Proceedings of 14th IEEE International Conference
on CYBER Technology in Automation, Control, and Intelligent Systems
July 16 - 19, 2024, Copenhagen, Denmark

2024 IEEE 14th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER) | 979-8-3315-0605-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CYBER63482.2024.10749520

# Approach for Identification of Tomato Orientation and Picking Points

Huan Hu[1,2], Ying Yan[1], Junning Zhang[1*], ZhenHao Han[2,3]

*(1. Beijing Information Science and Technology University, Beijing 100192, China*

*2. State Key Laboratory of Agricultural Equipment Technology, Beijing 100083, China*

*3. Chinese Academy of Agricultural Mechanization Sciences Group Co., Ltd., Beijing 100083, China)*

*Abstract*— **The varying direction of growth of tomato plants in the field poses a huge challenge for robot harvesting. A two-stage positioning strategy has been proposed to address this issue. First, the Sim-YOLO-Pose algorithm based on key points is used to determine the direction of growth of tomatoes. It integrates the SimAM attention mechanism with the backbone network of YOLO, focusing more attention on the target fruit and solving the difficulty in recognizing it. Then, an instance segmentation-based approach called YOLACT is adopted to separate the tomato stems, further analyzing the picking points. Both the GAN network and traditional data augmentation algorithms are used during data preprocessing. After obtaining the positions of the key points, the 3D coordinates are estimated by combining the depth values acquired from the RealSense D435i camera, which are used to calculate the picking point combined the spatial position with orientation information. To validate the accuracy, our proposed approach was compared with several state-of-the-art methods including the traditional Mask R-CNN model, the Faster R-CNN model and the YOLO family models. Finally, the Sim-YOLO-Pose model was deployed on the edge device Jetson Xavier using TensorRT. The results showed that the accuracy rate and the mAP of the Sim-YOLO-Pose model was 94.7%, 97.2%, respectively. This work will provide theoretical references for target detection based on embedded edge computing in an unstructured environment.**

***Index Terms*—Tomato recognition and localization, GAN network, Pose estimation, Embedded terminal deployment**

## I. Introduction

Tomatoes are not only delicious and nutritious but also hold a pivotal position in global cuisine, agriculture, and the economy. Their importance cannot be overstated, and they continue to be a vital part of our daily diet. To mitigate the physical strain and labor costs associated with manual picking, tomato picking robots hold significant promise for enhancing agricultural production, reducing costs, and improving worker safety. Accurate recognition of tomato fruits and picking points is the technical prerequisite for non-destructive picking by robots[1]. However, there are difficulties in the identification and positioning of tomatoes due to unstructured growth, complex stem posture and irregular picking environment, which poses a great challenge for accurate recognition of target objects in color images[2-3].

In recent years, domestic and foreign scholars have fully studied the recognition and localization of fruits. Instance segmentation is the process of detecting pixel-wise masks of objects in images[4]. Wang et al.[5] argued that two dimensional instance segmentation allowed for 3D point cloud instance segmentation to be conducted from RGB-D images instead of more expensive lidar sensors, which was verified according to the PointSeg model composed of YOLACT++ and point cloud instance segmentation. The experiments showed that it has not only good real-time performance but also better instance segmentation accuracy . Liu et al. [6] proposed an efficient fruit detection algorithm framework OrangePointSeg. Experimental results showed that the detection speed of the improved algorithm is 44.63 frames/s and the average accuracy is 31.15%. Li et al.[7] proposed a novel method of tomatoes segmentation based on RGB-D depth images and K-means optimized SOM(Self-organizing map) neural network. The correct recognition rate was 87.2%.

Subsequently, how the robot performs the picking task is an important issue to consider. Knowing the stem orientation is important for orienting a snipper-like end-effector for an optimal cut. There are several methods existing for determining the orientation of objects within a scene, some of which are based on deep learning[8-9]. Zhang et al.[10] proposed a method of visual localization and picking posture estimation of string tomato based on instance segmentation. Field test verified that the average recognition rate of picking point was 98.07% and the picking rate can achieve 98.15%. Liu et al. [11] established the improved YOLO v8-pose model added Slim-neck module and CBAM attention mechanism module to identify red ripe strawberries and detect the key

Research supported by Opening Fund of National Key Laboratory of Agricultural Equipment Technology [Grant number NKL-2023-004], Key Research and Cultivation Projects of Beijing Universities under Project [Grant number 2021YJPY201] and National Key Research and Development Program "Intergovernmental International Science and Technology Innovation Cooperation" Special Project [Grant number SQ2023YFE0101940].

* Corresponding author: Junning Zhang, Professor.
E-mail: zhangjunning@bistu.edu.cn

979-8-3315-0605-6/24/$31.00 ©2024 IEEE

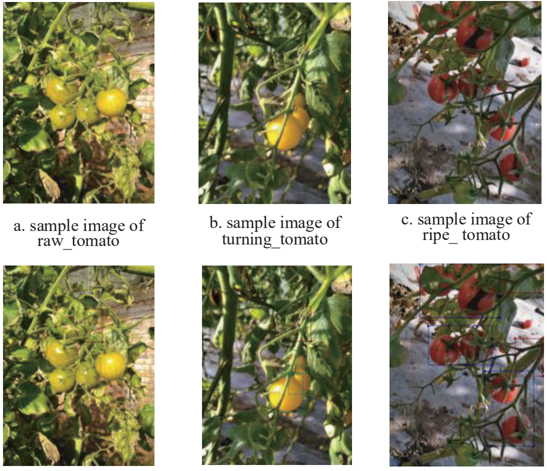
points of the stem. The mAP-kp of the improved YOLO v8-Pose can achieve 97.91%.

Most of the current research is focused on scenarios where tomatoes grow in a relatively regular pattern. However, the diverse growth postures of tomatoes in the field pose significant challenges for picking and localization. To address this issue, this paper proposes a two-stage picking strategy. The first step is to capture images from a farther distance. The posture of the detected tomatoes is estimated using YOLO-pose. The attention mechanism called SimAM[12] is added to improve the YOLO algorithm. Then, based on the estimated posture, the camera is moved to the vicinity of the target tomato. The more accurate posture is estimated and the fruit stem is segmented by YOLACT[13] algorithm. Finally, based on the identified picking point and posture, the robot is guided to pick the tomato. Meanwhile, the proposed algorithm is deployed on Jetson Xavier NX for verification, which can provide research support for the robotic operation of the facility environment based on the embedded edge computing.

## II. Tomato Image Acquisition and Processing

### A. Preparation of Tomato Dataset

1177 fruit images of cherry tomato fushan88 were collected from Datong Agricultural Hi-Tech Park in Shanxi Province in December 2022. Considering the universality of the detection algorithm, tomatoes with different levels of maturity were captured for training the model.


a. sample image of raw_tomato

b. sample image of turning_tomato

c. sample image of ripe_ tomato

d. annotation of raw_tomato

e. annotation of turning_tomato

f. annotation of ripe_tomato

Figure 1. Labeling tomatoes with different maturity and stems

Different maturity levels of tomatoes are classified into three categories: raw_tomatoes, turning_tomatoes and ripe_tomatoes, as shown in the Fig.1(a.b.c). The corresponding labeling results are shown in Fig.1(d.e.f). The Labelme tool was used to manually label tomatoes and stems among images. The VOC(Visual Object Classes) data format is used for annotation, generating XML files that contain detailed information about the images and the annotated objects within them[14]. Then these XML files are converted to the YOLO txt format[15]. The tomato is annotated by using the minimum bounding rectangle of the target. The area approximately 10cm above the tomato is labelled as the fruit stem.

### B. Data Augmentation and Preprocessing

Image preprocessing is used to constantly increase the layers during the training process, so that the model can quickly and accurately identify the target fruit and fruit stem[16-17].

In order to improve the model generalization ability, this paper combines GAN(Generative Adversarial Network) and traditional image enhancement methods to expand the dataset. The process of using GAN networks to expand training data is illustrated in Figure 2.

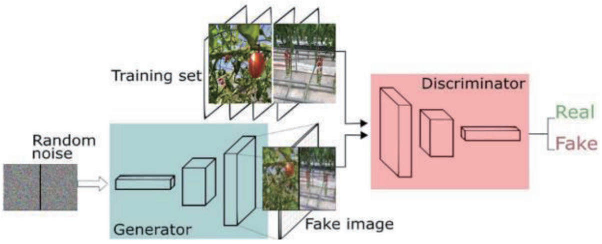


Figure 2. Dataset expansion using GAN networks

With the help of the idea provided by a “binary minimax game” problem, the GAN model is used. During the training, fix one party, like D first, then update the parameters of another model, like G, and iterate alternately to maximize the errors of the other party. Finally, G can estimate the distribution of the data samples to generate new data by adjusting the network structure[18]. In this study, the initial learning rate is set to 0.0002 and the number of iterations is 200 times.

Meanwhile, several traditional methods are used to enrich the background of the detected objects and enhance the diversity of the image data, which include mosaic data enhancement method, rotating the image 90 degree clockwise/counterclockwise and Hue method. The data augmentation algorithm proposed by GAN network is combined with these traditional algorithms, which uses the source view to adjust the Generator[19]. As a result, the dataset has been expanded to 3486 images, with 70% of the sample size used for the training set, 20% for the validation set, and 10% for the test set.

## III. Tomato Pose Estimation and Picking Point Localization

### A. Detecting Keypoints of Tomatoes Using YOLO-pose

As a heat-mapless joint detection method, YOLO-Pose groups the detected key points into a skeleton, which does not need the post-processing of the bottom-up method. Figure 3 describes its network structure. Tomato images are generally elliptical in shape, which can be described by the vertices of the major and minor axes. Given that the calyx at the base of the tomato is relatively easy to observe, we can replace one of the points with the two intersections between the calyx and the surface of the tomato. In summary, five points can be used to describe a tomato, which are used instead of the original 17 points in yolov7-pose algorithm, thereby reducing the total number of columns to 20. These columns include 1 column for representative class, 4 columns representing the bounding box position and the remaining 15 columns indicating the information of key points. Each point is represented by 3 values, with the first two representing the position and the last one indicating whether the point is visible.

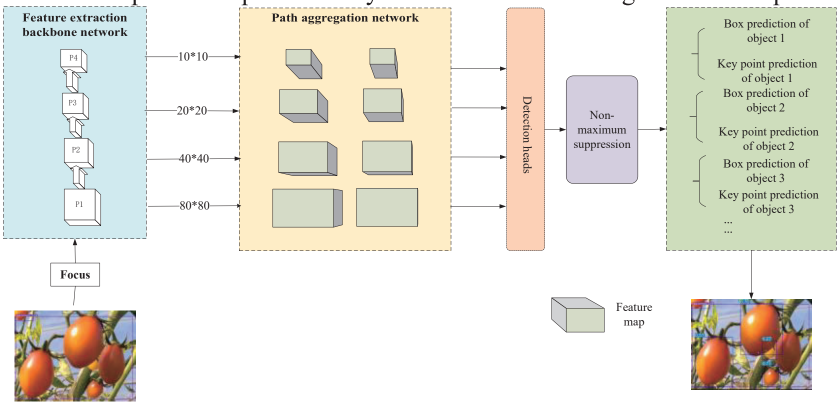

Feature extraction backbone network
P4
P3
P2
P1
Focus
10*10
20*20
40*40
80*80
Path aggregation network
Detection heads
Non-maximum suppression
Box prediction of object 1
Key point prediction of object 1
Box prediction of object 2
Key point prediction of object 2
Box prediction of object 3
Key point prediction of object 3
...
...
Feature map

Figure 3. Keypoints detection based on YOLO-Pose

## B. Introducing SimAM to Improve the Model

In order to improve the interference suppression ability of the model, a 3D non-parametric SimAM attention mechanism is introduced in the feature pyramid network, which can evaluate the feature weights more comprehensively and efficiently compared with the parametric attention mechanism to enhance tomato features and attenuate background interference. Based on the theory of visual neural science, SimAM performs more significantly than its neighboring neurons, producing spatial inhibition towards those neighboring neurons. When processing relative tasks concerned with vision, these neurons with more information should be assigned higher weights. During the detection task of the tomato cluster, these neurons are mainly responsible for extracting key features of tomato targets.

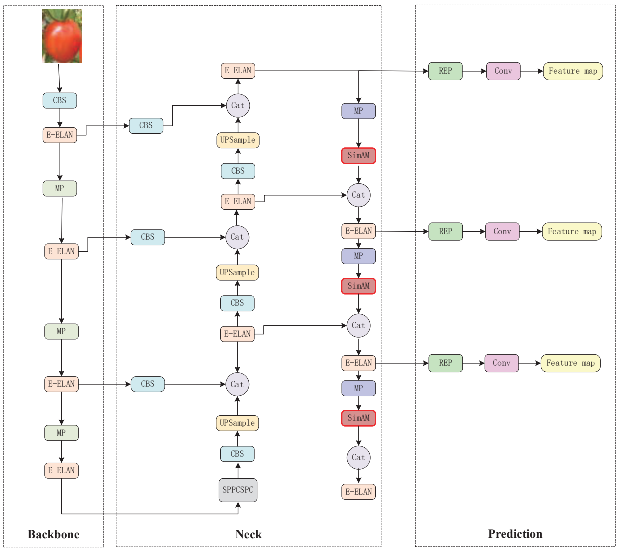


Figure 4. Network architecture after integrating SimAM

With the key target features highlighted and irrelevant information weakened, the SimAM module evaluates the importance of each independent neuron without introducing additional learnable parameters while preserving computational efficiency based on the energy function in neuroscience theory[19]. As depicted in Figure 4, the SimAM attention mechanism is presented in the feature pyramid structure to optimize the features of the backbone network and reduce the interference of background detection in an end-to-end manner.

The improved model is able to preserve more useful features during residual fusion and reduce feature loss, facilitating subsequent localization and classification.

## C. Stem Recognition Algorithm Based on YOLACT

YOLACT is a classic one-stage instance segmentation method[20] and faster than a two-stage instance split network[21] . It divides the instance segmentation task into two parallel processes, firstly using the prototype mask branch to generate a prototype mask, then predicting the mask coefficient of each instance and generating the instance mask.

Since the detection test of tomatoes are carried out in a wide-area context, which may cause interference towards stem resulting from the exist of background, it's vital to establish a dataset including images clipped in the range of 2-3cm. Using the completed fruit labeling boundary box in the target detection dataset, the length and width of the boundary are enlarged by 10%, so that the key information around the target such as the stem, calyx and other key information conducive to judgement is included in the box, as shown in Figure 5.

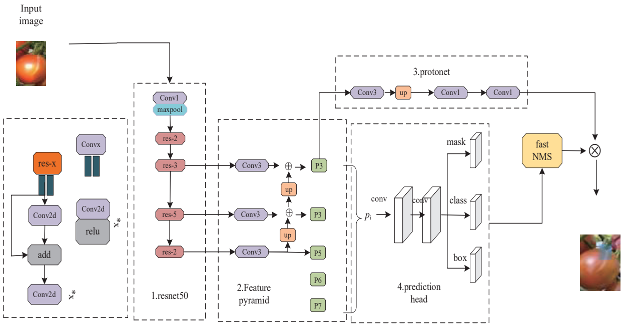


Figure 5. Network structure of YOLACT

## D. Pose Estimation of Tomato Fruit

Keypoints are used to estimate the orientation when tomatoes are seen up closely. The orientation is defined as the direction from the bottom of tomato to the calyx of its flower and a triangle is formed to indicate the final direction of the fruit as shown in Figure 6.

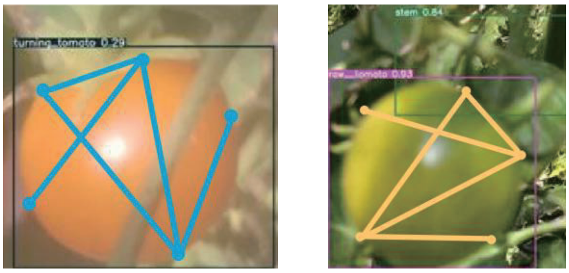


Figure 6. Result of detected keypoints

The spatial coordinates ($x_c$,$y_c$,$z_c$) of keypoints are calculated based on the camera projection model shown as follows:

$$\begin{bmatrix} \mathrm{u} \\ v \\ 1 \end{bmatrix} = \frac{1}{z_c} \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \qquad (1)$$

Where ($c_x$, $c_y$) represents the pixel coordinates of the image center, (u, v) represents the pixel-based location of the detected keypoint. The $f_x$ and $f_y$ represent the normalized focal lengths on the x-axis and y-axis, respectively.

Our system utilizes the RealSense D435i depth camera, so the depth $z_c$ can be obtained through the camera's API function. And then $x_c$ and $y_c$ are derived as follows:

$$\begin{cases} \mathrm{x}_c = \dfrac{z_c u - c_x z_c}{f_x} \\ y_c = \dfrac{z_c v - c_y z_c}{f_y} \end{cases} \qquad (2)$$

The estimation result of a single tomato pose was shown in Figure 7. The two vertices of the base of the triangle are defined as $P_1$ and $P_2$. $P_3$ is the midpoint between $P_1$ and $P_2$. It indicates the connection point between the tomato and its stem. The vector $\overrightarrow{q_1}$ is used to represent the growth posture of tomato, which is formed by the vertex of the triangle and the middle point of the bottom.

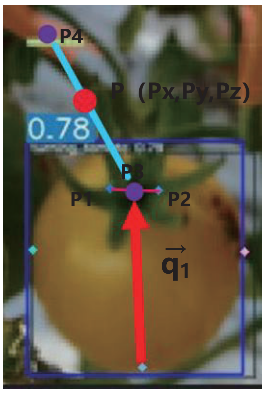


Figure 7. Estimation of single tomato pose

## E. Picking Point Localization

A two-step detection strategy is employed as presented in Figure 8. Firstly, the tomato plants are captured from a farther distance, thus the 3D position and posture of the detected tomatoes are estimated. Subsequently, the camera is moved closer to one group of tomatoes, thus keypoints are detected to further determine the harvesting posture. The stem is segmented using YOLACT algorithm.

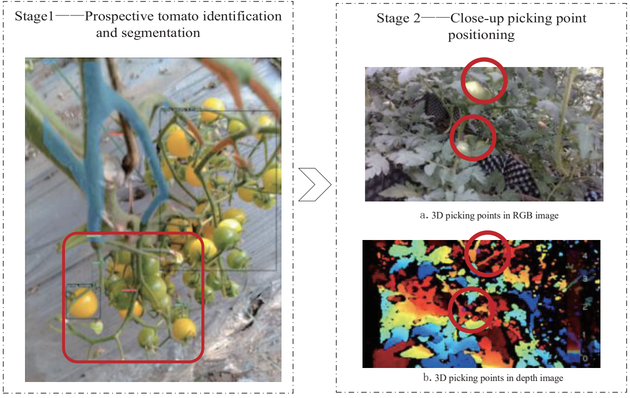

Stage1——Prospective tomato identification and segmentation
Stage 2——Close-up picking point positioning
a. 3D picking points in RGB image
b. 3D picking points in depth image

Figure 8. Flow diagram of tomato determination of 3D picking points

As Figure 7 shows, the centroid point of the fruit stem is further defined as $P_4$. The picking point P is defined as the midpoint between $P_3$ and $P_4$. The depth value of P (Pz) is obtained from the depth map provided by the RGB-D camera. Finally, the three-dimensional spatial coordinates (Px, Py) of the cut-off point are calculated according to Equation 2. Additionally, the color information and depth information in the RGB-D image of fruit stem were fused to calculate the extract depth value $P_z$ of the picking point.

# IV. Test Results Analysis and Deployment

## A. Construction of the Test Platform

The software environment of the experiment platform is based on Windows 10 version, Python 3.7, Tensorflow 2.11.0, and OpenCV_Python to build a framework for deep learning. The hardware environment of the experiment (GPU) is an Intel(R) Xeon(R) CPU E5-2620 v4@2.10GHz, and two NVIDIA TITAN V graphics cards with 24 GB. Resnet101 is the backbone network. Parameters about the model are illustrated in Table I below.

TABLE I. IPARAMETER VALUES OF TEST MODEL

| Attributes | Values |
|---|---|
| **Image_Shape** | [640 640 3] |
| **RPN_Anchor_Scales** | (96, 192, 384, 768, 1536) |
| **Num_classes** | 4 |
| **Number of iterations** | 200 |
| **Optimizer** | SGD |
| **Batch size** | 64 |

## B. Evaluation Indicators

This study adapts precision(P), Recall rate(R), and mean Average Precision(mAP) as evaluation indices of target recognition for tomato and fruit stem target recognition. The indicators are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$mAP = \frac{1}{C}\sum_{k=i}^{N} P(k)\Delta R(k) \tag{5}$$

TP represents the positive sample number predicted as positive by the model, FN represents the number of positive samples predicted as negative by the model, and FP represents the negative sample number predicted as positive. C is the number of categories. N is the number of reference thresholds, and k is the threshold. P(k) is the accuracy rate, and R(k) is the recall rate.

### C. *Comparison of Training Results and Analysis*

In order to verify the effectiveness of Sim-YOLO-Pose compared with K-means algorithm in YOLOv2 and YOLOv3, a comparison diagram of experimental results was conducted, as shown in Figure 9. It can be seen that YOLOv2 and YOLOv3 both accurately generated three tomato anchor frames of different maturity stages. Compared with the above two clustering methods, YOLOv5 uses genetic algorithm to evaluate the fitness of anchor frames. The comparison between Faster R-CNN and Sim-YOLO-Pose is also shown below. There is some false detection in the recognition resullt of Faster R-CNN. It shows that the Sim-YOLO-Pose model's accuracy is 94.7% and the recall rate is 91.5%. By contrast, the prediction effect of Sim-YOLO-Pose model is more consistent with the field scenarios in a comprehensive view.

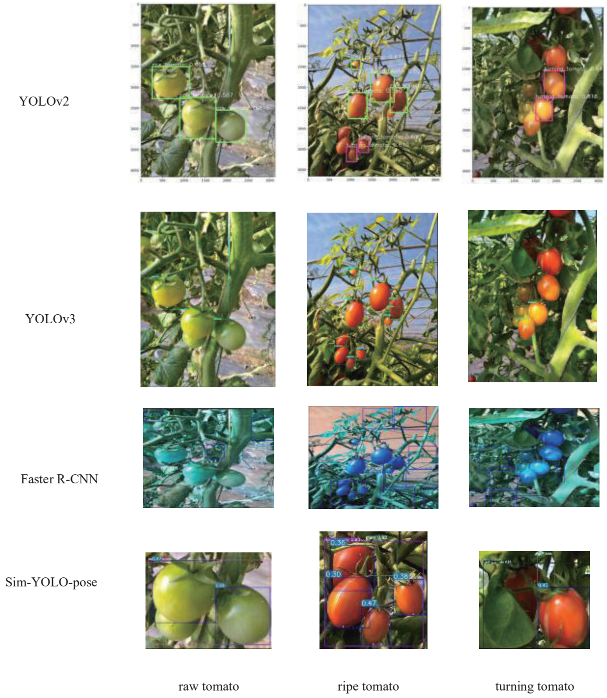


Figure 9. Recognition results of different YOLO series using K-means algorithm

The experimental results after 300 iterations are shown in Table II. Sim-YOLO-Pose improves the mAP by 0.45% than YOLOv5-pose model while 0.004% than YOLOv8-pose model. And the weight size of the Sim-YOLO-Pose is 16.1MB, which is smaller than YOLOv8's 23.1MB although it is bigger than YOLOv5's 14.7MB.

TABLE II. ANALYSIS RESULTS OF DIFFERENT VERSIONS TO ESTIMATE THE POSE OF TOMATOES

| Network name | Elapsed time of 300 iterations （h） | Evaluation index | | | Weights size(MB) |
|---|---|---|---|---|---|
| | | Precision rate (P,%) | Recall rate (R,%) | Mean Average Precision (mAP@.5, %) | |
| YOLOv5-pose | 0.849 | 0.926 | 0.836 | 0.93 | 14.7 |
| YOLOv8-pose | 1.053 | 0.961 | 0.911 | 0.971 | 23.1 |
| **Sim-YOLO-Pose** | **0.773** | **0.951** | **0.906** | **0.975** | **16.1** |

To further verify the rationality of the improved model, ablation experiments were conducted for each improvement. The test results are shown in Table III below. After replacing the key points with 5, the mAP was significantly improved. After adding the attention mechanism, the accuracy and recall rate were significantly improved.

TABLE III. ABLATION RESULTS

| Network name | replacement of keypoints | Hyperparameter optimization | Learning epoches | Precision | Recall | MAP50 |
|---|---|---|---|---|---|---|
| YOLOv7-pose_300 | √ | - | 300 | 0.927 | 0.79 | 0.716 |
| YOLOv7-pose_500 | √ | - | 500 | 0.983 | 0.77 | 0.67 |
| Sim-YOLOv7-pose | - | √ | 100 | 0.641 | 0.25 | 0.031 |
| **Sim-YOLO-Pose** | √ | √ | **200** | **0.951** | **0.906** | **0.975** |

### D. *Analysis of Stem Recognition Results*

In order to test the accuracy of the proposed method in identifying picking points, the fruit stem in the test sample was marked artificially with effective region. If the picking point is located within the region of the fruit stem, the picking point is considered effective; otherwise, the picking point identification is invalid. The marking results of the pickable fruit stem are shown in Table IV. Results showed that the experiment performed well in choosing the real stem according to the detected tomatoes, and the success rate can achieve 86.7% overall.

TABLE IV. PICKABLE STEM DATA SET

| Variety | Fushan88 tomatoes | | | Provence tomatoes |
|---|---|---|---|---|
| | ripe_tomatoes | turning_tomatoes | raw_tomatoes | |
| Number of stems | 156 | 58 | 46 | 118 |
| Number of pickable stems | 136 | 54 | 41 | 97 |

### E. *Embedded Terminal Deployment Test*

In practical applications, the Sim-YOLO-Pose model is deployed on the Jetson Xavier NX that is a versatile and powerful platform that brings the capabilities of AI and machine learning to the edge.

After training the model on deep learning workstation, it was optimized into the TensorRT-supported engine format to achieve more efficient inference. The optimization process utilizes TensorRTX, which is an open-source project dedicated to the deployment and optimization of deep learning models. The proposed algorithm operates at a speed of 30 frames per second, meeting the harvesting requirements. The detection results of the three categories are shown in the Figure 9. 347 images were selected from the test set for calculation, and 314 images were successfully detected. The success rate calculated according to the total data set was 90.3%, in which the number of red, yellow and green tomatoes were 189, 50 and 108. Accordingly, the number of succesfully detected tomatoes were 174,48 and 92, respectively. The success rates of three kinds of tomatoes were 92.1%, 96% and 85.2%, respectively. As for the images that failed to detect key points, it's analyzed that most of them were caused by the stem blocking more than half the volume of the fruit or due to the interference of neighbored fruits and steel mesh belts.

## V. Conclusion

This study proposes an end-to-end tomato and fruit stem detection network based on the Sim-YOLO-Pose model. It can simultaneously detect tomato and their thin fruit stems in near-color background environment. Furthermore, the YOLACT algorithm is used to segment the fruit stems that are suitable for picking, and the final cutting point is determined by combining the key points with the segmented stem. By combining the depth images provided by RealSense, the spatial cutting points and posture are estimated, which are then used to guide the robot in executing tomato picking. The specific experimental results are as follows:

1) The attention mechanism SimAM module is added to the backbone network, realizing an accuracy of 95.2% towards tomato and stem synchronously.

2) Yolo-pose method is utilized to estimate the 3D pose of the tomato verified by Realense D435i camera based on Jetson Xavier NX edge terminal, of which the mAP towards key points detection can achieve 97.2%.

However, the method in this paper also has shortcomings in actual harvesting operations. In the future, further pruning and quantifying of proximal fruit and stem models will be considered to make the model detection more accurate. And binocular cameras can be installed at the end of the robot arm to study the close-range fruit-stem matching method, and plan the picking sequence considered the occlusion of each fruit in single cluster and based on the maturity determination results.

## Acknowledgment

This work was supported by Opening Fund of National Key Laboratory of Agricultural Equipment Technology [Grant number NKL-2023-004], Key Research and Cultivation Projects of Beijing Universities under Project [Grant number 2021YJPY201] and National Key Research and Development Program "Intergovernmental International Science and Technology Innovation Cooperation" Special Project [Grant number SQ2023YFE0101940].

## References

[1] X.F. Liang, C.Q. Jin, M.D. Ni, et al. Acquisition and experiment on location information of picking point of tomato fruit cluster[J]. Transactions of the Chinese Society of Agricultural Engineering,2018,34(16):163-169.

[2] J.W. Yan, Y. Zhao, L.W. Zhang, et al. Recognition of Rosa roxbunghii in natural environment based on improved Faster RCNN[J]. Transactions of the Chinese Society of Agricultural Engineering, 2019,35(18):143-150.

[3] X.Z. Wang, X. Han, H.P. Mao. Vision-based detection of tomato main stem in greenhouse with red rope. Transactions of the Chinese Society of Agricultural Engineering, 2012,28(21).135-141.

[4] M. Hussain, L. He, J. Schupp, et al. Green fruit segmentation and orientation estimation for robotic green fruit thinning of apples[J]. Computers and Electronics in Agriculture,2023,207:107734.

[5] Z.T. Wang, Y.T. Xu, J.Y. Yu, et al. Instance segmentation of point cloud captured by RGB-D sensor based on deep learning[J]. International journal of computer integrated manufacturing,2021,34(9):950-963.

[6] D.E. Liu, L. Zhu, W.Z. Ji, et al. Real-time identification, localization, and grading method for navel oranges baased on RGB-D camera[J].Transactions of the chinese society of agricultural engineering, 2022,38(14):154-165.

[7] H. Li, H.X. Tao, L.H. Cui, et al. Recognition and localization method of tomato based on SOM — K-Means algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(1):24-29.

[8] J.Y. Choi, B.J. Lee, B.T. Zhang. Human body orientation estimation using convolutional neural network[J]. arXiv preprint arXiv:1609.01984,2016.

[9] K. Haras, R. Vemulapalli, R. Chellappa. Designing deep convolutional neural networks for continuous object orientation estimation[J]. arXiv:1702.01499,2017.

[10] Q. Zhang, Y.S. Pang, B. Li. Visual positioning and picking pose estimation of tomato clusters based on instance segmentation [J]. Transactions of the Chinese Society for Agricultural Machinery, 2023,54(10):205-215.

[11] M.C. Liu, C.Y. Chu, M.S. Cui, et al. Red ripe strawberryrecognition and stem key point detection based on improved YOLO v8-Pose[J]. Transactions of the chinese society for agricultural machinery, 1-12[2024-03-20].

[12] L.X. Yang, R.Y. Zhang, L.D. Li, et al.SimAM: A simple, parameter-free attention module for convolutional neural networks[C]//International conference on machine learning. PMLR, 2021: 11863-11874.

[13] D. Bolya, C. Zhou, F.Y. Xiao, et al. YOLACT: real-time instance segmentation [C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE,2019:9157-9166.

[14] M. Everingham, L. Van Gool, C.K. Williams, et al. The pascal visual object classes(voc) challenge[J]. International journal of computer vision,2010,88:303-338.

[15] N. Kounalakis, E. Kalykakis, M. Pettas, et al. Development of a Tomato Harvesting Robot: Peduncle Recognition and Approaching[C]. Turkey: 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications(HORA).IEEE.

[16] Z.T. Ning, L.F. Luo, J.X. Liao, et al. Recognition and the optimal picking point location of grape stems based on deep learning[J].Transactions of the Chinese Society of Agricultural Engineering,2021,37(9):222-229.

[17] X.M. Chen, L.C. Wang, J. Zhang, et al. Research on seafood target detection algorithm based on YOLOv5 and ASFF algorithm[J]. Radio Engineering, 2023:1-11.

[18] T.H. Li, M. Sun, Q.H. He, et al. Tomato recognition and location algorithm based on improved YOLOv5[J]. Computers and Electronics in Agriculture,2022:1-11..

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative Adversarial Networks[C].Advances in Neural Information Processing Systems (NIPS), 2014, 27:2672-2680.

[20] X.M. Chu, H. Lin and Z.Y. Wang. K-means clustering analysis of massive AIS data based on Spark platform[J]. Traffic Science and Technology,2021(03):138-141.

[21] Y.Z. Fang, Y.N. Song, R.H. Xu, et al. Research on remote sensing image detection based on lightweight Yolact[J]. Wireless Internet Technology,2023,20(10):117-121.