Biosystems Engineering 246 (2024) 122–134

Contents lists available at ScienceDirect

Biosystems Engineering

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Real-time detection of mature table grapes using ESP-YOLO network on embedded platforms

Jiaoliao Chen [a,*], Huan Chen [a], Fang Xu [a], Mengnan Lin [a], Dan Zhang [b], Libin Zhang [a]

[a] Key Laboratory of Special Purpose Equipment and Advanced Processing Technology, Ministry of Education & Zhejiang Province, Zhejiang University of Technology, Hangzhou, 310012, China
[b] Department of Mechanical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

ARTICLE INFO

Keywords:
Object detection
Efficient layer aggregation networks
Lightweight network
Deep learning
Attention mechanism

ABSTRACT

The real-time and high-precision detection methods on embedded platforms are critical for harvesting robots to accurately locate the position of the table grapes. A novel detection method (ESP-YOLO) for the table grapes in the trellis structured orchards is proposed to improve the detection accuracy and efficiency based on You Only Look Once (YOLO), Efficient Layer Shuffle Aggregation Networks (ELSAN), Squeeze-and-Excitation (SE), Partial Convolution (PConv) and Soft Non-maximum suppression (Soft_NMS). According to cross-group information interchange, the channel shuffle operation is presented to modify transition layers instead of the CSPDarkNet53 (C3) in backbone networks for the table grape feature extraction. The PConv is utilised in the neck network to extract the part channel's features for the inference speed and spatial features. SE is inserted in backbone networks to adjust the channel weight for channel-wise features of grape images. Then, Soft_NMS is modified to enhance the segmentation capability for densely clustered grapes. The algorithm is conducted on embedded platforms to detect table grapes in complex scenarios, including the overlap of multi-grape adhesion and the occlusion of stems and leaves. ELSAN block boosts inference speed by 46% while maintaining accuracy. The mAP @0.5:0.95 of ESP-YOLO surpasses that of other advanced methods by 3.7%–16.8%. ESP-YOLO can be a useful tool for harvesting robots to detect table grapes accurately and quickly in various complex scenarios.

Nomenclature

Abbreviations	
AP	average precision
C3	CSPDarkNet53
CBAM	Convolutional Block Attention Module
ELAN	Efficient Layer Aggregation Networks
ELSAN	Efficient Layer Shuffle Aggregation Networks
FLOPs	floating-point operations
FN	False Negative
FP	False Positive
IoU	Intersection over Union
mAP	mean average precision
NMS	Non-Maximum Suppression
PConv	Partial Convolution
SE	Squeeze-and-Excitation
Soft_NMS	Soft Non-maximum suppression
TP	Ture Positive
YOLO	You Only Look Once
Symbols	

(*continued on next column*)

(*continued*)

Abbreviations	
α - *IoU*	Loss Function
σ	Penalty Coefficient.
b_i	Compared Bounding Boxes
M	Bounding Box with The Highest Confidence
N_t	Threshold Generally
S_i	Confidence Score of The Bounding Box

1. Introduction

Table grapes are widely planted in the world, especially in China (Khan et al., 2020; Wu et al., 2016). However, the harvest is mainly done by hand with labour-intensive and time-consuming. The harvesting robot for table grapes is an increasingly hot issue to replace skilled humans. The detection is a complicated process for table grape harvesting robots due to ensure table grape integrity. However, some

* Corresponding author.
E-mail address: jlchen@zjut.edu.cn (J. Chen).

https://doi.org/10.1016/j.biosystemseng.2024.07.014
Received 3 January 2024; Received in revised form 16 July 2024; Accepted 17 July 2024
Available online 31 July 2024

detection methods are not easy to apply on table grapes harvesting robots due to the low detection permission in various complex scenarios, including overlap and occlusion. In addition, the detection methods are required to be run on embedded platforms for the harvesting robot because of the limited space, which easily results in low inference rates. It is essential to investigate the detection method for the performance improvement of detection accuracy in the harvesting robot.

Many detection methods were proposed to solve the problems in fruit detection, such as visual inspection, hyperspectral analysis, acoustic detection and chemical analysis (Abbaszadeh et al., 2013; Egea et al., 2021; Marin-San Roman et al., 2023). It is difficult for acoustic detection and chemical analysis to locate the fruits due to the low recognition ability for the fruits location with ambient interferences. A few researchers employed hyperspectral analysis in fruit detection to improve the location performance (Safren et al., 2007; Y. Zhang et al., 2019). Okamoto and Lee (2009) designed a hyperspectral-based detection method to detect small fruits by comparing the hyperspectral information of citrus leaves and fruits. Wouters et al. (2013) developed a multispectral vision sensor with informative wavebands for the location of flower buds on trees. Steinbrener et al. (2019) presented a hyperspectral-based convolutional neural network to classify fruits and vegetables. However, these hyperspectral analysis methods are difficult to be integrated into a real-time recognition system due to the large computation burden. Moreover, the hyperspectral camera is extremely expensive for a large-scale application in harvesting robots. The visual inspection is a common approach for fruit identification with the low cost of apparatus, consistency, high speed and accuracy, such as image processing and machine learning (Taghadomi-Saberi & Hemmat, 2015).

Image processing is a general method to detect objects, which can be used to recognise and locate fruits. Many researchers proposed colour-based, geometric-based and texture-based segmentation methods to detect various fruits, such as tomatoes, red apples, peaches, mangoes and pineapples (Chaivivatrakul & Dailey, 2014; B. Li et al., 2010; Payne et al., 2013; Qiang et al., 2014; R. Zhou et al., 2012). Xiong et al. (2017) proposed an Otsu algorithm with the Canny edge detection to determine the location of swaying grapes in space. Ji et al. (2016) presented a segmentation method with an adaptive histogram equalisation to distinguish between apples and branches. Jin et al. (2022) designed a vision method to accurately locate the picking point of table grapes via feature extraction from far-range images and close-range images. Xiong et al. (2018) developed a segmentation method for the swaying litchi clusters to detect harvesting points with fuzzy C-means clustering and binocular visual stereo matching. However, image processing is difficult to apply in harvesting robots due to the uncertain thresholds resulted from complex scenarios with lighting conditions and plant growth.

Currently, many machine learning methods for the harvest robots were explored to locate the fruits (Bulanon et al., 2004; Domingues et al., 2022; Kuznetsova et al., 2020; Linker et al., 2012). Tang et al. (2017) proposed a convolutional neural network with a k-means feature to recognise weeds in soybean seedling fields. Luo et al. (2018) used k-means clustering and profile analysis to detect the cutting points of the overlapping grapes. However, these conventional machine learning methods easily result in low-precision detection in complex scenarios due to the feature extraction by human intervention. Yuhao Bai et al. (2022) designed a multi-network fusion algorithm with a data processing, target recognition network and semantic segmentation network for the detection of matured cucumbers in a non-structural scene. Yifan Bai et al. (2024) proposed a YOLOv7 algorithm with a Swin Transformer to detect strawberries with a mAP of 92.1%. Although some detection methods are used to detect individual fruit, these approaches struggle with table grape clusters due to the complex geometric features in the obscure scenario. Xu et al. (2022) presented Mask R–CNN to accurately recognise cherry tomatoes with a fruit and stem accuracy of 93.76% and 89.34%, respectively. Q. Zhang and Gao (2020) improved Faster R–CNN with multi-scale feature maps to extract small stalks for grape clusters received an average precision of 92.07%. Although these methods perform high detection accuracy, they cannot be applied to the harvesting robot because of the large computational burden. Thus, it is necessary to study lightweight YOLO, especially for embedded systems. Shang et al. (2023) proposed a lightweight YOLOv5s algorithm with Shuffle Net and Ghost Net to detect apple flowers at 2.48 fps on embedded platforms. H. Li et al. (2021) presented a lightweight YOLOv4-tiny algorithm to detect the grapes with a mAP of 92.1%. Although they used lightweight modules to reduce the parameters, these algorithms still obtain low detection accuracy with low inference speed for grape harvesting robots.

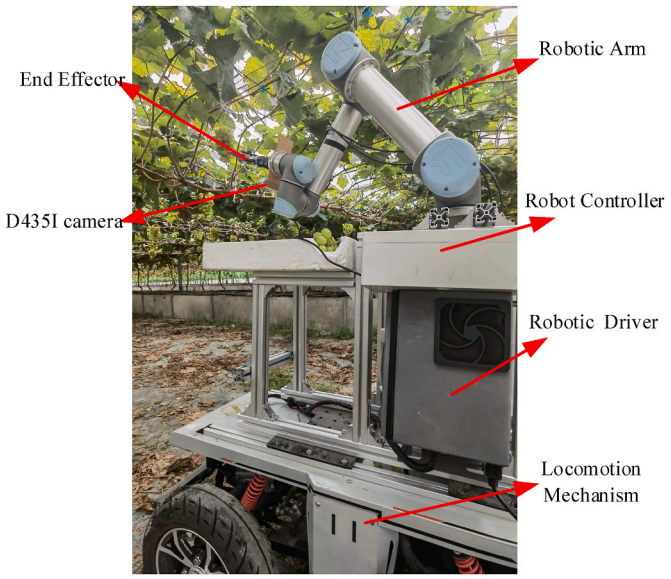

Fig. 1. Hardware of harvesting robot.

The objective of this study is to investigate a novel detection algorithm (ESP-YOLO) on embedded platforms for the grape harvesting robot to improve detection accuracy and speed. Efficient Layer Shuffle Aggregation Networks (ELSAN) are proposed to replace C3 blocks in backbone networks to enhance feature extraction of table grapes. Pconv is used to replace the convolution operations in the network to reduce parameter numbers and capture spatial features. ELSAN and Pconv are combined to try to improve inference speed on embedded systems. Squeeze-and-excitation (SE) is integrated to improve YOLO for channel-wise feature interaction. According to the channel-wise and spatial features, the location and quantities of SE in ESP-YOLO are optimised to reduce Instance-level noise. SE and Soft Non-maximum suppression (Soft_NMS) enhance the segmentation ability of table grapes to improve recognition precision under overlapping and occluding conditions. The experiments are conducted to verify the performances of ESP-YOLO on embedded platforms with grape images involved with the overlap, occlusion and distant shot.

2. Materials and methods

2.1. System overview

The hardware configuration for the harvesting robot includes an

Table 1
Experimental platform.

Items	Train Parameters	Test Parameters
CPU	Intel(R) Core (TM) i5-13600 K	ARMv8 Processor
GPU	NVIDIA GeForce RTX 3070	NVIDIA Tegra X1
Operating System	Windows 10	Ubuntu18.04
Acceleration environment	CUDA11.7	CUDA 10.2
Development Platform	torch-1.12.1	torch-1.7.0

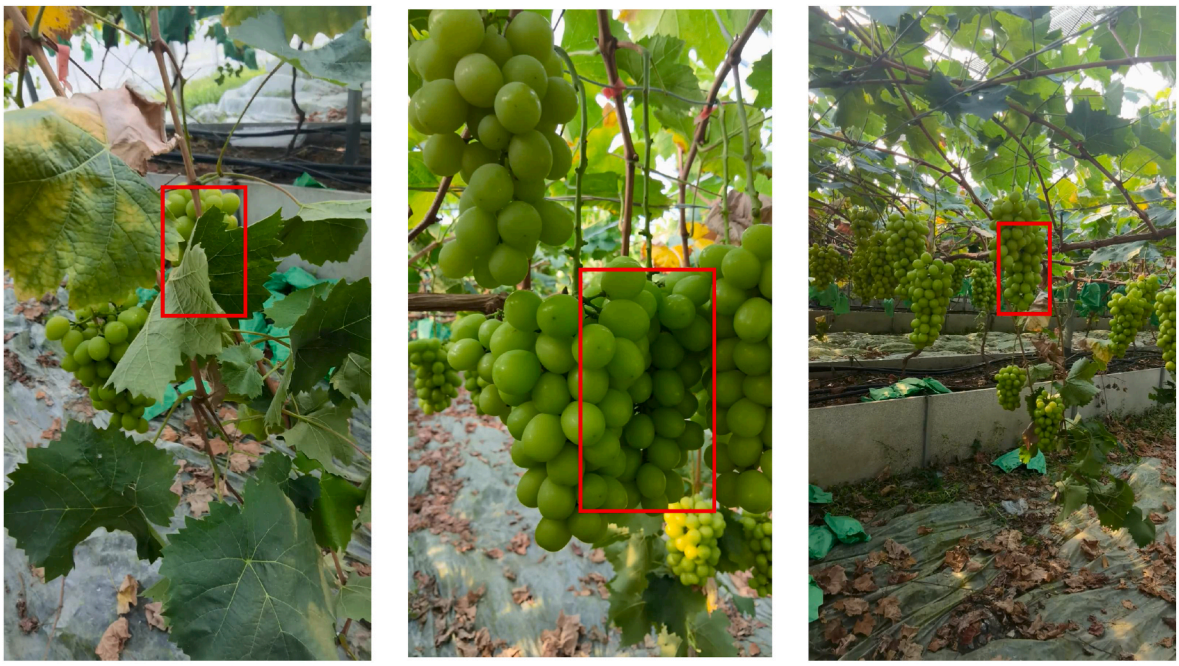

Fig. 2. Grapes images in the trellis structured orchards including occluded grapes(a), overlapped grapes(b) and distant shot of grapes(c).

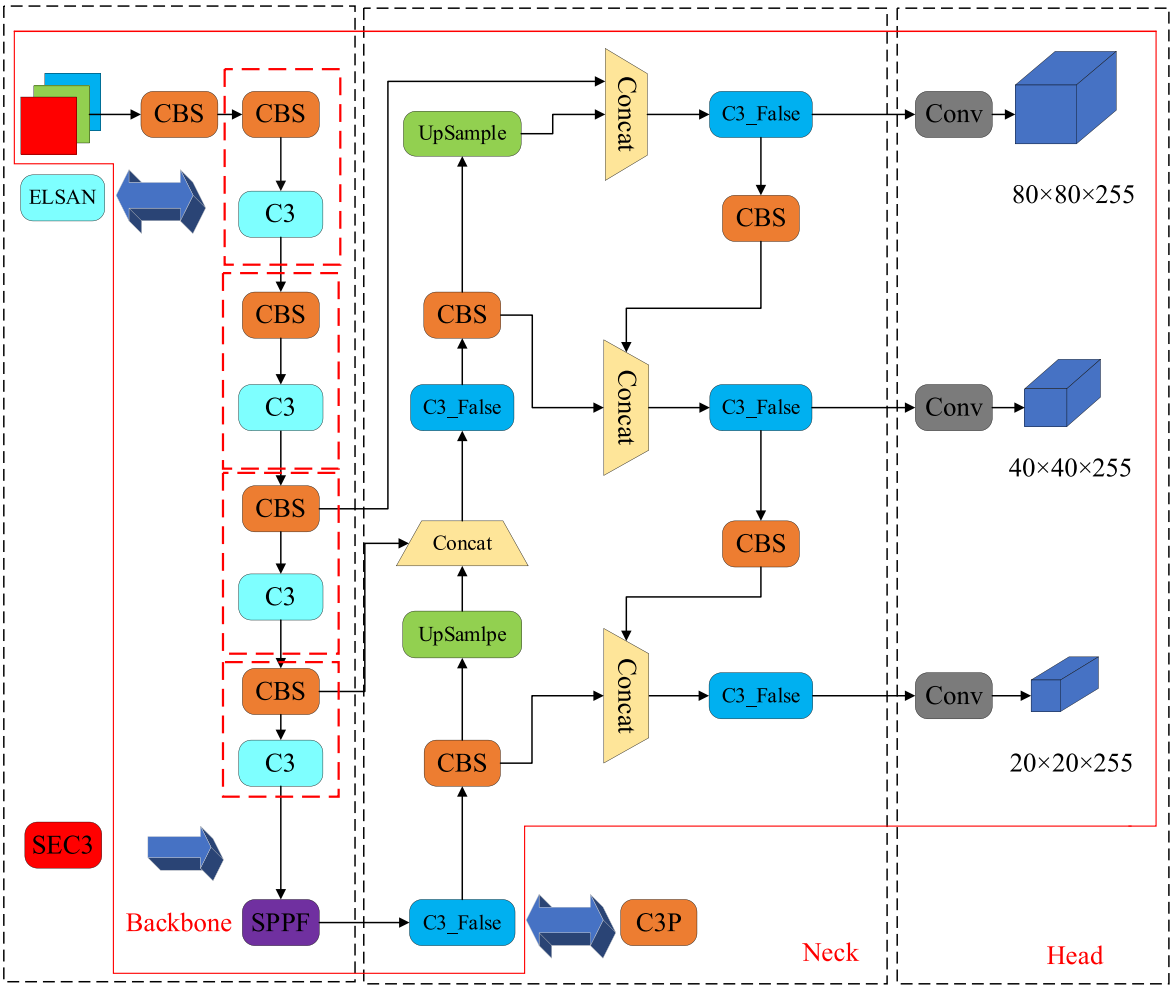

Fig. 3. Structure diagram of ESP-YOLO.

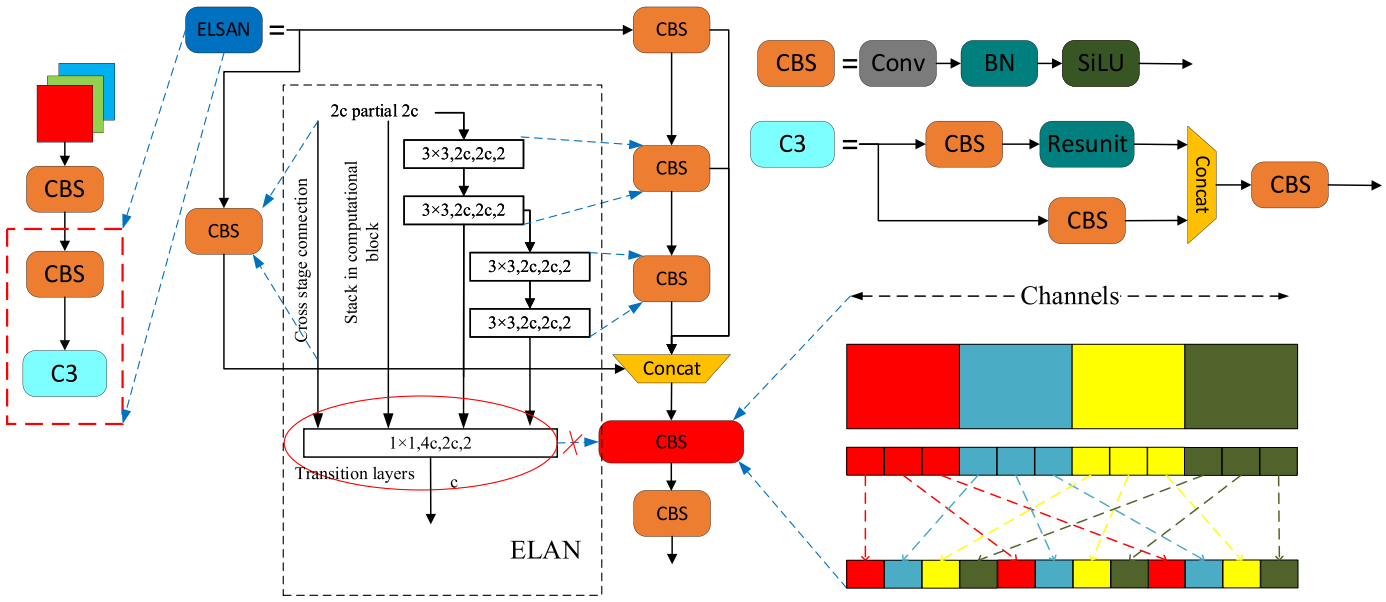

Fig. 4. Structure diagram of ELSAN.

Intel Realsense D435I camera, Jetson Nano, UR5 robotic arm and a wheeled-locomotion mechanism, as illustrated in Fig. 1. The D435I depth camera supplies RGB images and depth images of the grapes for the embedded system. ESP-YOLO which trained on a computer determines the three-dimensional coordinates via the real-time image from the camera in the Jetson Nano. The main specifications of the computer and the embedded platforms are in Table 1. Then, the grapes position information is served as the input for the control system of the robotic arm and wheeled locomotion mechanism to accomplish the harvesting task. The harvesting robot in real-scenario experiments can be used to verify the proposed method.

2.2. Image acquisition and labelling

In the study, the table grapes are planted in trellis-structured orchards. The Shine Muscat grapes are chosen to be recognised because of the separation difficulty from the background with their similar green colour. The images of table grapes during the harvest period in the greenhouse were acquired by Huawei Mate 40 and Intel Realsense D435I camera on 22 July and 7 August 2022 in Nanxun County, Huzhou City, Zhejiang Province, China. The dataset was collected on different dates to minimise the impact of subtle variations in the grapes during the maturity stage. Since the depth camera cannot obtain high-resolution images, we chose the images from mobile phones as the training and validation datasets. The grape images are taken from the spot distances in different directions with 0.5–0.8 m (close shot) and 0.8–2.5 m (distant shot) under the condition of various scenarios, including the overlap and occlusion (Fig. 2). 1431 images of table grapes with clear target contours and textures are collected as the experimental dataset. 477 Images in the experimental dataset are collected by a random rotated mobile phone with the angle range of -30-30° to add the rotation features of grape images in the dataset, which can be used to decrease the impact on image rotation. Additionally, 270 images of table grapes from the depth camera are used for transfer learning to verify the detection algorithm on the harvesting robot in real scenarios. The public dataset of grapes is chosen to test the robustness of the algorithm (Santos et al., 2020).

In the experimental dataset, 332 images are selected as the test dataset, including the scenario of overlap, occlusion and distant shot. The rest 952 and 147 images are randomly selected as the training dataset and the validation set, respectively. Since the dataset with sufficient sample images is a prerequisite for the deep neural networks, some image processing techniques such as noise enhancement, brightness change, rotation and mirroring are employed to augment data for the set diversity of the table grape. Furthermore, the grape images are rotated for augmented data to reduce the impact of different camera angles on identification accuracy. The grapes images were manually labelled via LabelImg with well-fitted rectangular frames. According to our skill knowledge, we depicted the actual size of the targets blocked by branches and leaves or other grapes in the solid box (Fig. 2). The labelled files are stored in the PASCAL VOC format for the dataset.

The training and validation datasets and the labelled file are used to train ESP-YOLO. The test dataset is used to evaluate the performance of ESP-YOLO. The images in the test dataset with overlap, occlusion and distant shots scenarios are selected to construct three separate sub-test datasets, including the overlapping dataset, occlusion dataset and distant shot dataset.

2.3. ESP-YOLO construction

The accuracy, real-time and lightweight of the model are the key factors in the vision system of the harvesting robots. YOLOv5 is a small size model with fast inference speed and an acceptable detection accuracy in computers. YOLOv5 structure divides into the backbone network for feature extraction, the neck network for feature fusion and the head network for anticipation in the solid red box in Fig. 3.

In backbone networks, CBS and C3 blocks are used for downsampling and feature extraction. However, YOLOv5 is hard to run on embedded platforms due to the low inference speed resulted from the burdensome backbone network. In the study, ELSAN replaced CBS and C3 to extract the deeper features information with the lower floating-point operations (FLOPs) (in the dotted red box in Fig. 3) and SE is used to heighten the weight of useful features in the backbone network. In the neck network, C3P enhances the inference speed with high detection accuracy instead of C3_False.

2.4. Lightweight network

In this study, ELSAN is modified to replace the CBS and C3 with many res units and transition layers in Fig. 4. The res unit generally needs a large training set to extract features effectively. However, it is a complicated task to acquire the huge training set of table grapes. Moreover, the res unit and transition layers can weaken the ability for

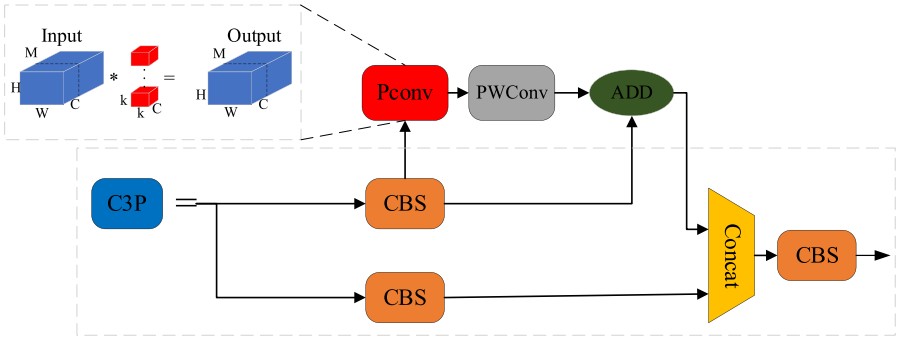

Fig. 5. Structure diagram of C3P.

parallel computing resulted from raised memory access, which is an important factor for the embedded platforms.

Efficient Layer Aggregation Networks (ELAN) consist of the cross-stage connection and the stack computational block. The cross-stage connection block can achieve the shortest path for faster inference speed (Wang, Liao, & Yeh, 2023). The stack computational block can enrich the gradient information to improve the learning capacity. Thus, ELAN can replace C3 to enhance the network depth for the feature extraction capabilities. However, YOLO sets channel numbers for each layer, which results in redundant transition layers in ELAN and CBS in the backbone network. The redundant operations in these transition layers greatly extend the computational burden, herein, these redundant transition layers in ELSAN are deleted to reduce FLOPs. Owing to the weaker interaction of channels resulted from the removal of transition layers, the channel shuffle operation is inserted in ELSAN to remedy cross-group information interchange.

The lightweight convolution is a popular method to deal with redundant feature maps, such as PConv (Chen et al., 2023). PConv efficiently extracts features via the convolution of a subset of the total channels. Although PConv can ignore some features, it can improve inference speed for applications on embedded platforms with limited memory and computation resources. The feature map of input size $H\times W\times M$ is split into the feature maps of size $H\times W\times C$ (part A) and $H\times W\times (M-C)$ (part B). After convolution operations, part A is directly connected to part B for the feature map in Fig. 5. However, PConv is not applied in the backbone network to extract the features due to the reduced information in the inter-channel. C3P based on PConv and PWConv is inserted in the neck network to capture spatial features. The channel number C in C3P can affect FLOPs and the spatial features. Then, C is set as $M/4$ (Chen et al., 2023). Thus, the YOLO network is improved by using ELSAN and C3P for a lightweight network on embedded platforms.

2.5. Channel attention mechanism

The instance-level noise easily occurs in the images for the over-lapping grapes due to the intra-class feature coupling and the inter-class feature boundary blurring. Since ELSAN reduces the information of feature extractions, the instance-level noise in the images is hard to remove, such as overlap, occlusion and small grapes. Then, the attention mechanism should increase the weight of the useful features to reduce the instance-level noise with the fewer feature information.

Based on global pooling and sigmoid, the squeeze-and-excitation (SE) block is able to get global information and adjust the feature weights adaptively(Hu et al., 2020). SE is a universal block in the network to extract more useful features, which is integrated into C3 (SEC3) to connect YOLO blocks. Although SEC3 can increase the computational volume slightly, the appropriate locations and quantity of SEC3 in the networks drop down the interferences of instance-level noise significantly. Meanwhile, the depth of channel-wise features

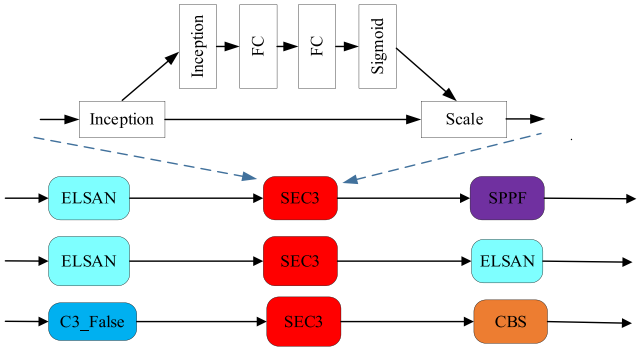

Fig. 6. Structure diagram of SE.

relies on SEC3 locations. According to the YOLO structure characters, SEC3 is inserted between ELSAN and SPPF(ESEP-YOLO), in each down-sampling stage (EDP-YOLO) and each down-sampling and up-sampling stage (EDUP-YOLO), respectively (Fig. 6). The three improved YOLO networks with SCE3 the are compared on detection accuracy in the experiments.

2.6. Soft Alpha_NMS algorithm

A series of prediction boxes as the detected table grapes can be obtained by using the improved YOLO network. In the inference phase, Non-Maximum Suppression (NMS) algorithm can be used to calculate their Intersection over Union (IoU) of prediction boxes for detection results with the highest IoU. However, some prediction boxes of the table grapes with low IoU are also recognised accurately and cannot be removed, especially in occlude and overlap scenarios. Thus, Soft_NMS is more suitable for occlude and overlapping scenes because of the retained prediction boxes with higher IoU than the threshold, instead of the highest IoU in NMS (Bodla, Singh, Chellappa, & Davis, 2017a). Nevertheless, Soft_NMS increases the post-processing time during the inference phase. To enhance the efficiency and robustness of Soft_NMS, α-IOU has been introduced to determine the confidence score of the bounding box S_i (He et al., 2021). Soft Alpha_NMS is defined as,

$$S_i=\begin{cases} S_i & \alpha-IoU(M,b_i)<N_t \\ S_i e^{-\frac{\alpha-IoU(M,b_i)^2}{\sigma}} & \alpha-IoU(M,b_i)\geqslant N_t \end{cases} \tag{1}$$

where M is the bounding box with the highest confidence, b_i denotes the compared bounding boxes in the current target, N_t is the threshold generally set to 0.5, α - IoU (M, b_i) is the loss function and σ is the penalty coefficient.

The positions and quantities of SE have been determined according to the comparison of detection precision from datasets of table grapes.

Table 2
The hyperparameters in the training process.

Hyperparameter	Epochs	Batch Size	Momentum	Weight Decay	Learning Rate
Value	120	4	0.937	0.0005	0.01

Then, the improved SE is integrated with ELSAN, C3P and Soft Alpha_NMS (ESP-YOLO) is proposed to detect the table grapes on embedded platforms.

2.7. Evaluation indicators

ESP-YOLO is trained by the training and validation sets on a computer to generate weight files. The setting hyperparameters in train process are listed in Table 2. Then, the weights file is transferred to an embedded platform to recognise the table grapes in the experiments. The ablation experiment is conducted to assess the performance of the different modules in Section 2. The performance of ESP-YOLO is accessed on test datasets in terms of detection precision and speed, including occlusion subset, overlap subset, distant shot subset and et al. Its robustness is also tested on publicly available datasets. Finally, the proposed method is verified by a harvesting robot of the table grapes in actual scenarios.

The performance of detection algorithms can be evaluated by using *recall*, *precision*, *F1-score*, *AP*, inference speed and total network parameters, which are defined as,

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathit{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$F1 - \mathit{score} = \frac{2\mathit{Precision} \times \mathrm{Recall}}{\mathit{Precision} + \mathrm{Recall}} \tag{4}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{5}$$

$$mAP = \frac{\sum_{i=1}^{k} AP}{k} \tag{6}$$

where true positive (TP), false positive (FP) and false negative (FN) are the number of true positive samples, false positive samples and false negative samples, respectively. *AP* is the area enclosed by the *Precision-Recall* (*PR*) curve and the coordinate axis with the range of 0–1. *R* is the integral variable of the product of *recall* and *precision* and *mAP* is the average *AP* under all categories. The *mAP*@0.5 is the *mAP* as the IoU threshold is 0.50. The *F1-score* and *mAP* perform comprehensive metrics of recognition precision, recall and localisation accuracy. Then, the performance of ESP-YOLO can be accessed in terms of the *F1-score*, *mAP*@0.5 and *mAP*@0.5:0.95.

3. Results

3.1. Network improvement ablation experiments

ELSAN only replaces the C3 and CBS blocks in the backbone network (E-YOLO). According to the test dataset, E-YOLO results are showed in Table 3. The inference speed of E-YOLO is 48% faster than YOLOv5s, thanks to the reduction in computational burden resulted from the removal of redundant transition layers and residual units. Although ELSAN obtains larger FLOPs in compare to Mobilenet (Howard et al., 2019), the inference speed of ELSAN is faster than Mobilenet. The detection accuracy of E-YOLO is nearly identical to that of YOLOv5s, whereas the other lightweight networks lead to a decrease in the detection accuracy.

In the test dataset, these results of ESEP-YOLO, EDP-YOLO and EDUP-YOLO are shown in Table 4 and Fig. 7. ESEP-YOLO improves the mAP@0.5 and mAP@0.5:0.95 of 1.2% and 2.9% than EDP-YOLO and EDUP-YOLO due to the fewer false alarm in the solid box (Fig. 7). The reason is that too many SEC3 blocks in EDP-YOLO and EDUP-YOLO may focus excessively on some specific local information instead of global information. Although EDP-YOLO and EDUP-YOLO identify more grapes within the dashed boxes in Fig. 7 (b) and (c), their localisation of the table grapes is not totally accurate. Thus, an SEC3 block between ELSAN and SPPF effectively extracts single-layer channel-wise features with deep semantic information.

A comparison of various modified strategies is presented in Table 5. The combined ELSAN and C3P accelerate the inference speed by 68.9% in contrast to YOLOv5s due to the improvement to diminish FLOPs in Section 2.4. C3P or SE separately inserted in E-YOLO does not increase accuracy. However, YOLOv5 combined with SE and C3P increases mAP@0.5:0.95 by 1.6%. Although Soft Alpha_NMS inevitably increases the inference time, the mAP@0.5:0.95 is improved by 6.2% due to the more accurate criteria for NMS in Section 2.6. Table 6 can be concluded that Soft_NMS can achieve better detection accuracy if α is 1. Despite of a 12% decrease in the inference speed of ESP-YOLO in contrast to EP-YOLO, the mAP@0.5:0.95 and inference speed are improved by 9.8% and 48.6% in contrast to those of YOLOv5s, respectively. Therefore, ESP-YOLO is an efficient network for table grapes detection concerning high accuracy and speed.

3.2. Results in deference scenarios

According to the overlap, occlusion and distant shot test datasets, the results of ESP-YOLO and YOLOv5s are shown in Table 7 and Fig. 8. In all test datasets, ESP-YOLO significantly improves mAP@0.5 by 1% and mAP@0.5:0.95 by 10% for table grapes detection. In the overlap test dataset, the mAP@0.5:0.95 of ESP-YOLO is increased by 4.6% due to the reduced miss detections in the dotted box (Fig. 8) resulted from the separation capabilities of SE and Soft Alpha_NMS. In the distant shot, compared with YOLOv5s, ESP-YOLO achieves a 4.5% increase of mAP@0.5. ESP-YOLO's detection accuracy is better than YOLOv5 in the

Table 4
Detection results of ESP-YOLO, EDP-YOLO and EDUP-YOLO.

Model	mAP@0.5 (%)	mAP@0.95 (%)	F1-score	FLOPs (G)	Inference time (ms)
ESEP-YOLO	**89.4**	**59.5**	**0.876**	**9.5**	**86.2**
EDP-YOLO	88.3	57.8	0.868	12.5	92
EDUP-YOLO	88.3	57.8	0.862	15.4	95

Table 3
Detection results of four different backbone networks.

Backbone	mAP @0.5 (%)	mAP @0.5:0.95 (%)	F1-score	FLOPs (G)	Memory access(M)	Parameters (M)	Inference Time (ms)
C3	88.6	**57.6**	**0.874**	16.4	156.4	6.69	136.2
Mobilenet	87.3	55.9	0.852	**6.1**	156.4	**3.36**	93.6
Ghostnet	**88.9**	57.5	0.868	8.2	155.5	3.51	117.9
ELSAN	88.7	**57.6**	0.873	10.8	**105.7**	5.26	**93.0**

Fig. 7. Detection results of ESP-YOLO with (a) before SPPF, (b) in down-sampling and (c) in up-sampling and down-sampling.

Table 5
Ablation study of ESP-YOLO.

Model	name	mAP@0.5 (%)	mAP @0.5:0.95 (%)	F1-score	Inference time(ms)
YOLOv5s	YOLOV5s	88.6	57.6	0.874	136
+ELSAN	E-YOLO	88.7	57.6	0.873	93.0
+ELSAN + C3P	EP-YOLO	88.7	57.6	0.872	**80**
+ELSAN + SE	ES-YOLO	88.2	58.0	0.875	95
+SE + C3P	SEP-YOLO	88.9	58.5	**0.88**	108.1
+ELSAN + C3P + SE	ESEP-YOLO	**89.4**	59.6	0.877	86.2
+ELSAN + C3P + SE + Soft_NMS	ESP-YOLO	**89.4**	**63.3**	0.876	90

Table 6
Detection results of different α

α	0.5	1	2	3	4	5
mAP@0.5	88.7	89.4	88.5	89.4	88.9	89.1
mAP@0.5:0.95	62.1	63.4	62.8	63.2	62.9	62.6

complex scenario.

According to the public grape dataset, the detection results for ESP-YOLO and YOLOV5s are shown in Fig. 9. The red box, red circle and red dashed circle in Fig. 9 illustrates that ESP-YOLO performs superior detection accuracy compared to YOLOv5 in overlap, occlusion and distant shot scenarios, which is similar to custom datasets results. Meanwhile, ESP-YOLO is more generalisation and robust than YOLOv5.

3.3. Comparison of different methods

According to the test dataset, the results of ESP-YOLO, Faster R–CNN (Ren et al., 2015), SSD (Liu et al., 2016), YOLOv5s, YOLOv5-Mobile (Zeng, Li, Song, Zhong, & Wei, 2023), YOLOv7-Tiny (Gu et al., 2023), YOLOv8n (Yang et al., 2023), YOLO-Grape (Li et al., 2021), YOLOv5-CFD (Zhu et al., 2023) and SM-YOLOv4 (Qiu et al., 2022)are shown in Table 8. YOLOv8n, YOLO-Grape, YOLOv5-CFD, and SM-YOLOv4 are run on Jetson Xavier Nx due to their high hardware and software requirements (Python 3.8).

The model sizes of detection algorithms based on YOLO are less than 30 MB except YOLOv5-CFD, which is easy to be operated on embedded platforms. Then, the slow inference speed of algorithms is a key factor in the deployment of embedded platforms. The inference time of Faster-RCNN and SSD for each image is more than 1000ms, which cannot be used on embedded platforms. The detection speed of ESP-YOLO is much greater than other table grape detection algorithms. Compared to other lightweight methods, the inference speed of ESP-YOLO is faster by 2.8%–83% on the Jetson Nano benefited from the improvements to match the embedded platforms in Section 2.4.

The mAP@0.5:0.95 of ESP-YOLO surpasses that of YOLOv7-Tiny by 8.8% benefited from the precision segmentation of overlapping grapes in the dotted box and solid box (Fig. 10). The mAP@0.5:0.95 of ESP-YOLO is enhanced by 13.4% in contrast to YOLOv5-Mobile due to the less miss detection in distant shot scenarios. The mAP@0.5:0.95 of ESP-YOLO is 3.7%–16.7% than other table grape detection algorithms due to the increased segmentation capacity via SE, C3P and Soft Alpha_NMS. The mAP@0.5:0.95 of ESP-YOLO is 5.5% higher than that of YOLOv8n with the same detection speed. Thus, Considering the detection speed and accuracy, ESP-YOLO outperforms these advanced algorithms.

3.4. Real scenarios experiments

The real-time embedded system runs on a harvesting robot to pick the table grapes. The trained ESP-YOLO is implemented into the embedded system to detect table grapes in a variety of work tasks, such

Table 7
Detection results of ESP-YOLO and YOLOv5s in different scenarios.

Test dataset	Model	mAP@0.5(%)	mAP@0.5:0.95(%)	Precision(%)	Recall(%)	F1-score
all	YOLOv5s	88.6	57.6	93.2	**83**	87.4
	ESP-YOLO	**89.4**	**63.4**	**93.7**	82.3	**87.6**
overlap	YOLOv5s	83.4	51.9	81.5	83.4	82.4
	ESP-YOLO	**90.3**	**54.3**	**89.5**	**88.5**	**88.9**
occlusion	YOLOv5s	89.8	59.3	**92.3**	85.0	88.5
	ESP-YOLO	**91.2**	**59.6**	91.4	**88.3**	**89.8**
distant shot	YOLOv5s	81.6	50.2	84.1	76.7	80.2
	ESP-YOLO	**85.3**	**51.1**	**89.6**	**76.7**	**82.6**

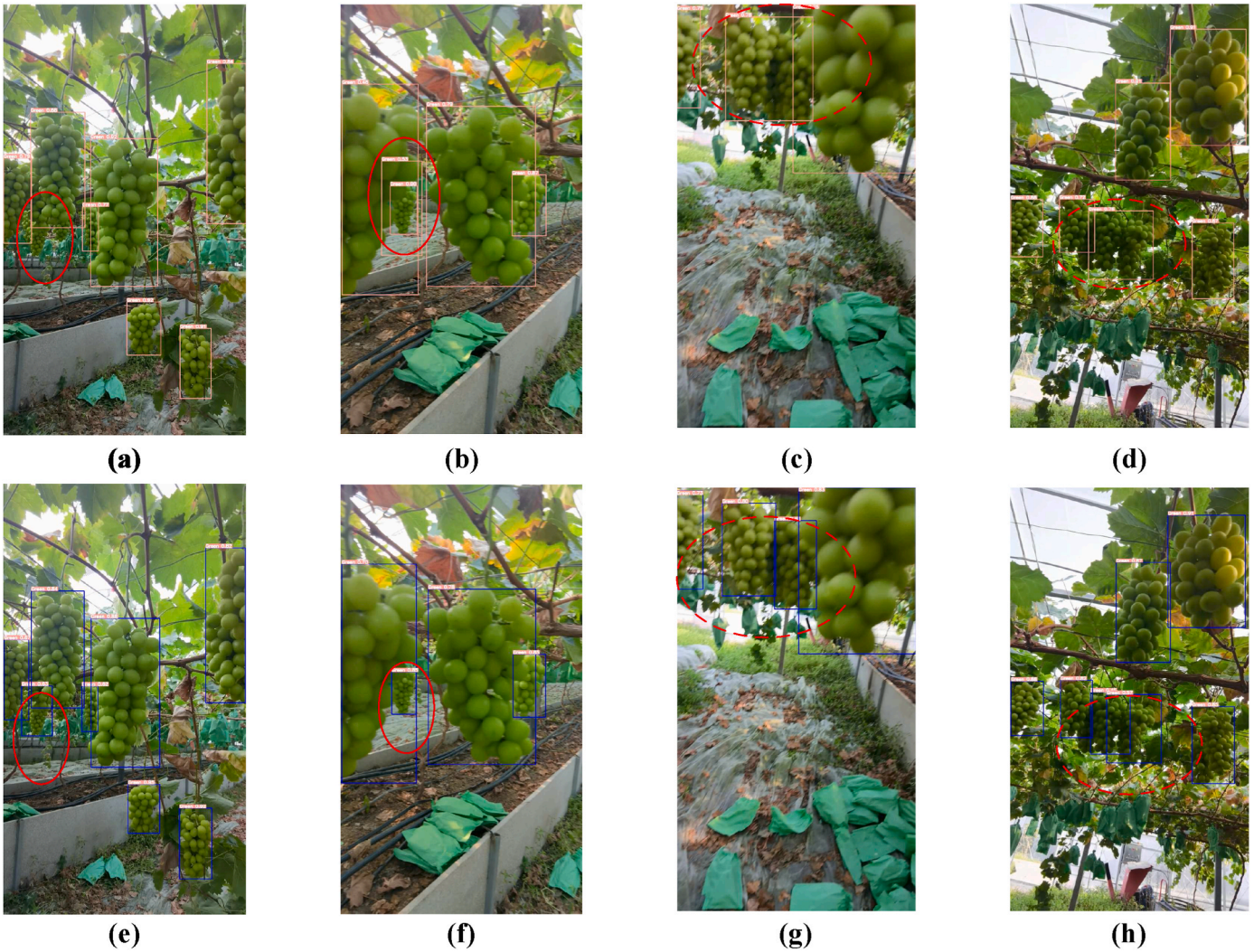

Fig. 8. Detection results of YOLOv5s and ESP-YOLO. (a) and (b) YOLOv5s in distant shot scenarios; (c) and (d) YOLOv5s in overlap scenarios; (e) and (f) ESP-YOLO in distant shot scenarios; (g) and (h) ESP-YOLO in overlap scenarios.

as moving and harvesting operations. The frame rate of D435I camera is configured as 20 fps, which continuously transmits grape images after preprocessing as input for ESP-YOLO. Image comparison methods and masking of image operation are used to preprocess the images from D435I to improve detection speed. Finally, Edge detection in the depth images is used to identify picking points and the Visual Serving control method is employed to execute the harvesting tasks (Y.-R. Li et al., 2023; Yan et al., 2021).

ESP-YOLO and YOLOv5s are conducted on a harvesting robot in a grape greenhouse, which results are shown in Fig. 11, Fig. 12 and Table 9. The average detection time of the system is 45 ms.

The mAP@0.5:0.95 of ESP-YOLO is 5.7% higher than that of YOLOv5s due to a reduction in miss error detection in the solid box (Fig. 11). Fig. 12 demonstrates ESP-YOLO accurately recognises and locates the grape clusters in various scenarios including overlap, front light, backlight, distant shot et al. The table grapes are accurately identified and segmented, except for those far beyond the harvesting robot's workspace. Therefore, ESP-YOLO on an embedded platform can detect the table grapes accurately and quickly for the harvesting task of a robot.

4. Discussion

The aforementioned results demonstrate that the proposed method

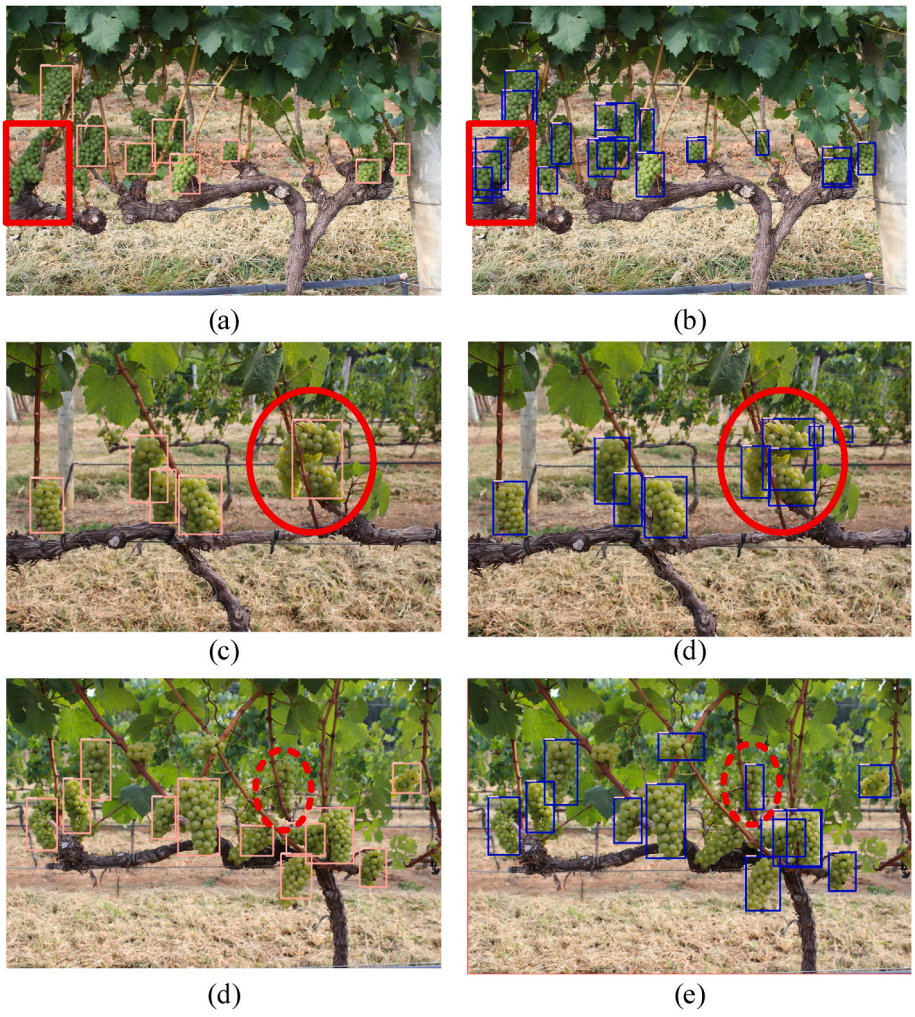

Fig. 9. Detection results of YOLOv5s and ESP-YOLO in public grape dataset.

Table 8
Comparison results of different methods.

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Inference time (ms)	Model size (MB)
Faster-RCNN	86.3	57.7	5030	108.2
SSD	88.5	55.7	6800	91.1
YOLOv5s	88.6	57.6	136	13.7
YOLOv5-Mobile	87.3	55.8	93.1	7.1
YOLOv7-Tiny	89.0	58.2	160.3	23.1
YOLOv8n*	88.4*	60.0*	**34.1***	**5.9**
YOLO-Grape	87.8*	61.0*	54.8*	28.7
YOLOv5-CFD	86.6*	54.2*	124*	33.0
SM-YOLOv4	88.2*	58.0*	103*	26.9
ESP-YOLO	**89.4(89.4*)**	**63.3(63.3*)**	**90**(36*)	11.2

Note: * The results are run on Jetson Xavier Nx.

in this study achieves high accuracy and inference speed. Subsequent sections will delve into the significance of these findings and explore their potential applications. To gain deeper insights into ESP-YOLO's performance, we conduct an extensive analysis of the experimental data with respect to detection accuracy and inference speed.

4.1. Evaluation in overlap scenarios

The efficiency and accuracy of grape harvesting robots primarily depend on the capability of grapes reconisation and location in their vision systems. Owing to the irregular shapes of grapes unliked the apples, mangoes, and kiwis. It is a challenging task for grape harvesting robots to detect grapes in overlapping scenarios. The attention mechanism and NMS are able to enhance the grapes identification in overlapping scenarios due to instance-level noise suppression. CBAM can extract spatial and channel features, which is utilised to identify overlapping grapes (Zhu et al., 2023). Another study utilised Soft-NMS and SE attention mechanisms to obtain a notable 91.08% mAP in grape detection (H. Li et al., 2021). However, these studies ignore the differences between the role of channel attention mechanisms and spatial attention mechanisms, which resulted in the lower accuracy in Table 8.

It is necessary for the methods to effectively extract textural and shape features because of the separation difficulty from the background with their similar green colour. In this study, SE and Soft Alpha_NMS are applied to segment the overlapping grapes. Table 10 demonstrates that the mAP@0.5 of ESP-YOLO surpasses that of YOLO-CFD and YOLO-Grape by 1.6%–4.2%. This improvement is attributed to the channel features extracted from the deep network and the spatial features extracted from the shallow network. Moreover, the results of the ablation experiments show that channel features combined with spatial features enhance the algorithm's ability to detect grapes (Table 5). The experiments demonstrate that the ESP-YOLO can mitigate the influence

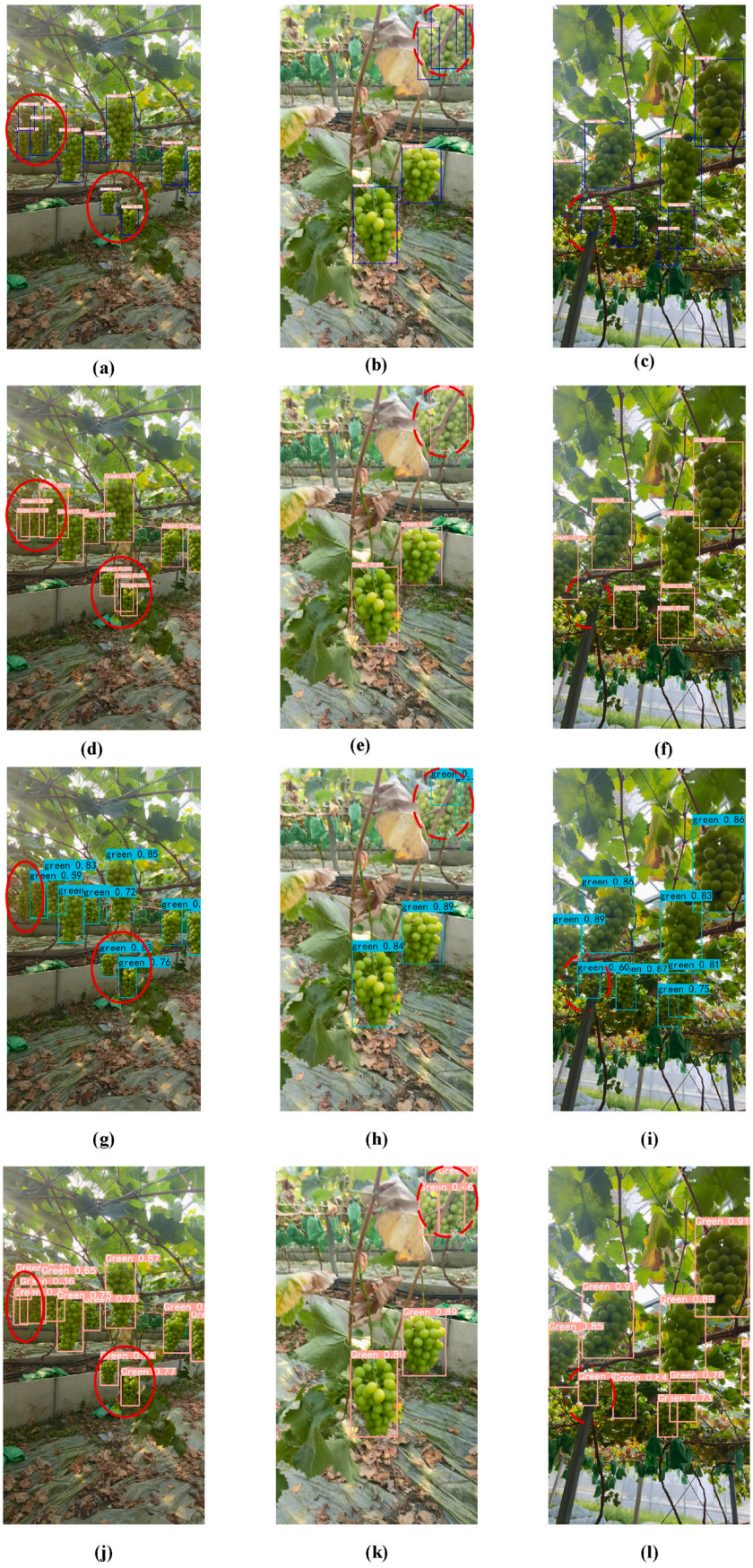

Fig. 10. Detection results under various complex scenarios with ESP-YOLO (a)–(c), YOLOv5-Mobile (d)–(f), YOLOv7-Tiny (g)–(i), YOLOv8n (j)–(l).

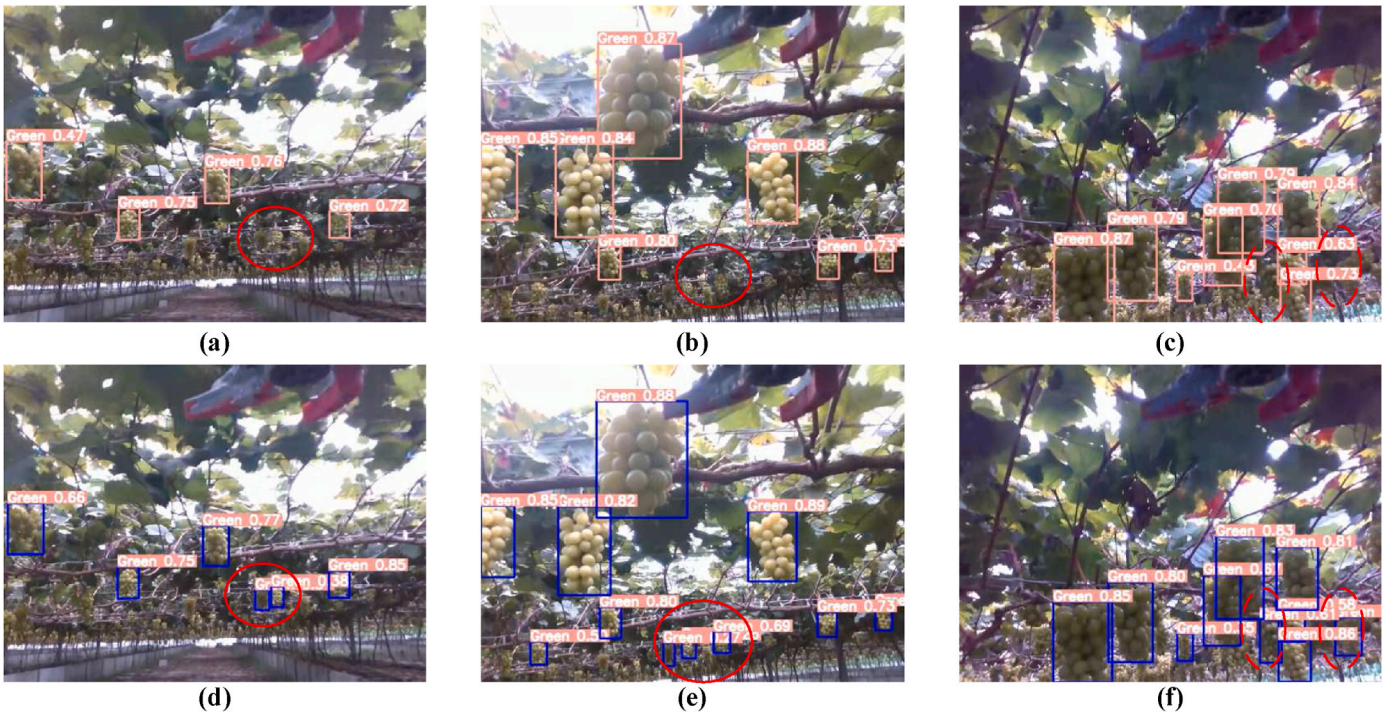

Fig. 11. Detection results of harvesting robot with YOLOv5s and ESP-YOLO. (a) and (b) YOLOv5s in distant shot scenarios; (c) YOLOv5s in overlap scenarios; (d) and (e) EP-YOLO in distant shot scenarios; (f) EP-YOLO in overlap scenarios.

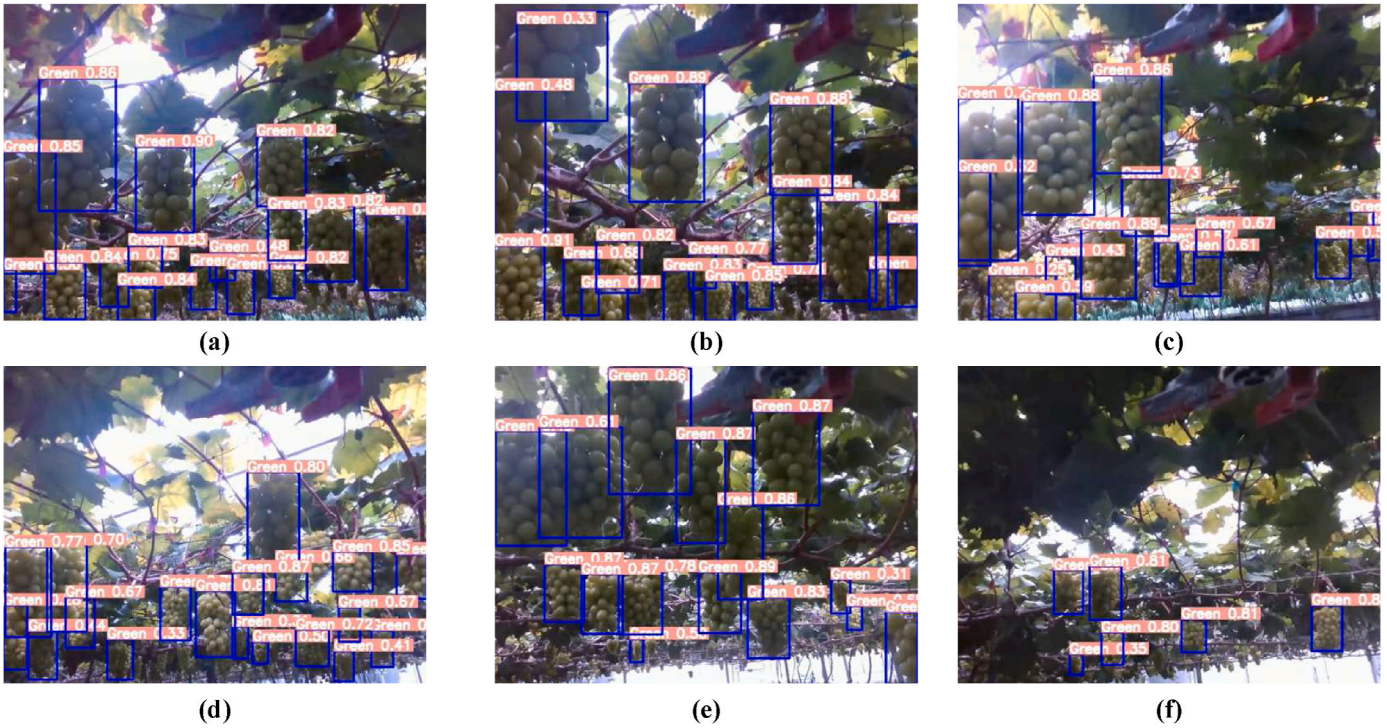

Fig. 12. Detection results of EP-YOLO under different real scenarios. (a) and (b) in overlap scenarios; (c) and (d) in front light scenarios; (e) in backlight scenarios; (f) in distant shot scenarios.

Table 9
Detection results of YOLOv5s, E-YOLO, EP-YOLO and ESP-YOLO.

Model	mAP@0.5(%)	mAP@0.5:0.95(%)	F1-score
YOLOv5s	92	59.2	0.90
ESP-YOLO	**96.3**	**62.6**	**0.93**

of target overlap and occlusion, overcome interference caused by target-background colour similarity, and effectively detect grapes in complex scenes.

Table 10
Detection results of YOLO-CFD, YOLO-Grape and ESP-YOLO in overlap scenarios.

Model	Precision(%)	Recall(%)	mAP@0.5(%)	mAP@0.5:0.95(%)
YOLO-CFD	85.8	84	87.5	49.6
YOLO-Grape	88.6	84.7	89.7	50.2
ESP-YOLO	**91.4**	**88.3**	**91.2**	**59.6**

4.2. Evaluation of the speed

Taking into account of the cost and portability, embedded platforms are used as the computing core for the agricultural machinery. Consequently, current research focuses on the improvement of detection speed. Ghost Net is integrated into YOLO to efficiently detect white asparagus with minimal computational resources (P. Zhang et al., 2024). Another study highlighted LES-YOLO's ability to identify pine cones at 115 frames per second on standard computers (Cui et al., 2023). These studies confirm the lightweight models' effectiveness such as Ghost Net in the FLOPs reduction. However, these studies ignore the effect of memory accesses on the inference speed (Ma et al., 2018). In this study, aiming to the poor performance of embedded platforms, the memory access is reduced by removing the transition layer in ELSAN. Table 3 shows the inference speed of Ghost Net is lower than that of ELSAN due to the decreased memory. The experiments prove that ESP-YOLO can improve the inference speed and quickly detect grapes on embedded systems.

4.3. Current deficiencies and future studies

Although promising results are found in the preliminary models for grapes detection exhibit, this study still has limitations. It is important to note that while these overlapping grapes are successfully identified, there is no guarantee of the identification accuracy in the situations of more than three clusters of overlapping grapes. Our future research is to reconstruct grape information via multiple recognitions with the moving camera installed in the robot arm to drop down the false detections in the complex scenarios. Furthermore, the reconstructed grape information can be applied for the picking point to correct the robot's trajectory, thereby improving the precision and speed of the harvesting robot.

5. Conclusions

In the study, ESP-YOLO based on ELSAN, SE and C3P is proposed to detect table grapes with higher accuracy, faster speed and lightweight in complex scenarios. ELSAN is superior to the Mobilenet and Ghostnet in the detection accuracy of table grapes and inference speed on embedded platforms. ELSAN can be used as a lightweight block to decrease FLOPs with high detection accuracy. C3P in the YOLO Neck can slightly improve the inference speed. Combined with C3P, SEC3 inserted between ELSAN and SPPF in YOLO can reduce instance-level noise effectively and improve detection accuracy. Soft Alpha_NMS can improve detection accuracy in dense scenarios.

Compared with the advanced detection methods, ESP-YOLO achieves higher detection accuracy and speed in complex scenarios, such as overlap, occlusion and distant shot. Furthermore, ESP-YOLO is validated on an embedded platform in a grape harvesting robot with a detection accuracy of 96.3% and a detection speed of 45 ms/image. Therefore, ESP-YOLO can be acted as a useful tool for harvesting robots to detect table grapes accurately and quickly in various complex scenarios.

CRediT authorship contribution statement

Jiaoliao Chen: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Conceptualization. **Huan Chen:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology. **Fang Xu:** Writing – original draft, Resources, Investigation, Data curation, Conceptualization. **Mengnan Lin:** Software, Investigation. **Dan Zhang:** Supervision, Resources, Data curation, Conceptualization. **Libin Zhang:** Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing 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.

Acknowledgment

This work was supported by Research on Public Welfare Technology Application Projects of Zhejiang Province (Grant no. LGG22E090003, Grant no. LGG18E050023) and National Key Research and Development Program of China (Grant no. 2018YFB1309404).

References

Abbaszadeh, R., Rajabipour, A., Mahjoob, M., Delshad, M., & Ahmadi, H. (2013). Evaluation of watermelons texture using their vibration responses. *Biosystems Engineering, 115*(1), 102–105. https://doi.org/10.1016/j.biosystemseng.2013.01.001

Bai, Y., Guo, Y., Zhang, Q., Cao, B., & Zhang, B. (2022). Multi-network fusion algorithm with transfer learning for green cucumber segmentation and recognition under complex natural environment. *Computers and Electronics in Agriculture, 194*. https://doi.org/10.1016/j.compag.2022.106789

Bai, Y., Yu, J., Yang, S., & Ning, J. (2024). An improved YOLO algorithm for detecting flowers and fruits on strawberry seedlings. *Biosystems Engineering, 237*, 1–12. https://doi.org/10.1016/j.biosystemseng.2023.11.008

Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017a). Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision* (pp. 5561–5569). https://doi.org/10.1109/ICCV.2017.593

Bulanon, D. M., Kataoka, T., Okamoto, H., & Hata, S.-i. (2004). Development of a real-time machine vision system for the apple harvesting robot. In *Paper presented at the SICE 2004 annual conference*.

Chaivivatrakul, S., & Dailey, M. N. (2014). Texture-based fruit detection. *Precision Agriculture, 15*(6), 662–683. https://doi.org/10.1007/s11119-014-9361-x

Chen, J., Kao, S.-h., He, H., Zhuo, W., Wen, S., Lee, C.-H., & Chan, S.-H. G. (2023). Run, don't walk: Chasing higher FLOPS for faster neural networks. In *Paper presented at the proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Cui, M., Lou, Y., Ge, Y., & Wang, K. (2023). LES-YOLO: A lightweight pinecone detection algorithm based on improved YOLOv4-tiny network. *Computers and Electronics in Agriculture, 205*. https://doi.org/10.1016/j.compag.2023.107613

Domingues, T., Brandao, T., & Ferreira, J. C. (2022). Machine learning for detection and prediction of crop diseases and pests: A comprehensive survey. *Agriculture-Basel, 12*(9). https://doi.org/10.3390/agriculture12091350

Egea, M. B., Bertolo, M. R. V., Oliveira Filho, J. G.d., & Lemes, A. C. (2021). A narrative review of the current knowledge on fruit active aroma using gas chromatography-olfactometry (GC-O) analysis. *Molecules, 26*(17). https://doi.org/10.3390/molecules26175181

Gu, B., Wen, C. J., Liu, X. Z., Hou, Y. J., Hu, Y. H., & Su, H. Q. (2023). Improved YOLOv7-tiny complex environment citrus detection based on lightweighting. *Agronomy-Basel, 13*(11). https://doi.org/10.3390/agronomy13112667

He, J., Erfani, S., Ma, X., Bailey, J., Chi, Y., & Hua, X. S. (2021). Alpha-IoU: A family of power intersection over union losses for bounding box regression. In *Paper presented at the 35th conference on neural information processing systems (NeurIPS), electr network*.

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M. X., … Ieee. (2019). Searching for MobileNetV3. Oct 27-Nov 02. In *Paper presented at the IEEE/CVF international conference on computer vision*. Seoul, SOUTH KOREA: ICCV. https://doi.org/10.1109/iccv.2019.00140.

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-Excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(8), 2011–2023. https://doi.org/10.1109/tpami.2019.2913372

Ji, W., Qian, Z., Xu, B., Tao, Y., Zhao, D., & Ding, S. (2016). Apple tree branch segmentation from images with small gray-level difference for agricultural harvesting robot. *Optik, 127*(23), 11173–11182. https://doi.org/10.1016/j.ijleo.2016.09.044

Jin, Y., Yu, C., Yin, J., & Yang, S. X. (2022). Detection method for table grape ears and stems based on a far-close-range combined vision system and hand-eye-coordinated picking test. *Computers and Electronics in Agriculture, 202*. https://doi.org/10.1016/j.compag.2022.107364

Khan, N., Fahad, S., Naushad, M., & Faisal, S. (2020). Grape production critical review in the world. https://doi.org/10.2139/ssrn.3595842.

Kuznetsova, A., Maleva, T., & Soloviev, V. (2020). Using YOLOv3 algorithm with pre- and post-processing for apple detection in fruit-harvesting robot. *Agronomy-Basel, 10*(7). https://doi.org/10.3390/agronomy10071016

Li, H., Li, C., Li, G., & Chen, L. (2021). A real-time table grape detection method based on improved YOLOv4-tiny network in complex background. *Biosystems Engineering, 212*, 347–359. https://doi.org/10.1016/j.biosystemseng.2021.11.011

Li, Y.-R., Lien, W.-Y., Huang, Z.-H., & Chen, C.-T. (2023). Hybrid visual servo control of a robotic manipulator for cherry tomato harvesting. *Actuators, 12*(6). https://doi.org/10.3390/act12060253

Li, B., Wang, M., & Wang, N. (2010). Development of a real-time fruit recognition system for pineapple harvesting robots. In *Paper presented at the 2010 Pittsburgh, Pennsylvania, june 20-june 23, 2010*. https://doi.org/10.13031/2013.29922

Linker, R., Cohen, O., & Naor, A. (2012). Determination of the number of green apples in RGB images recorded in orchards. *Computers and Electronics in Agriculture, 81*, 45–57. https://doi.org/10.1016/j.compag.2011.11.007

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot MultiBox detector. In *Paper presented at the 14th European conference on computer vision (ECCV)*. https://doi.org/10.1007/978-3-319-46448-0_2. Amsterdam, NETHERLANDS.

Luo, L., Tang, Y., Lu, Q., Chen, X., Zhang, P., & Zou, X. (2018). A vision methodology for harvesting robot to detect cutting points on peduncles of double overlapping grape clusters in a vineyard. *Computers in Industry, 99*, 130–139. https://doi.org/10.1016/j.compind.2018.03.017

Ma, N. N., Zhang, X. Y., Zheng, H. T., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In *Paper presented at the 15th European conference on computer vision (ECCV), Munich, Germany*. https://doi.org/10.1007/978-3-030-01264-9_8

Marin-San Roman, S., Fernandez-Novales, J., Cebrian-Tarancon, C., Sanchez-Gomez, R., Diago, M. P., & Garde-Cerdan, T. (2023). Monitorization of varietal aroma composition dynamics during ripening in intact Vitis vinifera L. Tempranillo blanco berries by hyperspectral imaging. *Journal of Agricultural and Food Chemistry*, 2616–2627. https://doi.org/10.1021/acs.jafc.2c074252616

Okamoto, H., & Lee, W. S. (2009). Green citrus detection using hyperspectral imaging. *Computers and Electronics in Agriculture, 66*(2), 201–208. https://doi.org/10.1016/j.compag.2009.02.004

Payne, A. B., Walsh, K. B., Subedi, P. P., & Jarvis, D. (2013). Estimation of mango crop yield using image analysis - segmentation method. *Computers and Electronics in Agriculture, 91*, 57–64. https://doi.org/10.1016/j.compag.2012.11.009

Qiang, L., Cai, J., Bin, L., Lie, D., & Zhang, Y. (2014). Identification of fruit and branch in natural scenes for citrus harvesting robot using machine vision and support vector machine. *International Journal of Agricultural and Biological Engineering, 7*(2), 115–121. https://doi.org/10.3965/j.ijabe.20140702.014

Qiu, C., Tian, G., Zhao, J., Liu, Q., Xie, S., & Zheng, K. (2022). Grape maturity detection and visual pre-positioning based on improved YOLOv4. *Electronics, 11*(17). https://doi.org/10.3390/electronics11172677

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*. https://doi.org/10.1109/TPAMI.2016.2577031

Safren, O., Alchanatis, V., Ostrovsky, V., & Levi, O. (2007). Detection of green apples in hyperspectral images of apple-tree foliage using machine vision. *Transactions of the Asabe, 50*(6), 2303–2313. https://doi.org/10.13031/2013.24083

Santos, T. T., de Souza, L. L., dos Santos, A. A., & Avila, S. (2020). Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Computers and Electronics in Agriculture, 170*. https://doi.org/10.1016/j.compag.2020.105247

Shang, Y., Xu, X., Jiao, Y., Wang, Z., Hua, Z., & Song, H. (2023). Using lightweight deep learning algorithm for real-time detection of apple flowers in natural environments. *Computers and Electronics in Agriculture, 207*. https://doi.org/10.1016/j.compag.2023.107765

Steinbrener, J., Posch, K., & Leitner, R. (2019). Hyperspectral fruit and vegetable classification using convolutional neural networks. *Computers and Electronics in Agriculture, 162*, 364–372. https://doi.org/10.1016/j.compag.2019.04.019

Taghadomi-Saberi, S., & Hemmat, A. (2015). Improving field management by machine vision-A review. *Agricultural Engineering International: CIGR Journal, 17*(3).

Tang, J., Wang, D., Zhang, Z., He, L., Xin, J., & Xu, Y. (2017). Weed identification based on K-means feature learning combined with convolutional neural network. *Computers and Electronics in Agriculture, 135*, 63–70. https://doi.org/10.1016/j.compag.2017.01.001

Wang, C.-Y., Liao, H.-Y. M., & Yeh, I.-H. (2023). Designing Network Design Strategies Through Gradient Path Analysis. *Journal of Information Science and Engineering, 39*(3), 975–995. https://doi.org/10.6688/jise.202307.

Wouters, N., De Ketelaere, B., De Baerdemaeker, J., & Saeys, W. (2013). Hyperspectral waveband selection for automatic detection of floral pear buds. *Precision Agriculture, 14*(1), 86–98. https://doi.org/10.1007/s11119-012-9279-0

Wu, Y. S., Duan, S. Y., Zhao, L. P., Gao, Z., Luo, M., Song, S. R., … Wang, S. P. (2016). Aroma characterization based on aromatic series analysis in table grapes. *Scientific Reports, 6*. https://doi.org/10.1038/srep31116

Xiong, J., He, Z., Lin, R., Liu, Z., Bu, R., Yang, Z., … Zou, X. (2018). Visual positioning technology of picking robots for dynamic litchi clusters with disturbance. *Computers and Electronics in Agriculture, 151*, 226–237. https://doi.org/10.1016/j.compag.2018.06.007

Xiong, J., He, Z., Tang, L., Lin, R., & Liu, Z. (2017). Visual localization of disturbed grape picking point in non-structural environment. *Transactions of the Chinese Society for Agricultural Machinery, 48*(4), 29–33. https://doi.org/10.6041/j.issn.1000-1298.2017.04.003, 81.

Xu, P., Fang, N., Liu, N., Lin, F., Yang, S., & Ning, J. (2022). Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation. *Computers and Electronics in Agriculture, 197*. https://doi.org/10.1016/j.compag.2022.106991

Yan, J., Wang, P., Wang, T., Zhu, G., Zhou, X., & Yang, Z. (2021). Identification and localization of optimal picking point for truss tomato based on mask r-cnn and depth threshold segmentation. In *Paper presented at the 2021 IEEE 11th annual international conference on CYBER Technology in automation, control, and intelligent systems (CYBER)*. https://doi.org/10.1109/CYBER53097.2021.9588274

Yang, G. L., Wang, J. X., Nie, Z. L., Yang, H., & Yu, S. Y. (2023). A lightweight YOLOv8 tomato detection algorithm combining feature enhancement and attention. *Agronomy-Basel, 13*(7). https://doi.org/10.3390/agronomy13071824

Zeng, T., Li, S., Song, Q., Zhong, F., & Wei, X. (2023). Lightweight tomato real-time detection method based on improved YOLO and mobile deployment. *Computers and Electronics in Agriculture, 205*. https://doi.org/10.1016/j.compag.2023.107625.

Zhang, P., Dai, N., Liu, X., Yuan, J., & Xin, Z. (2024). A novel lightweight model HGCA-YOLO: Application to recognition of invisible spears for white asparagus robotic harvesting. *Computers and Electronics in Agriculture, 220*, Article 108852.

Zhang, Q., & Gao, G. (2020). Prioritizing robotic grasping of stacked fruit clusters based on stalk location in RGB-D images. *Computers and Electronics in Agriculture, 172*. https://doi.org/10.1016/j.compag.2020.105359

Zhang, Y., Gao, J., Cen, H., Lu, Y., Yu, X., He, Y., & Pieters, J. G. (2019). Automated spectral feature extraction from hyperspectral images to differentiate weedy rice and barnyard grass from a rice crop. *Computers and Electronics in Agriculture, 159*, 42–49. https://doi.org/10.1016/j.compag.2019.02.018

Zhou, R., Damerow, L., Sun, Y. R., & Blanke, M. M. (2012). Using colour features of cv. 'Gala' apple fruits in an orchard in image processing to predict yield. *Precision Agriculture, 13*(5), 568–580. https://doi.org/10.1007/s11119-012-9269-2

Zhu, Y., Li, S., Du, W., Du, Y., Liu, P., & Li, X. (2023). Identification of table grapes in the natural environment based on an improved Yolov5 and localization of picking points. *Precision Agriculture, 24*(4), 1333–1354. https://doi.org/10.1007/s11119-023-09992-w