learning algorithms for yield prediction. ASABE Annual International Meeting, 2017, paper number: 1700076.
- Farjon G, Huijun L, Edan Y. Deep-learning-based counting methods, datasets, and applications in agriculture: a review. Precision Agriculture, 2023: 1-29.
- Girshick R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
- Hafeez A, Husain M A, Singh S P, et al. Implementation of drone technology for farm monitoring & pesticide spraying: A review. Information processing in Agriculture, 2022.
- He K, Gkioxari G, Dollár P, et al. Mask r-cnn. Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
- He L, Fang W, Zhao G, et al. Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods. Computers and Electronics in Agriculture, 2022, 195: 106812.
- Huang Y, Jing J, Wang Z. Fabric defect segmentation method based on deep learning. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-15.
- Huang Z, Huang L, Gong Y, et al. Mask scoring r-cnn. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6409-6418.
- Jha K, Doshi A, Patel P, et al. A comprehensive review on automation in agriculture using artificial intelligence. Artificial Intelligence in Agriculture, 2019, 2: 1-12.
- Jia W, Zhao D, Liu X, et al. Apple recognition based on K-means and GA-RBF-LMS neural network applied in harvesting robot. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(18): 175-183.
- Jia W, Tian Y, Luo R, et al. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. Computers and Electronics in Agriculture, 2020, 172: 105380.
- Jia W, Zhang Y, Lian J, et al. Apple harvesting robot under information technology: A review. International Journal of Advanced Robotic Systems, 2020, 17(3): 925310.
- Jia W, Zhang Z, Shao W, et al. RS-Net: robust segmentation of green overlapped apples. Precision Agriculture, 2022, 23(2): 492-513.
- Kamilaris A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey. Computers and electronics in agriculture, 2018, 147: 70-90.
- Kang H, Chen C. Fruit detection and segmentation for apple harvesting using visual sensor in orchards. Sensors, 2019, 19(20): 4599.
- Kirillov A, Wu Y, He K, et al. Pointrend: Image segmentation as rendering. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9799-9808.

- Klingenberg S, Fischer R, Zettler I, et al. Facilitating learning in immersive virtual reality: Segmentation, summarizing, both or none?. *Journal of Computer Assisted Learning*, 2023, 39(1): 218-230.
- Koirala A, Walsh K B, Wang Z, et al. Deep learning—Method overview and review of use for fruit detection and yield estimation. *Computers and electronics in agriculture*, 2019, 162: 219-234.
- Kong T, Sun F, Liu H, et al. Foveabox: Beyond anchor-based object detection. *IEEE Transactions on Image Processing*, 2020, 29: 7389-7398.
- Lian J, Jia W, Zareapoor M, et al. Deep learning based small surface defect detection via exaggerated local variation-based generative adversarial network. *IEEE Transactions on Industrial Informatics*, 2020, 16(2): 1343-1351.
- Liu M, Jia W, Wang Z, et al. An accurate detection and segmentation model of obscured green fruits. *Computers and Electronics in Agriculture*, 2022, 197: 106984.
- Luo Z, Yang W, Yuan Y, et al. Semantic segmentation of agricultural images: a survey. *Information Processing in Agriculture*, 2023.
- Maheswari P, Raja P, Apolo-Apolo O E, et al. Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques—a review. *Frontiers in Plant Science*, 2021, 12: 684328.
- Meshram A T, Vanalkar A V, Kalambe K B, et al. Pesticide spraying robot for precision agriculture: A categorical literature review and future trends. *Journal of Field Robotics*, 2022, 39(2): 153-171.
- Palacios F, Diago M P, Melo-Pinto P, et al. Early yield prediction in different grapevine varieties using computer vision and machine learning. *Precision Agriculture*, 2023, 24(2): 407-435.
- Saleem M H, Potgieter J, Arif K M. Automation in agriculture by machine and deep learning techniques: A review of recent developments. *Precision Agriculture*, 2021, 22: 2053-2091.
- Song Y, Glasbey C A, Horgan G W, et al. Automatic fruit recognition and counting from multiple images. *Biosystems Engineering*, 2014, 118: 203-215.
- Sun M, Zhao R, Yin X, et al. FBoT-Net: Focal bottleneck transformer network for small green apple detection. *Computers and Electronics in Agriculture*, 2023, 205: 107609.
- Tang Y, Zhou H, Wang H, et al. Fruit detection and positioning technology for a *Camellia oleifera* C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. *Expert systems with applications*, 2023, 211: 118573.
- Tian H, Wang T, Liu Y, et al. Computer vision technology in agricultural automation—A review. *Information Processing in Agriculture*, 2020, 7(1): 1-19.
- Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF international conference on computer vision*. 2019: 9627-9636.
- Wang Z, Jia W, Mou S, et al. KDC: A green apple segmentation method. *Spectroscopy and Spectral Analysis*, 2021, 41(9): 2980.
- Wang Z, Zhang Z, Lu Y, et al. SE-COTR: A novel fruit segmentation model for green apples application in complex

orchard. Plant Phenomics, 2022, ID: 0005.

Yang R, Song L, Ge Y, et al. Boxsnake: Polygonal instance segmentation with box supervision. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 766-776.

Yu L, Xiong J, Fang X, et al. A litchi fruit recognition method in a natural environment using RGB-D images. Biosystems Engineering, 2021, 204: 50-63.

Yu Y, Zhang K, Yang L, et al. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Computers and Electronics in Agriculture, 2019, 163: 104846.

Zhang C, Zou K, Pan Y. A method of apple image segmentation based on color-texture fusion feature and machine learning. Agronomy, 2020, 10(7): 972.

Zhang G, Lu X, Tan J, et al. Refinemask: Towards high-quality instance segmentation with fine-grained features. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 6861-6869.

Zhang Z, Jia W, Shao W, et al. Green apple detection based on optimized FCOS in orchards. Spectroscopy and Spectral Analysis, 2022, 42(2): 647-653.

Zheng Y, Sui X, Jiang Y, et al. SymReg-GAN: Symmetric image registration with generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5631-5646.

Zhou H, Wang X, Au W, et al. Intelligent robots for fruit harvesting: Recent developments and future challenges. Precision Agriculture, 2022, 23(5): 1856-1907.

Zhu C, He Y, Savvides M. Feature selective anchor-free module for single-shot object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 840-849.

Zhu N, Liu X, Liu Z, et al. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. International Journal of Agricultural and Biological Engineering, 2018, 11(4): 32-44.