Relative Position Detection of Clustered Tomatoes Based on BlendMask-BiFPN

Caiping Guo, Can Tang, Yehong Liu, Xin Wang*, Shumao Wang
Beijing Key Laboratory of Optimized Design for Modern Agricultural Equipment, College of Engineering, China Agricultural University (East Campus), Beijing, China
*guocaiping@cau.edu.cn; cantang@cau.edu.cn; liuyehong@cau.edu.cn; *wangxin117@cau.edu.cn; wangshumao@cau.edu.cn*

***Abstract*—In robotic harvesting, maneuvering around obstacles to position manipulators is challenging, especially in unstructured environments. This study proposes a method to detect the relative position of tomato bunches to the main stem position using the BlendMask-BiFPN algorithm. Initial comparative tests between full-stem and partial-stem labelling strategies revealed that the latter produced more complete peduncle masks, which guided our choice for subsequent experiments. Significant modifications to the BlendMask algorithm included the integration of a ResNet-101-BiFPN backbone, which improved the feature fusion network of the model. The revised model demonstrated high efficiency in pinpointing the relative positions of clustered tomatoes, achieving 91.3 % AR_{50}^{mask} and 84.8 % AP_{50}^{mask} for the detection of tomato bunches. Comparisons with Mask RCNN, YOLACT, YOLACT++, and YOLOv8 showed that the BlendMask-BiFPN model outperforms these alternatives, suggesting its potential for more effective robotic harvesting in complex agricultural scenarios.**

***Index Terms*—Automation; BlendMask-BiFPN; Deep learning; Neural networks; Robot; Tomato harvesting.**

I. INTRODUCTION

As automation technology for tomato cultivation in Chinese greenhouses continues to advance, total tomato production and unit yield will continue to increase in the future [1]. However, tomato harvesting at this stage is still high in labour costs, which is highly labour intensive. Therefore, it is inevitable to solve the problem of labour substitution in the picking process and develop the robot technology for picking tomatoes [2].

In robotic fruit picking, it is very important to correctly find and locate the fruit. Researchers have recently made good progress in this area. Li *et al.* [3] created a new network, MTA-YOLACT, that does three things at once for robots picking fruits: it finds groups of fruits and separates the main stems and peduncles. This network worked well on pictures of cherry tomato plants. Then, Rong, Hu, Hu, and Xu [4] created a model that can tell apart tomato fruits, main stems, and calyxes in difficult situations. This model uses the Swin Transformer V2 to perform the object detection task. Yan, Wang, Wang, Zhu, Zhou, and Yang [5] used a mix of Mask RCNN and a special way to separate parts according to depth to help find where to pick tomatoes. They were able to find the right depth information for fruit stems 87.3 % of the time. Zhang, Chen, Li, and Xu [6] used the YOLOv4 model to quickly find important parts in the bunches of tomatoes and the main stems that can be picked. They combined depth and colour information in their images to find picking points. Their system could recognise these points in 54 ms with a 93.83 % success rate.

Fruit picking robots are usually operating in complex unstructured natural environments. Branches, immature fruits, and other debris distributed around mature fruits become obstacles that hinder the movement of the robotic arm [7]. Furthermore, the size of the tomato bunches is different and the growth posture is changeable, so even if the peduncles of the clustered tomatoes can be detected successfully, the picking will fail due to the estimation of the the bad cutting pose of the robot arm, the obstruction of branches, and other problems, which will affect the success rate and efficiency of the picking. The success of a robot picking tomato bunches depends on its ability to accurately position its end effector to cut the peduncle. Therefore, it is crucial to acquire precise information on the location of both the main stem and the target fruit. These data enable the robotic arm to dynamically plan and update its trajectory, adapting to natural variability. Moreover, it is essential to ensure that the end effector avoids collisions with the tomato bunches as it approaches [8].

In addressing the challenge of robotic harvesting, integration of posture information with tomato location data has been shown to significantly enhance the robot harvest success rate. Zhang, Gao, Zhou, Zhang, Zou, and Yuan [9] developed a method to assess tomato posture, enabling the detection of tomato bunches and their orientation in complex settings. Their experiments achieved a keypoint detection success rate of 94.02 %. Du, Meng, Ma, Lu, and Cheng [10] introduced the YOLO-lmk model, an end-to-end solution for the detection of 3D poses of individual tomatoes within clusters. This innovative algorithm performs both bounding box and keypoint detection tasks for tomatoes in a single model and when combined with point cloud data, facilitates comprehensive 3D pose detection of tomatoes. Kim, Lee, Kim, and Kim [11] used a four-keypoint system to determine the posture of the tomato fruiting bodies. Their method offers

Manuscript received 26 March, 2024; accepted 2 July, 2024.

the advantage of maintaining consistent computational demands regardless of the number of objects present in the scene, providing multiple tomato-peduncle poses. Furthermore, Zhang *et al.* [12] created a specialised keypoint network (KPN) for tomatoes and introduced an innovative keypoint processing pipeline. This advancement not only improves keypoint localisation accuracy, but also effectively addresses the issue of inaccurate predictions stemming from poor quality source data.

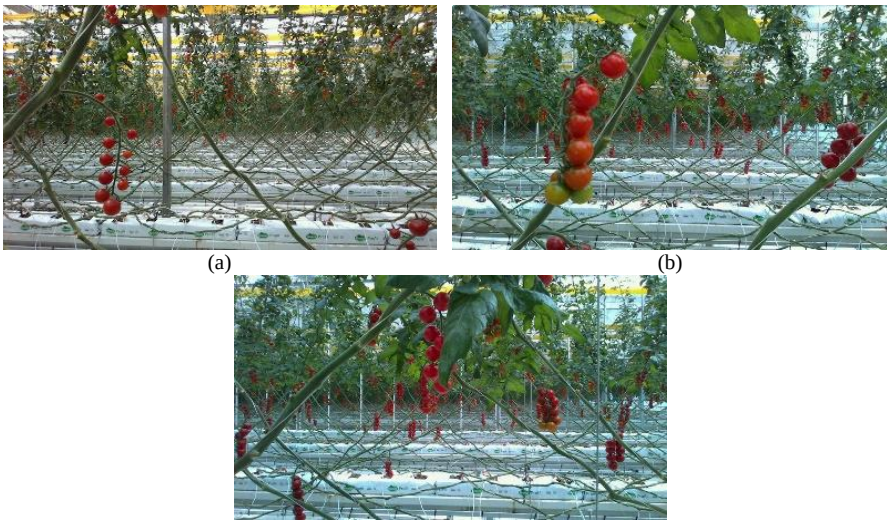
Current robotic fruit picking methods often overlook the crucial aspect of the relative positioning between the peduncle and the main stem. This oversight can lead to collisions between the robot and the tree trunk during harvesting. Therefore, it is important to develop a robust tomato bunch relative position detection algorithm for visual systems. To address this, a BlendMask-BiFPN algorithm was proposed for the relative position detection of tomato bunches. This algorithm identifies the relative positions of the tomato bunches by classifying them according to the position of their peduncles relative to the camera and the main stem. This classification not only guides the robotic picker but also supports the technical development to harvest tomatoes in various positions. It enables planning an efficient picking path based on the detected positions of the tomato clusters. As a result, this approach significantly reduces the likelihood of collisions between the fruit harvesting robot and the main stem, thus improving the success rate of picking and minimising damage to the tomatoes during the harvesting process. In summary, the main contributions of this paper are as follows.

1. The BlendMask-BiFPN model is proposed to identify the relative position of the tomato bunches with the main stem and the camera, improving the detection accuracy.
2. Extensive experiments have been performed on the clustered tomato data set. We show that the proposed model outperforms the original BlendMask and other instance segmentation algorithms in terms of accuracy and robustness.

II. MATERIALS AND METHODS

A. Data Acquisition

The tomato bunch data set used in this paper was collected in the Beijing Chaoyang Agricultural Garden from April 12 to 20, 2023. In this study, we captured images of tomato bunches using a RGB-D camera RealSense L515 that can obtain RGB-D images with a resolution of 1280×720 pixels. The RGB-D camera was photographed at a distance of 400 mm to 520 mm from the plants. To ensure sample diversity, the images include two weather conditions (cloudy and sunny) and three periods of the day (morning, noon, and afternoon). Therefore, the data set includes different lighting conditions, different shooting angles, and different poses of tomato bunches. Figure 1 shows some samples of the data set in different environments.



(a) (b) (c)

Fig. 1. Tomato bunch samples with different growing circumstances: (a) Backlight conditions; (b) Frontlight conditions; (c) Occlusion by leaves.

B. Labelling Strategy

The LabelMe annotation tool marks tomato bunches images in this study. Considering that this paper aims to identify the relative position of clustered tomatoes, the main stem, peduncles, and the top tomato in the tomato bunch are annotated as a sample. Based on the position information of the clustered tomatoes relative to the depth camera and the main stem, the samples are divided into eight categories for labelling. The schematic diagram of the relative position classification of the tomato bunches is shown in Fig. 2 and the specific category division rules are as follows: the samples located on the upper left of the picture and on the right side of the main stem are labelled as “r-upl”; the samples located on the upper right of the picture and on the right side of the main stem are labelled as “r-upr”; the samples located in the lower left of the picture, on the right side of the main stem are labelled as “r-lowl”; the samples located in the lower right of the picture, on the right side of the main stem are labelled as “r-lowr”; the samples located in the upper left of the picture, on the left side of the main stem are labelled as “l-upl”;the samples located in the upper right of the picture, on the left side of the main stem are labelled as “l-upr”; the samples located in the lower left of the picture, on the left side of the main stem are labelled as “l-lowl” and the samples located in the lower right of the picture, on the left side of the main stem are labelled as “l-lowr”. Two different strategies were used for the labelling. As shown in Fig. 3, the full-stem labelling method is to mark the entire main stem of tomato bunches in the image and the partial-stem labelling method is

to label the main stem centred on the midpoint of the connection between the peduncle and the main stem, and the marking length of the main stem is 30 mm.

Fig. 2. Schematic diagram of the relative position classification of tomato bunches.

(a) (b)

Fig. 3. Examples of different labelling strategies for tomato bunches: (a) The full-stem labelling method; (b) The partial-stem labelling method.

C. *Improvement of the BlendMask Network Architecture*

1. *Overall network structure*

BlendMask is an innovative one-stage instance segmentation network that uniquely integrates top-down and bottom-up methodologies [13]. Using a top-down approach, it produces dense instance masks via a sliding window, while its bottom-up method generates dense-pixel-embedded features, grouping them through sophisticated techniques. This network improves the extraction of low-level detailed features by incorporating a bottom module into the FCOS anchor-free detection framework [14]. BlendMask is inspired by the fusion techniques found in FCIS [15] and YOLACT [16], introducing a "blender module" to integrate high- and low-level features more effectively. The architecture of BlendMask encompasses a detection module to assign bounding boxes to each detected object and a mask branch. This branch adeptly combines low-level spatial data with high-level semantic information to create precise instance masks. For feature extraction, the detection module employs ResNet in conjunction with a feature pyramid network, feeding these extracted features into the FCOS detector for target identification. The mask branch comprises three key components: a bottom module for score map prediction, a top layer, for instance, attention forecasting, and the blender module, which seamlessly merges the scores and attentions. As a member of the dense-pixel prediction category, BlendMask transcends the limitations of resolution typically imposed by top-level sampling, resulting in superior mask quality.

2. *BiFPN backbone for feature fusion*

In recognition of the relative position of tomato bunches, the colour of the main stems and peduncles of clustered tomatoes is similar to that of the background, which can easily affect the recognition accuracy. Therefore, it is necessary to refine the features to guide the model to pay more attention to the targets and improve the recognition results of the model. Taking into account the above, BiFPN [17] is introduced into BlendMask. The network structures of BiFPN and BlendMask-BiFPN are shown in Figs. 4 and 5, respectively.

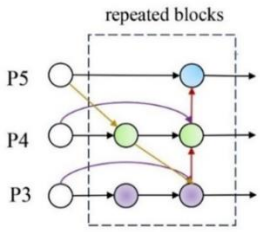


Fig. 4. Structure of the BiFPN network.

The BlendMask uses FPN for feature extraction and fusion, which only contains top-down feature fusion and is limited by the one-way transmission of information flow, resulting in lower recognition accuracy for the relative position of the tomato bunches. Unlike the FPN structure, BiFPN removes nodes with one single input based on PANet [18], improving the efficiency of the model. Furthermore, BiFPN uses weighted feature fusion to fuse input feature layers of different resolutions, and the weights corresponding to input layers of different resolutions are also different. By automatically learning the weight parameters of each input layer through the network, the overall feature information can be better represented. It helps identify and distinguish targets

with similar features, which is more in line with the requirements of this study. The BiFPN network uses the fast normalised fusion method to fuse weighted features, as shown in (1)

$$o = \sum_i \frac{\omega_i}{\varepsilon + \sum_j \omega_j} \times I_I, \quad (1)$$

where I_I represents the input feature map of the i^{th} level, ω_i and ω_j are the learnable weights, and the ReLU activation function is used to scale the learnable weights to [0, 1]. $\varepsilon = 0.0001$ is used to ensure numerical stability. o is the output feature.

3. *FCOS detector*

FCOS is a one-stage anchor free object detector that aims to solve object detection with a pixel-wise approach, similar to semantic segmentation [14]. FCOS uses the centre point of an object to define whether its position is positive and regresses four distances from the centre point to the object boundary. Instead of tiling multiple anchors for each location, FCOS tiles only one anchor point per pixel, reducing the number of design parameters that must be carefully adjusted. The hyperparameters associated with the anchor boxes severely affect the detection performance, so their elimination improves the generalisability of the model. Additionally, anchor-free detectors avoid the computations related to anchor boxes, such as IoU overlap and matching between anchors and ground-truth boxes. As a result, a relatively simple model is obtained, which allows faster training and inference times, as well as a lower memory requirement.

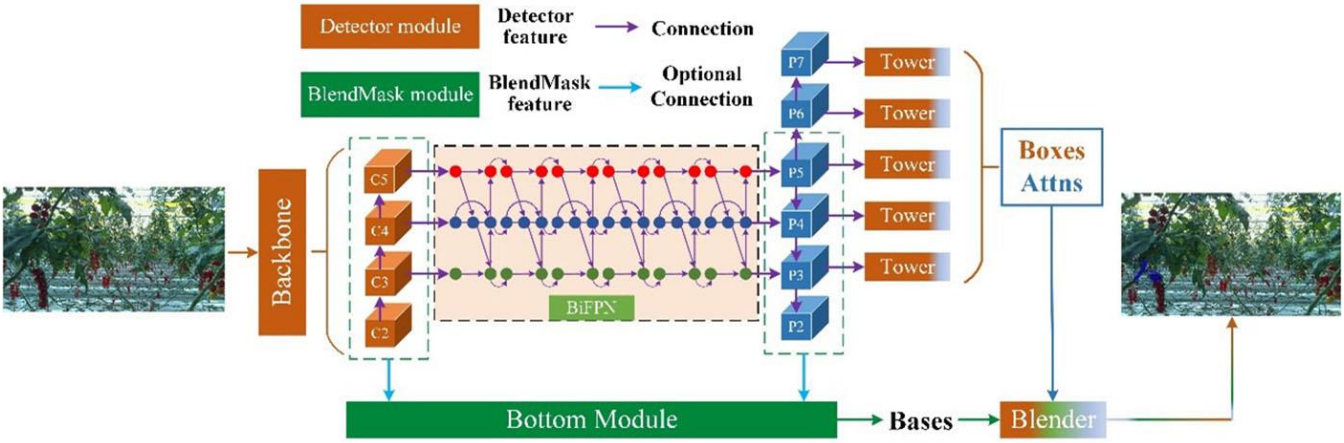


Fig. 5. Using BiFPN to improve BlendMask.

D. Evaluation Metrics

COCO metrics are used to evaluate the semantic segmentation performance of tomato bunches. Mask evaluation metrics include average precision (AP), AP_{50}, AP_{75}, AR, AR_{50}, and AR_{75}, as shown in Table I.

TABLE I. PERFORMANCE METRIC OF TOMATO BUNCHES INSTANCE SEGMENTATION.

Metric	Description
AP	AP is an average of 10 precision values on IoU = 0.5:0.05:0.95
AP_{50}	AP at IoU = 0.5
AP_{75}	AP at IoU = 0.75
AR	AR is an average of 10 recall values on IoU = 0.5:0.05:0.95
AR_{50}	AR at IoU = 0.5
AR_{75}	AR at IoU = 0.75

E. Experimental Setup

This experiment is based on the open source detection toolbox Detectron2 and AdelaiDet. The server platform is configured with Intel R Core TM i7-10700KF CPU @ 3.80 GHz processor, 32 GB RAM, and 10 GB NVIDIA GeForce PTX 3080 GPU. The software environments used in the experiment are Windows10, Pytorch, CUDA, and CUDNN. The learning momentum is 0.9, the batch size is 2, the learning rate is set to 0.0025, and the weight decay rate is 0.0001. BlendMask is trained using the warm-up learning rate strategy. Among them, the feature fusion network is FPN.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Comparison of Experimental Results of Different Annotation Strategies

To compare the impact of different annotation strategies on relative position detection of tomato bunches, this paper uses the BlendMask model to observe training of the clustered tomato data set using two labelling strategies with 20,000 iterations. When the number of iterations is 13,000 and 17,000, the learning rate is reduced by 10 times. The experimental and test results are shown in Table II and Fig. 6.

TABLE II. COMPARISON OF TWO LABELLING STRATEGY EXPERIMENTS.

Annotation strategy	Training set	AP_{50}^{box}	AP_{50}^{mask}
Full-stem labelling method	517	83.7 %	84.2 %
Partial-stem labelling method	517	54.2 %	53.1 %

It can be seen from Table II that the experimental data of the full-stem labelling method are better than those of the partial-stem labelling method. The reason may be that the proportion of target pixels in the image is small, resulting in poor detection results. Song *et al.* [19] showed that wires were segmented more difficultly than branches due to the smaller proportion of wire pixels. And similar conclusions could be drawn in [20]. In our study, the pixel ratio of the partial-stem labelling method was significantly lower than that of the full-

stem labelling method. Therefore, the full-stem labelling method can achieve better detection results. However, by observing the prediction results of the two labelling methods in Fig. 6, we found that the mask of peduncle is incomplete in the prediction results of the full-stem labelling method, which is because the proportion of peduncle to the target pixel in the full-stem labelling method is smaller than that of the partial-stem labelling method. Considering that the incomplete peduncle mask is unfavourable for the positioning of the picking point, we finally selected the partial-stem labelling tomato bunches data set and expanded it to 1272 images. Then we divided the data set into a training set, a validation set, and a test set in a ratio of 7:2:1 for all subsequent experiments.

(a) (b)

Fig. 6. Prediction results of two annotation methods: (a) Full-stem labelling method; (b) Partial-stem labelling method.

B. Comparative Experimental Results of BlendMask-BiFPN

To verify the effectiveness of the model improvement method proposed above, we used three ResNet-50-FPN, ResNet-101-FPN, and ResNet-101-BiFPN backbone networks to carry out ablation experiments based on the BlendMask model. The same training, validation, and test sets were used, and 50,000 iterations were set to train the network.

The training loss curves of different methods were convergent, as shown in Fig. 7, where different line colours represented trained models of different methods, respectively. At the initial stage of the loss curve, the loss values of three models rapidly decreased within approximately 50,000 iterations. We tested these trained models, and the results are shown in Table III.

We can get the following conclusions from Table III. Compared to BlendMask with ResNet-50-FPN, the AR^{mask}, AR^{mask}_{50}, AR^{mask}_{75}, AP^{mask}, AP^{mask}_{50}, and AP^{mask}_{75} values of BlendMask with ResNet-101-FPN increased by 4.0 %, 1.6 %, 7.2 %, 5.9 %, 6.2 %, and 9.5 %, respectively. The results show that a deeper network is beneficial to extract richer features. After replacing the BlendMask backbone network ResNet-101-FPN with ResNet-101-BiFPN, AP^{mask}_{50} increased from 58.5 % to 84.8 %, and AR^{mask}_{50} increased from 6.6 % to 91.3 %. The experimental results in Table III show that the improvement strategy is effective and can optimise the comprehensive performance of the model.

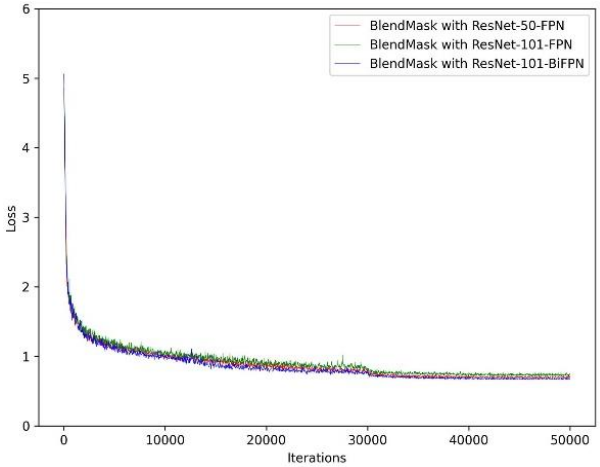


Fig. 7. Training loss curves of the different methods introduced into BlendMask.

To further verify the accuracy of the classification results obtained from the model, this experiment also recorded the AP^{mask} and AP^{mask}_{50} values of the relative positions of each cluster of tomatoes, as shown in Table IV. After replacing

ResNet-101-FPN with ResNet-50-FPN, the AP^{mask} and AP^{mask}_{50} values for each relative position of tomato bunches are higher than BlendMask with ResNet-50-FPN, and the AP^{mask}_{50} values of other categories have reached more than 40 % except for the l-lowl category. When BiFPN is introduced further, the AP^{mask}_{50} and AP^{mask} values are increased in all eight categories.

TABLE III. BACKBONE COMPARISON ON BLENDMASK FOR 5 K ITERATIONS.

Backbone	AR^{mask}	AR^{mask}_{50}	AR^{mask}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
ResNet-50-FPN	51.3 %	83.1 %	60.6 %	29.4 %	52.3 %	32.9 %
ResNet-101-FPN	55.3 %	84.7 %	67.8 %	35.3 %	58.5 %	42.4 %
ResNet-101-BiFPN	**58.9** %	**91.3** %	**73.2** %	**51.1** %	**84.8** %	**62.9** %

TABLE IV. INSTANCE SEGMENTATION RESULTS AT DIFFERENT BACKBONE.

Indicator	Backbone	l-lowl	l-lowr	l-upl	l-upr	r-lowl	r-lowr	r-upl	r-upr
AP^{mask}	ResNet-50-FPN	16.1 %	26.2 %	31.4 %	21.0 %	33.9 %	30.1 %	41.4 %	35.4 %
	ResNet-101-FPN	22.7 %	27.9 %	40.1 %	28.3 %	35.4 %	38.5 %	45.1 %	44.0 %
	ResNet-101-BiFPN	**48.7** %	**46.5** %	**57.1** %	**42.8** %	**48.8** %	**51.1** %	**59.0** %	**55.1** %
AP^{mask}_{50}	ResNet-50-FPN	28.1 %	48.4 %	50.3 %	43.2 %	61.8 %	56.4 %	71.3 %	58.8 %
	ResNet-101-FPN	38.1 %	47.5 %	65.5 %	54.3 %	59.2 %	61.8 %	72.5 %	68.9 %
	ResNet-101-BiFPN	**85.1** %	**80.6** %	**84.2** %	**76.8** %	**83.3** %	**88.7** %	**91.1** %	**88.4** %

Moreover, among these eight categories, r-upl has the highest AP^{mask}_{50} values of 91.1 %. Therefore, it is concluded that BiFPN can effectively improve the feature extraction ability of the network.

C. Comparison of Different Algorithms

To further verify the performance of the BlendMask-BiFPN instance method, this paper evaluates the method using the parameters AP^{box}, AP^{box}_{50}, AP^{mask}, and AP^{mask}_{50}, and compares with the existing main-stream instance segmentation algorithm. The specific results are shown in Table V.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT INSTANCE SEGMENTATION ALGORITHMS.

Method	AP^{box}	AP^{box}_{50}	AP^{mask}	AP^{mask}_{50}	Time (ms)
Mask RCNN	47.1 %	63.4 %	32.9 %	59.4 %	269
YOLACT	56.4 %	84.5 %	39.7 %	76.8 %	138
YOLACT++	57.2 %	84.7 %	40.5 %	79.3 %	141
YOLOv8	64.6 %	81.6 %	49.3 %	79.9 %	**79**
BlendMask-BiFPN	**70.0** %	**89.1** %	**51.1** %	**84.8** %	253

In Table V, the box indicators for each model are slightly higher than those of the mask, and our model achieves the best results in terms of AP^{box}, AP^{box}_{50}, AP^{mask}, and AP^{mask}_{50}. The AP^{mask}_{50} of BlendMask-BiFPN is 84.8 %, which is 25.4 %, 8.0 %, 5.5 %, and 4.9 % higher than Mask RCNN, YOLACT, YOLACT++, and YOLOv8, respectively. Among the five segmentation networks, Mask RCNN has the worst performance with AP^{mask}_{50} value of only 65.1 %. It can be seen from the comparison results that the BlendMask-BiFPN proposed in this study has the best performance to detect the relative position of the tomato bunches, but it remains challenging to achieve the trade-off between accuracy and speed. It should be noted that the detection time of BlendMask-BiFPN is 253 ms per image on average, which is not conducive to real-time detection of the relative position. The processing time of YOLOv8 is 79 ms, but unfortunately it only has a 79.9 % AP accuracy, which is lower than BlendMask-BiFPN.

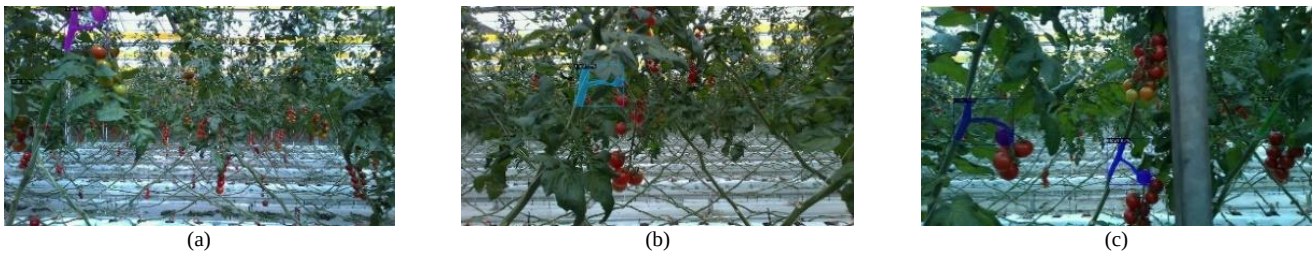
D. Qualitative Analysis

1. *Performance of the Proposed Model under Different Illumination Angles*

To evaluate the performance of the proposed model under different illumination angles, this study conducted experiments using 176 tomatoes with frontlight conditions and 181 with backlight conditions. In Table VI, 158 out of the 176 tomato bunches were correctly identified relative position under frontlight conditions, corresponding to 165 of 181 under backlight conditions. The relative position of the tomato bunches was misidentified at 5.7 % and 3.3 % under frontlight and backlight conditions, respectively. This means that some of the relative position categories of the tomato bunches are not recognised or that the background of the main stems, leaves, etc., is incorrectly identified as some kind of relative position due to its similar colour. The above results indicate that the proposed model is not affected by changes in lighting within the greenhouse environment. Some examples of the results are shown in Fig. 8.

TABLE VI. THE DETECTION RESULTS OF THE PROPOSED METHOD UNDER DIFFERENT ILLUMINATION ANGLES.

Illumination angles	Tomato Count	Correctly Identified		Falsely Identified		Missed	
		Amount	Rate (%)	Amount	Rate (%)	Amount	Rate (%)
Frontlight	176	158	89.8	10	5.7	8	4.5
Backlight	181	165	91.2	6	3.3	10	5.5



(a) (b) (c)

(d) (e) (f)

Fig. 8. Some examples of the detection results under different illumination angles: (a)–(c) frontlight conditions; (d)–(f) backlight conditions.

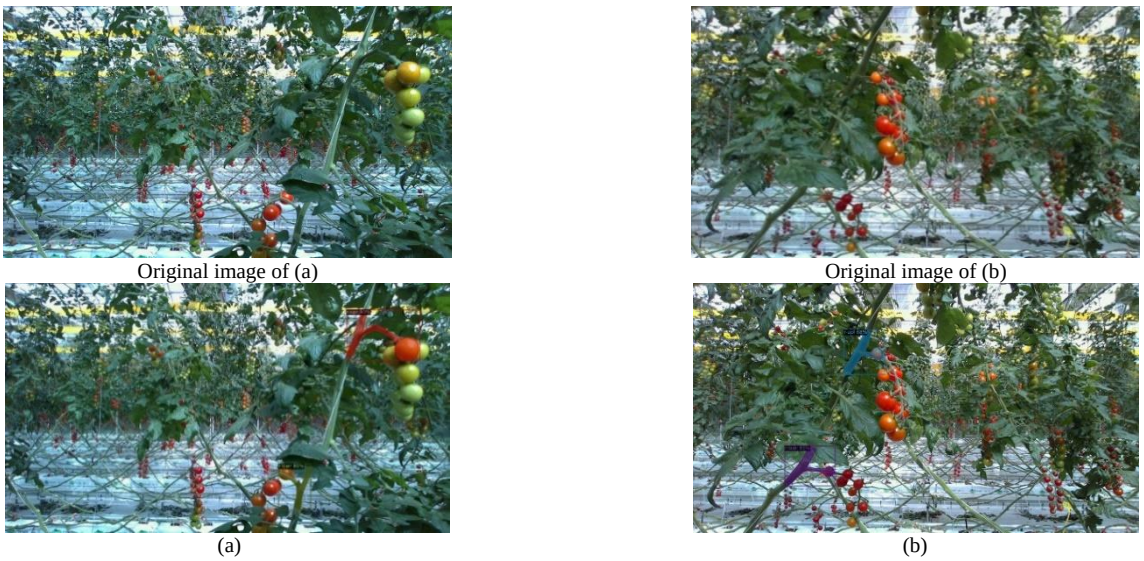
2. *Performance of the Proposed Model under Different Occlusion Condition*

The presence of obstacles such as leaves and main stems will affect the detection of the relative position of tomato bunches. To evaluate the performance of the proposed model under different occlusion conditions, it is divided into light and moderate occlusion according to the degree of occlusion of the marked area. Moderate cases are 30 %~60 % of the marked area blocked by leaves, other tomatoes. Cases with less than 30 % blockage were identified as slight.

The results are shown in Table VII, and the detection performance of the relative position of the tomato bunches under slight occlusion is slightly better than that under moderate occlusion. In the case of moderate occlusion, the presence of the labelled region is very different from the full labelled region, which explains some of the loss of semantic information. Besides, the proportion of unrecognised tomato bunches relative to the relative position of the main stem and camera is much higher than that of the incorrectly identified relative position. This may be due to the fact that the colour of the main stem and peduncle in the labelled area is similar to that of the leaves, making it easy to be unidentified by some occlusion. Figure 9 shows some examples of the detection results for both cases.

TABLE VII. THE DETECTION RESULTS OF THE PROPOSED METHOD UNDER DIFFERENT OCCLUSION CONDITIONS.

Occlusion conditions	Tomato Count	Correctly Identified		Falsely Identified		Missed	
		Amount	Rate (%)	Amount	Rate (%)	Amount	Amount
Slight	180	149	82.8	6	Slight	180	149
Moderate	122	89	73.0	6	Moderate	122	89



Original image of (a) Original image of (b)

(a) (b)

Fig. 9. Some examples of the detection results under different occlusion conditions: (a) the slight occlusion conditions; (b) the moderate occlusion conditions.

IV. DISCUSSION

Our method can provide pose classification for multiple tomato bunches in an image simultaneously. Most previous studies have focussed on estimating the peduncle pose of a single tomato cluster in the image. The inference time is longer because the process consists of a multistage framework. In tomato pose estimation, we use the method of instance segmentation for detection. Compared to the key point detection method, our algorithm can more accurately describe the shape of tomatoes, main stems, and peduncles, which is conducive to guiding the picking robot to avoid obstacles. In addition, point clouds obtained by commercial cameras are sparse and incomplete, often with zero and infinite values, which can lead to incorrect positioning of picking points, resulting in pick failure or manipulator damage. The detection accuracy of the instance segmentation method does not depend on accurate point cloud information, the efficiency is high, and the environmental influence is low. Section III showed that only 890 images were needed as a training set and the accuracy of detecting the relative position of tomato bunches reached 84.8 %. Compared to other instance segmentation data sets with tens of thousands of images, such as the MSCOCO data set, 890 images are a small training set. However, it can satisfy BlendMask-BiFPN training.

In conclusion, although we have explored a method for relative position detection of tomato bunches, the proposed

algorithm is still in the preliminary stage. The algorithm proposed in this study did not correctly predict the relative position of the tomato bunches when the occluded area exceeded 30 %, and only 73.0 % of the relative position of the clustered tomatoes could be identified when the occlusion rate was 30 % to 60 %. Additionally, concerning the generality of the model, the model trained on this data set can be used in other greenhouses, but with reduced accuracy. The model is only applicable in greenhouses that have tomato varieties and environments similar to the training set.

V. CONCLUSIONS

Aiming at the problem that tomato bunch picking robots are prone to collision in unstructured environments, this study proposed a relative position detection method for clustered tomatoes based on BlendMask. To compare the impact of the two annotation strategies on the recognition results, this paper used the BlendMask model to train the clustered tomato data set with the two annotation strategies. The results showed that the recognition performance of the full-stem labelling method is better than that of the partial-stem labelling method, but the mask of the peduncle in the test result images with the full-stem labelling method is incomplete, which is not conducive to the later tomato bunch picking. Therefore, the data set labelled with the partial-stem labelling method was selected for subsequent experiments.

To improve the feature fusion capability, the BlendMask-BiFPN instance method is proposed to segment tomato bunches with an eight-class relative position of tomato bunches by pixel-wise annotations. The verification results showed that the BlendMask-BiFPN model achieves significant performance in the self-built tomato bunches plant image data set. The values of AR_{50}^{mask} and AP_{50}^{mask} were 91.3 % and 84.8 %, respectively. The detection results of the BlendMask-BiFPN algorithm proposed in the study were compared with the other four algorithms, and the results showed that the AP_{50}^{mask} of the BlendMask-BiFPN model was improved by 25.4 %, 8.0 %, 5.5 %, and 4.9 %, respectively, when compared to Mask RCNN, YOLACT, YOLACT++, and YOLOv8. This indicates that the improved strategy in this paper can significantly improve the accuracy of detecting the relative position of clustered tomatoes. It can also detect the relative position of tomato bunches at different illumination angles and under slight occlusion with strong robustness. The limitation of the proposed method is that the generalisation of the model is general and only suitable for tomato bunches varieties similar to the data set and environments similar to the greenhouse environment of the data set. Future work will focus on addressing these limitations by expanding the diversity of the data set to improve the detection accuracy.

NOMENCLATURE

BiFPN	Bidirectional Feature Pyramid Network
FCIS	Fully Convolutional Instance-aware Semantic Segmentation
FCOS	Fully Convolutional One-Stage Object Detection
FPN	Feature Pyramid Network
IoU	Intersection over Union
MASK RCNN	Mask Region Convolutional Neural Networks
PANet	Path Aggregation Network
ReLU	Rectified Linear Unit
ResNet	Residual Neural Network
RGB-D	An image data type (RGB stands for red, green, and blue colour channels; D stands for depth information)
YOLACT	You Only Look At Crops Transforms
YOLACT++	A version that has been improved from the original YOLACT model
YOLO	You Only Look Once
YOLOv8	The eighth-generation version of the YOLO family of object detection models

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] J. Rong, H. Zhou, F. Zhang, T. Yuan, and P. Wang, "Tomato cluster detection and counting using improved YOLOv5 based on RGB-D fusion", *Computers and Electronics in Agriculture*, vol. 207, art. ID 107741, 2023. DOI: 10.1016/j.compag.2023.107741.

[2] Z. Jin, W. Sun, J. Zhang, C. Shen, H. Zhang, and S. Han, "Intelligent tomato picking robot system based on multimodal depth feature analysis method", *IOP Conf. Series: Earth and Environmental Science*, vol. 440, pp. 042074-1–042074-5. DOI: 10.1088/1755-1315/440/4/042074.

[3] Y. Li *et al.*, "MTA-YOLACT: Multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting", *European Journal of Agronomy*, vol. 146, art. 126812, 2023. DOI: 10.1016/j.eja.2023.126812.

[4] Q. Rong, C. Hu, X. Hu, and M. Xu, "Picking point recognition for ripe tomatoes using semantic segmentation and morphological processing", *Computers and Electronics in Agriculture*, vol. 210, art. 107923, 2023. DOI: 10.1016/j.compag.2023.107923.

[5] J. Yan, P. Wang, T. Wang, G. Zhu, X. Zhou, and Z. Yang, "Identification and localization of optimal picking point for truss tomato based on Mask R-CNN and depth threshold segmentation", in *Proc. of 2021 IEEE 11th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, 2021, pp. 899–903. DOI: 10.1109/CYBER53097.2021.9588274.

[6] Q. Zhang, J. Chen, B. Li, and C. Xu, "Method for recognizing and locating tomato cluster picking points based on RGB-D information fusion and target detection", *Transactions of the Chinese Society of Agricultural Engineering*, vol. 37, no. 18, pp. 143–152, 2021. DOI: 10.11975/j.issn.1002-6819.2021.18.017.

[7] H. Yang, L. Li, and Z. Gao, "Obstacle avoidance path planning of hybrid harvesting manipulator based on joint configuration space", *Transactions of the Chinese Society of Agricultural Engineering*, vol. 33, no. 4, pp. 55–62, 2017. DOI: 10.11975/j.issn.1002-6819.2017.04.008.

[8] T. Kalampokas, E. Vrochidou, G. A. Papakostas, T. Pachidis, and V. G. Kaburlasos, "Grape stem detection using regression convolutional neural networks", *Computers and Electronics in Agriculture*, vol. 186, art. 106220, 2021. DOI: 10.1016/j.compag.2021.106220.

[9] F. Zhang, J. Gao, H. Zhou, J. Zhang, K. Zou, and T. Yuan, "Three-dimensional pose detection method based on keypoints detection network for tomato bunch", *Computers and Electronics in Agriculture*, vol. 195, art. 106824, 2022. DOI: 10.1016/j.compag.2022.106824.

[10] X. Du, Z. Meng, Z. Ma, W. Lu, and H. Cheng, "Tomato 3D pose detection algorithm based on keypoint detection and point cloud processing", *Computers and Electronics in Agriculture*, vol. 212, art. 108056, 2023. DOI: 10.1016/j.compag.2023.108056.

[11] T. Kim, D.-H. Lee, K.-C. Kim, and Y.-J. Kim, "2D pose estimation of multiple tomato fruit-bearing systems for robotic harvesting", *Computers and Electronics in Agriculture*, vol. 211, art. 108004, 2023. DOI: 10.1016/j.compag.2023.108004.

[12] F. Zhang *et al*., "TPMv2: An end-to-end tomato pose method based on 3D key points detection", *Computers and Electronics in Agriculture*,

vol. 210, art. 107878, 2023. DOI: 10.1016/j.compag.2023.107878.
[13] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "BlendMask: Top-down meets bottom-up for instance segmentation", in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8570–8578. DOI: 10.1109/CVPR42600.2020.00860.
[14] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection", in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9626–9635. DOI: 10.1109/ICCV.2019.00972.
[15] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation", in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4438–4446. DOI: 10.1109/CVPR.2017.472.
[16] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation", in *Proc of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9156–9165. DOI: 10.1109/ICCV.2019.00925.
[17] Y. Zhang *et al.*, "Development of a cross-scale weighted feature fusion network for hot-rolled steel surface defect detection", *Engineering Applications of Artificial Intelligence*, vol. 117, part A, art. 105628, 2023. DOI: 10.1016/j.engappai.2022.105628.
[18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation", in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768. DOI: 10.1109/CVPR.2018.00913.
[19] Z. Song *et al.*, "Canopy segmentation and wire reconstruction for kiwifruit robotic harvesting", *Computers and Electronics in Agriculture*, vol. 181, art. 105933, 2021. DOI: 10.1016/j.compag.2020.105933.
[20] X. Zhang, M. Karkee, Q. Zhang, and M. D. Whiting, "Computer vision-based tree trunk and branch identification and shaking points detection in Dens-Foliage canopy for automated harvesting of apples", *Journal of Field Robotics*, vol. 38, no. 3, pp. 476–493, 2021. DOI: 10.1002/rob.21998.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) license (http://creativecommons.org/licenses/by/4.0/).