


Article

Intelligent Detection of Lightweight “Yuluxiang” Pear in Non-Structural Environment Based on YOLO-GEW

Rui Ren ^{1,2}, Haixia Sun ^{1,2}, Shujuan Zhang ^{1,2,*}, Ning Wang ³ , Xinyuan Lu ^{1,2}, Jianping Jing ^{1,2}, Mingming Xin ^{1,2} and Tianyu Cui ^{1,2}

¹ College of Agricultural Engineering, Shanxi Agricultural University, Jinzhong 030801, China

² Dryland Farm Machinery Key Technology and Equipment Key Laboratory of Shanxi Province, Taigu 030801, China

³ Department of Biosystems and Agricultural Engineering, Oklahoma State University, 111 Agricultural Hall, Stillwater, OK 74078, USA

* Correspondence: zsj2021@sxau.edu.cn; Tel.: +86-139-3549-1091

Abstract: To detect quickly and accurately “Yuluxiang” pear fruits in non-structural environments, a lightweight YOLO-GEW detection model is proposed to address issues such as similar fruit color to leaves, fruit bagging, and complex environments. This model improves upon YOLOv8s by using GhostNet as its backbone for extracting features of the “Yuluxiang” pears. Additionally, an EMA attention mechanism was added before fusing each feature in the neck section to make the model focus more on the target information of “Yuluxiang” pear fruits, thereby improving target recognition ability and localization accuracy. Furthermore, the CIoU Loss was replaced with the WIoUv3 Loss as the loss function, which enhances the capability of bounding box fitting and improves model performance without increasing its size. Experimental results demonstrated that the enhanced YOLO-GEW achieves an F1 score of 84.47% and an AP of 88.83%, while only occupying 65.50% of the size of YOLOv8s. Compared to lightweight algorithms such as YOLOv8s, YOLOv7-Tiny, YOLOv6s, YOLOv5s, YOLOv4-Tiny, and YOLOv3-Tiny; there are improvements in AP by 2.32%, 1.51%, 2.95%, 2.06%, 2.92%, and 5.38% respectively. This improved model can efficiently detect “Yuluxiang” pears in non-structural environments in real-time and provides a theoretical basis for recognition systems used by picking robots.

Keywords: “Yuluxiang” pear; non-structural environments; lightweight; YOLO-GEW



Citation: Ren, R.; Sun, H.; Zhang, S.; Wang, N.; Lu, X.; Jing, J.; Xin, M.; Cui, T. Intelligent Detection of Lightweight “Yuluxiang” Pear in Non-Structural Environment Based on YOLO-GEW. *Agronomy* **2023**, *13*, 2418. <https://doi.org/10.3390/agronomy13092418>

Academic Editor: Juncheng Ma

Received: 20 August 2023

Revised: 14 September 2023

Accepted: 18 September 2023

Published: 20 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Currently, smart orchards have emerged as an innovative agricultural production method. Among them, fruit harvesting plays a pivotal role in smart orchards as it directly impacts the quality and storage effectiveness of the fruits [1,2]. “Yuluxiang” pears, which are Chinese geographical indication protected products renowned for their wide adaptability and high nutritional value, have successfully been exported to multiple countries [3–5]. However, these pears are highly seasonal fruits with a short optimal harvesting window that requires timely picking. Currently, manual picking is predominantly relied upon for “Yuluxiang” pears in a non-structural environment, resulting in labor-intensive operations and low efficiency. With the continuous advancement of information technology, achieving intelligent picking robots for “Yuluxiang” pears to replace manual operations holds significant importance. Therefore, in order to enhance harvesting efficiency and improve quality aspects, fruit detection methods with superior stability, rapid detection speed, and precise recognition rates would be more advantageous.

The orchard environment is typically non-structural. When conducting fruit picking operations, “Yuluxiang” pears exhibit color characteristics that resemble leaves and tree canopies. Factors such as variations in lighting conditions, complex weather situations, and intricate backgrounds pose certain challenges for intelligent recognition of “Yuluxiang”

pears in non-structured environments. With the advancement of smart agriculture processes and the development of deep learning techniques, convolutional neural networks efficiently extract fruit image information within orchard environments. These networks integrate feature extraction with selection and classification into a unified model that ensures real-time detection while maintaining accuracy. Among them, the YOLO series of networks transforms object detection problems into regression problems and utilizes data from different positions on collected images to accomplish candidate box extraction and target recognition classification [6–9]. In recent years, it has been widely applied in fruit recognition in orchards [10–12]. The achievements of fruit target detection based on the YOLO algorithm are shown in Table 1.

Table 1. The achievements of fruit target detection based on the YOLO algorithm.

Fruit Type	Research Method	Purpose of Research	Research Achievement	Reference
Strawberry	YOLOv3	Utilized an enhanced YOLOv3 recognition approach to continuously detect and identify strawberries in complex environments.	Exhibiting robustness against occlusion, overlap, and density with a mAP of 87.51%. The accuracy rate for identifying ripe strawberries reached 97.14% with a recall rate of 94.46%, while the accuracy rate for identifying unripe strawberries was 96.51% with a recall rate of 93.61%.	[13]
Blueberry	YOLOv4-Tiny	Proposed integrating CBAM into the feature pyramid structure of the target detection network (I-YOLOv4-Tiny) to recognize blueberries at different maturity levels.	Achieving fruit detection accuracy of up to 96.24% within only an average detection time of 5.72 milliseconds. The memory size occupied by the network structure is merely 24.20 MB, satisfying both high precision requirements and ensuring fast response.	[14]
Jujube	YOLOv5s	Proposed L-YOLOv5s-RCA, a lightweight convolutional neural network, for jujube recognition.	The network incorporates detection layers of varying scales and the BiFPN structure to enhance detection accuracy, while introducing the Dual Coordinate Attention module for efficient operations. The model achieves a mAP of 97.2% with a size of 7.1 MB.	[15]
Citrus	YOLOv5s	Enhanced the visual saliency detection model by integrating it with YOLOv5s for the identification of citrus fruits in natural environments.	This approach achieves a mAP of 95.4% with a single image detection time of 70 ms.	[16]
Tomato	YOLOv8s	Addresses the limited level of automation in tomato harvesting within agriculture and proposes a method for automatic tomato detection based on YOLOv8s, DSCConv, DPAG, and FEM fusion techniques.	Test results demonstrate a mAP of 93.4%, while reducing the model size to 16 MB and achieving a real-time detection speed of 138.8 FPS.	[17]

Table 1. Cont.

Fruit Type	Research Method	Purpose of Research	Research Achievement	Reference
Apple	YOLOv5	Further improved the YOLOv5 architecture by combining feature pyramid networks and data augmentation, enabling it to detect smaller targets, improve feature quality in complex backgrounds, and achieve effective detection of apples.	The research findings demonstrated exceptional precision, recall, and F1 scores of 0.97, 0.99, and 0.98 respectively.	[18]
Green passion	YOLOv5	Conducted a study on green passion fruit and proposed a lightweight real-time detection model, MbECA-v5, specifically designed for complex environments.	The model achieved an average precision mean of 88.3%, computational complexity of 6.6 GFLOPs, volume of only 6.41 MB, and a real-time detection speed of 10.92 f/s on embedded devices.	[19]

However, despite the high accuracy and fewer parameters offered by improved target detection algorithms based on YOLO for fruits with similar colors in complex environments, they have not yet demonstrated effective results for “Yuluxiang” pears or efficient identification in non-structured environments using resource-limited picking robots.

In conclusion, this study focuses on the detection of “Yuluxiang” pears in non-structural orchards. To address challenges such as large model size and low detection accuracy in natural environments, a lightweight pear detection model called YOLO-GEW is proposed. This method is based on YOLOv8s as the base model and optimized using GhostNet as the backbone network to reduce computational complexity and improve detection accuracy. Additionally, an EMA attention mechanism is introduced in the neck part to enhance feature fusion capability. Furthermore, by utilizing WIoUv3 Loss as an improved loss function, the model’s fitting ability is enhanced. The proposed YOLO-GEW can effectively detect “Yuluxiang” pear fruits while reducing parameters and model size.

2. Materials and Methods

2.1. Construction of Data Set

This study focuses on the research of “Yuluxiang” pear fruits, and the collected images are from an experimental base of the Pomology Institute located in Jinzhong City, Shanxi Province. “Yuluxiang” pear is a hybrid of “Kuerle” pear as female and Snowflake pear as male. It belongs to the tree plant of the rose family. The cultivation mode is high stem open heart with soil deep application of organic fertilizer, expanding holes to improve soil permeability, and the implementation of inter-row grass. The average fruit weight is 250 g, and the yield per mu is 2000–3000 kg. The TNY-AL00 model camera with a focal length of 35 mm and a resolution of 3456 × 3456 pixels was utilized for capturing. The shooting period ranged from 8:00 to 18:00, while maintaining a distance range between the equipment and the fruits at 10–50 cm. During the image collection process, various weather conditions, time periods, and backgrounds were considered when capturing images of “Yuluxiang” pears. A total of 1050 experimental sample images were collected and divided into training set (734 images), validation set (211 images), and test set (105 images) in a ratio of 7:2:1. The Labelling tool was employed to annotate the minimum bounding rectangle for identifying “Yuluxiang” pears as “pear”. The dataset is illustrated in Figure 1.

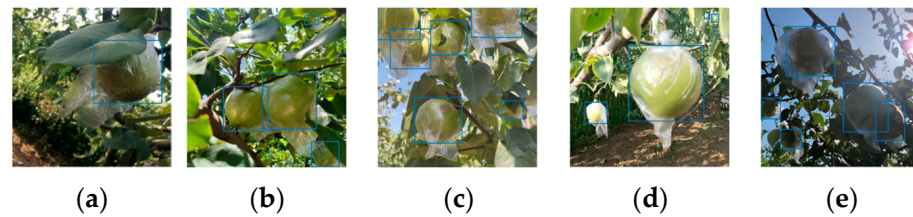


Figure 1. “Yuluxiang” pear dataset. (a) leaf occlusion; (b) fruit overlap; (c) density; (d) front lighting; (e) backlighting. The blue bounding boxes represent the true box of “Yuluxiang” pear manually calibrated.

2.2. YOLOv8 Model and Performance Comparison

The YOLOv8 model, which has been enhanced and refined based on previous iterations, showcases exceptional performance and versatility. It comprises four key components: the input module, backbone feature extraction network, neck network, and detection module. The input module encompasses functionalities such as image input, data augmentation, and adaptive anchor box calculation. The backbone feature extraction network leverages Conv + Bn + SiLU (CBL), CSPLayer_2Conv (C2F), and spatial pyramid pooling-fast (SPPF) structures to extract features from the input images. The neck network adopts the path aggregation network (PAN) structure, which enhances its ability to fuse features of objects at different scales. Concat indicates that it concatenates itself according to a certain dimension, which is usually used to merge two feature maps, upstands for deconvolution upsampling. The detection end decouples the processes of classification and detection, primarily including loss calculation and target detection box filtering. VFL Loss is utilized for classification loss, while DFL Loss + CIoU Loss is employed for regression loss. The obtained feature maps are decoded and predicted to output the class of detected targets and generate bounding boxes. YOLOv8 controls the depth and number of layers through adjustable parameters such as width (w), depth (d), and ratio (r). This model consists of five network structures with varying sizes: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. The network structure of YOLOv8 is shown in Figure 2.

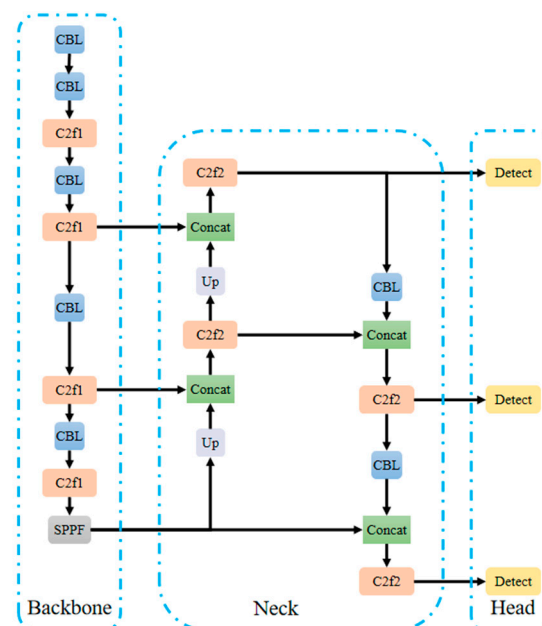


Figure 2. YOLOv8 basic network model structure.

The YOLOv8 model was initially evaluated on the constructed dataset to detect accurately and efficiently “Yuluxiang” pears in non-structural environments. Comparing the performance of the base models (YOLOv8x, YOLOv8l, and YOLOv8m), Table 2 demon-

strates minimal disparity in mAP values, with a maximum achievement of 87.72% by YOLOv8m. Additionally, among the five YOLOv8 models, YOLOv8n exhibits the smallest size (only 5.95 MB). Maintaining consistent layers, depth, and ratio comparisons with YOLOv8n, Yolov8s showcases an improved mAP by 0.99%. Despite its slightly lower mAP compared to that of YOLOV8m (a difference of only 1.21%), its model size is merely 21.48 MB (equivalent to just 43.29% of the size of YOLOV8m). Furthermore, when contrasted with both YOLOv8x and YOLOv8l variants, it is noteworthy that YOLOv8s reduces layer count by as much as 140.

Table 2. Performance comparison of YOLOv8 base models.

Model	d (Depth)	w (Width)	r (Ratio)	AP (%)	Model Size (MB)	Layer
YOLOv8n	0.33	0.25	2.0	85.52	5.95	225
YOLOv8s	0.33	0.50	2.0	86.51	21.48	225
YOLOv8m	0.67	0.75	1.5	87.72	49.62	295
YOLOv8l	1.00	1.00	1.0	87.51	83.60	365
YOLOv8x	1.00	1.25	1.0	87.40	130.39	365

To further monitor the dynamic trends of network training, Figure 3 displays the classification loss curve, bounding box loss, and mAP curve of five YOLOv8 models. The convergence and stabilization of these three curves after 100 epochs indicate that the models have achieved good training results without overfitting. It is worth noting that this study focuses solely on detecting target “Yuluxiang” pears in non-structural environments while considering both real-time performance and accuracy of model detection. Improvements were made based on the YOLOv8s model.

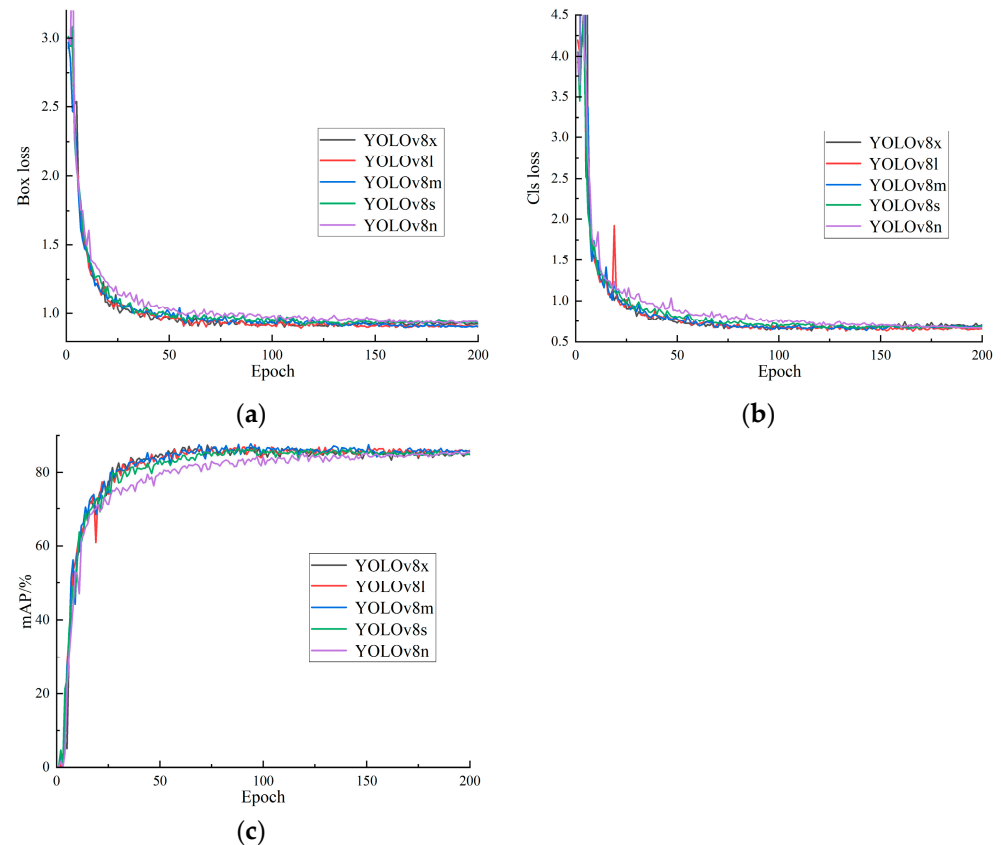


Figure 3. YOLOv8 network training trends. (a) Box loss; (b) Cls loss; (c) mAP.

2.3. Improved “Yuluxiang” Pear Detection Model

2.3.1. Lightweight Model

The computational costs of deep neural networks are typically high due to the inclusion of a large number of convolutional layers. In 2020, Huawei Noah’s Ark Lab proposed the GhostNet lightweight network in order to achieve model lightweighting and enhance detection efficiency. This network employs ghost modules to generate an equal number of feature maps as regular convolutional layers, which are then substituted with the original convolutional layers to reduce computational costs [20]. The ghost module compresses the input feature layer through non-linear convolution operations, followed by linear convolution operations that process the feature maps layer by layer, resulting in another set of feature maps. Finally, these two sets of feature maps are combined to obtain new feature maps.

Ghost bottleneck is obtained by stacking ghost modules. In the ghost bottleneck structure, the first ghost module acts as an expansion layer to increase the number of channels, while the second ghost module reduces the number of channels to match the shortcut path. The ghost bottleneck includes two bottleneck modules with different strides: a stride 1 bottleneck is used to expand the channel count, while a stride 2 bottleneck is used to reduce the channel count and maintain consistency with the output when connected. This design helps reduce model computation.

The GhostNet network is primarily composed of ghost bottleneck stacks, with the ghost module serving as building blocks. The first layer consists of a standard convolutional layer with 16 filters, followed by a series of gradually increasing ghost bottlenecks. These ghost bottlenecks are divided into different stages based on the size of their input feature maps. Except for the last bottleneck in each stage, which has a stride of 2, all other ghost bottlenecks have an applied stride of 1.

2.3.2. EMA Attention Mechanism

The attention mechanism is a technique that facilitates the learning process of the network model by assigning varying weights to different segments of input data, thereby enabling the model to prioritize important information and enhance its performance while mitigating overfitting scenarios [21–23]. Efficient multi-scale attention (EMA) transforms certain channels into batch dimensions and group channel dimensions into multiple sub-features, effectively preserving channel-specific information while reducing computational costs [24], thus ensuring an equitable distribution of spatial semantic features within each feature group.

The EMA model employs three parallel pathways to extract attention weight descriptors for grouping feature maps. Among these pathways, two are 1×1 branches, while the third pathway is a 3×3 branch. In the 1×1 branches, two 1D global average pooling operations are utilized to encode channels along two spatial directions. As for the 3×3 branch, it only utilizes one stacked 3×3 kernel to capture multi-scale feature representations. Additionally, EMA incorporates cross-spatial learning to aggregate cross-space information from different spatial dimensions, thereby achieving more comprehensive feature aggregation.

2.3.3. WIoU Loss Function

The loss function of bounding box regression (BBR) plays a crucial role in object detection, and its precise definition significantly enhances the performance of the model. In YOLOv8’s architecture, DFL Loss + CIoU Loss is adopted as the regression loss. However, CIoU only considers the overlapping area between predicted boxes and ground truth boxes, neglecting the intermediate region which may introduce biased evaluation results. Weighted intersection over union (WIoU) addresses this potential bias issue by incorporating weights for the region between predicted boxes and ground truth boxes [25]. WIoU has three versions: WIoUv1 constructs a boundary box loss based on an attention mechanism; WIoUv2 designs a monotonic static focus mechanism (FM); and WIoUv3 utilizes dynamic

non-monotonic FM for gradient gain allocation strategy. The spatial relationship between the ground truth box and the predictive box is shown in Figure 4.

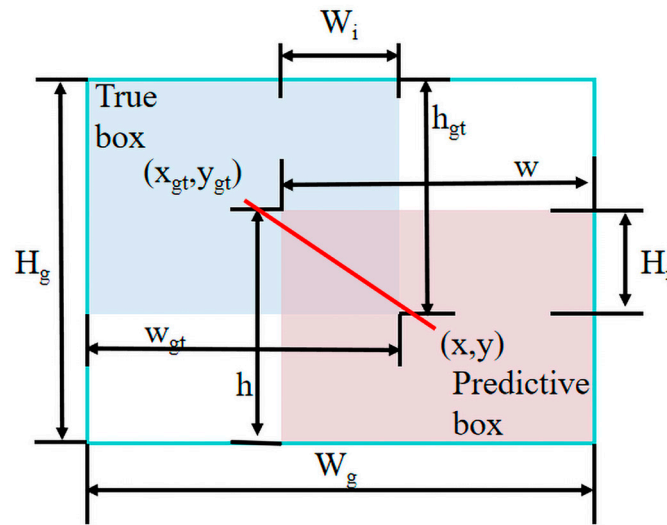


Figure 4. The spatial relationship between the ground truth box and the predictive box. W_g and H_g represent the width and height of the minimum rectangle formed by the predictive box and true box, while w and h represent the width and height of the predictive box. w_{gt} and h_{gt} denote the width and height of the true box, while W_i and H_i respectively indicate the width and height of the overlapping rectangle between predictive box and true box. The red line represents the distance between the center points of the two boxes.

The calculation formula for $WIoUv1$ is shown in Equations (1)–(3):

$$L_{IoU} = 1 - IoU = 1 - \frac{W_i H_i}{wh + w_{gt} h_{gt} - W_i H_i} \tag{1}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{2}$$

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{3}$$

where r stands for gradient gain, $r = L^{\gamma*}_{IoU} \in [0; 1]$. During the model’s training, the gradient gain decreases with the decrease of L_{IoU} , resulting in a slow convergence rate in the late stages of training. Therefore, the mean of L_{IoU} is introduced as the normalizing factor. The calculation formula for $WIoUv2$ is shown as Equation (4):

$$L_{WIoUv2} = \left(\frac{L_{IoU}}{\overline{L_{IoU}}}\right)^\gamma L_{WIoUv1} \tag{4}$$

$$r = \left(\frac{L^*_{IoU}}{\overline{L_{IoU}}}\right)^\gamma \tag{5}$$

where $\overline{L_{IoU}}$ is the running mean with momentum m . Dynamically updating the normalizing factor keeps the gradient gain r at a high level overall, which solves the problem of slow convergence in the late stages of training.

The abnormality degree of the anchor box β is represented by the ratio of L^*_{IoU} and $\overline{L_{IoU}}$, as shown in Formula (6):

$$\beta = \frac{L^*_{IoU}}{\overline{L_{IoU}}} \in [0, +\infty) \tag{6}$$

The non-monotonic focus coefficient was constructed using β and applied to WIoUv1 to obtain WIoUv3, as shown in Equation (7):

$$L_{WIoUv3} = rL_{WIoUv1}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \quad (7)$$

where the mapping of outlier degree is β and gradient gain is r , which is controlled by the hyper-parameters α and δ .

Since $\overline{L_{IoU}}$ is dynamic, the quality demarcation standard of anchor boxes is also dynamic, which allows WIoU v3 to make the gradient gain allocation strategy that is most in line with the current situation at every moment. The training process adopts an α value of 1.9 and a δ value of 3 to effectively allocate smaller gradient gains to anchor boxes with lower quality, thereby enhancing the fitting ability of the bounding box loss function for improved model performance.

2.4. Test Platform

The research was conducted on a Windows 10 operating system, equipped with an Intel Core i5-10400H CPU @ 2.90 GHz, 32 GB RAM, a 1 T hard drive, and a NVIDIA GeForce RTX 3060 GPU with a memory of 12,288.0 MB. The software used during the experiment primarily consisted of python version 3.8.8, torch version 1.12.1, CUDA version 11.3, and CUDNN version 8.6 for deep learning computations in pycharm2021.

The network input image size was $640 \times 640 \times 3$ pixels, stochastic gradient descent optimization (SGD) was used and the parameters were updated using a stochastic gradient descent optimizer with a momentum of 0.9. The initial learning rate is set to 0.01 with weight decay of 0.0005. Batch-size is 8, and 200 epochs are trained. All models use the same dataset and training parameters.

2.5. Evaluation Indicators

The evaluation of the model's performance in this study employed commonly used metrics for object detection algorithms, namely *F1* score and average precision (*AP*). Both *F1* and *AP* are measures that assess precision and recall, with specific calculation formulas shown in Equations (8)–(11):

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$AP = \int_0^1 Precision(Recall)d(Recall) \quad (10)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

where, the term *TP* represents the accurate detection of "Yuluxiang" pears by the model. *FP* indicates false detections, where a pear is mistakenly identified as "Yuluxiang". *FN* denotes undetected instances of "Yuluxiang" pears.

Additionally, complexity metrics such as floating point operations per second (FLOPs), model layers, and parameter count are computed to assess computational complexity. The size of the model is evaluated based on its dimensions.

3. Results and Analysis

3.1. Test Results and Analysis of Different Lightweight Backbone Feature Network Models

In order to lighten the network and improve the model's feature extraction ability, this study replaced the backbone networks of the YOLOv8s model with lightweight models GhostNet, MobileNetv3 [26], and FasterNet [27] for experimental analysis. The results are shown in Table 3. After optimization, the three lightweight models exhibit certain

improvements in accuracy and F1 score compared to YOLOv8s, while simultaneously reducing model size and FLOPs. Among them, GhostNet demonstrates superior precision and AP performance, with a 4.73% and 1.33% increase respectively over YOLOv8s, while maintaining a model size that is only 64.85% of YOLOv8s. Although MobileNetv3 has a smaller model size and computational complexity compared to YOLOv8s, it experiences a slight decrease in AP by 0.5%. FasterNet achieves the highest accuracy and F1 score among the four models, with the lowest computational complexity of 17.2 G. The replacement of GhostNet and FasterNet with YOLOv8s as the backbone improves the lightweight nature of the model and its ability to extract features like “Yuluxiang” pears. In terms of detection accuracy, GhostNet’s F1 score is only 0.19% lower than that of FasterNet, while GhostNet’s AP is 0.4% higher than that of FasterNet. In terms of lightweight networks, despite being only 0.4 G higher in FLOPs compared to FasterNet, GhostNet has a smaller model size by 2.77 MB. In consideration of the lightweight nature of the model and the feature extraction capability of the backbone network, this study selected GhostNet as the backbone for YOLOv8s to extract features of “Yuluxiang” pears, naming it YOLO-G. This model not only reduces model parameters but also suppresses useless features to improve model efficiency, laying a foundation for subsequent research.

Table 3. Test results of backbone feature networks of different lightweight models.

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)	Model Size (MB)	FLOPs (G)
YOLOv8s	86.59	77.12	81.58	86.51	21.48	28.5
GhostNet	91.32	76.52	83.28	87.84	13.93	17.6
MobileNetv3	88.73	77.22	82.58	86.01	14.93	21.7
FasterNet	87.51	79.79	83.47	87.44	16.70	17.2

3.2. Effects of Adding Attention Mechanism to Neck on Model Performance

In order to enhance the feature fusion capability of the lightweight YOLO-G model for “Yuluxiang” pear, this study introduced an EMA attention mechanism prior to fusing features in the neck section of the model. This modification enables the model to prioritize target information relevant to “Yuluxiang” pear fruit, thereby enhancing target recognition and localization accuracy. Four EMA modules are added at positions shown in Figure 5 in the neck.

The effectiveness of EMA in enhancing model detection performance was validated through a comparative analysis with YOLO-G under identical experimental conditions. Furthermore, the impacts of various attention mechanisms, such as the convolutional block attention module (CBAM) [28], coordinate attention (CA) [29], squeeze and excitation (SE) [30], along with the EMA attention module, on feature fusion within the model were thoroughly examined. The results are shown in Table 4.

According to Table 4, the YOLO-G model with the added CBAM attention module shows no significant change in F1 and AP performance, however it results in a slight increase in model size and parameters. This suggests that the CBAM attention module has minimal impact on the performance of YOLO-G. With the addition of the SE attention module, there is a 0.33% improvement in F1 score but a 0.26% decrease in AP score. Amongst the four added attention mechanisms, CA has the smallest increase in model size and parameters, with an increase of only 0.1 MB and 0.04 M, respectively, achieving improvements in recall, F1, and AP compared to YOLO-G. The EMA attention mechanism achieves the highest F1 score at 84.50% and an AP value of 88.43%. Adding the EMA attention mechanism only increases model size and parameters by 0.04 MB and 0.03 M respectively, resulting in higher precision, F1 score, and AP compared to adding CA attention mechanism by 1.17%, 0.3%, and 0.28%, respectively. Experiments show that adding EMA attention mechanism to lightweight models’ neck part can effectively improve detection performance without significantly increasing model size or parameters.

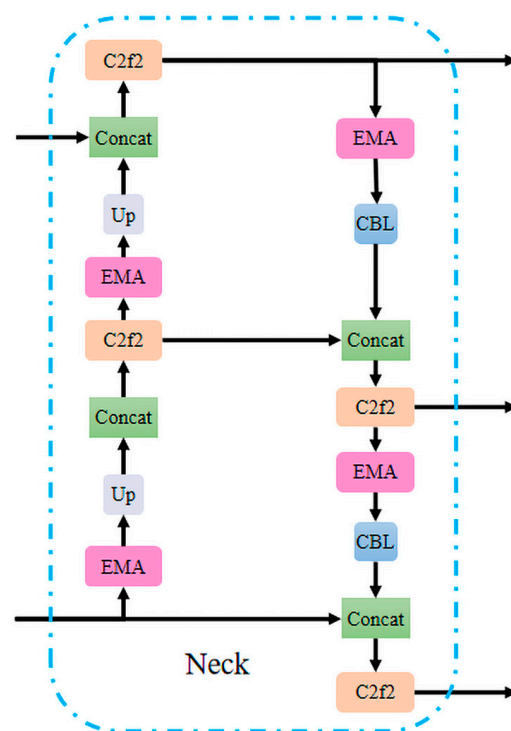


Figure 5. Location of EMA module addition in neck.

Table 4. Influences of different attention mechanisms on model feature fusion.

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)	Model Size (MB)	Parameters (M)
YOLO-G	91.32	76.52	83.28	87.84	13.93	7.12
CBAM	89.01	78.22	83.27	87.82	14.72	7.53
SE	88.81	78.98	83.61	87.58	14.03	7.17
CA	89.33	79.62	84.20	88.15	14.03	7.16
EMA	90.50	79.24	84.50	88.43	14.07	7.19

3.3. Ablation Test

In order to verify the rationality of the improved YOLOv8s model, an ablation experiment was conducted to validate the effectiveness of each improvement point. The results of the ablation experiment are shown in Table 5. By improving the backbone to GhostNet on the original YOLOv8s model, it is defined as YOLO-G model. In the YOLO-G neck, the EMA attention mechanism is added and defined as YOLO-GE model. The model that improves the loss function in YOLO-G to WIoUv3 is defined as YOLO-GW, and finally, all improvement points are fused together and defined as YOLO-GEW.

Table 5. Results of ablation test.

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)	Model Size (MB)
YOLOv8s	86.59	77.12	81.58	86.51	21.48
YOLO-G	91.32	76.52	83.28	87.84	13.93
YOLO-GE	90.50	79.24	84.50	88.43	14.07
YOLO-GW	90.44	78.72	84.17	88.36	13.93
YOLO-GEW	89.93	79.64	84.47	88.83	14.07

The results in Table 5 demonstrate the superior performance of the four enhanced models in terms of precision, F1 score, and AP. Among them, YOLO-G achieves the highest precision while maintaining a small model size. Furthermore, YOLO-G exhibits a 1.33% higher AP compared to YOLOv8s, confirming the effectiveness of replacing the backbone network for extracting “Yuluxiang” pear features and utilizing lightweight models. The model sizes of YOLO-GE and YOLO-GEW increased by only 0.14 MB compared to YOLO-G and YOLO-GEW, respectively. In comparison to YOLO-G, incorporating EMA attention mechanism into the neck layer of YOLO-GE leads to improvements in recall, F1 score, and AP; whereas adding the EMA attention mechanism further enhances recall, F1, and AP by 0.92%, 0.3%, and 0.47% respectively in the case of YOLO-GEW. This approach is feasible as it allows better integration of backbone networks for extracting “Yuluxiang” pear features without significantly increasing model size while enhancing target recognition ability and localization accuracy.

Compared to the YOLOv8s model, both YOLO-G and YOLO-GW achieved a reduction in size of 7.55 MB, while YOLO-GE and YOLO-GEW achieved a reduction of 7.41 MB. Furthermore, when compared to YOLO-G and YOLO-GE, both YOLO-GW and YOLO-GEW exhibited a decrease in accuracy by 0.88% and 0.57%, respectively, but demonstrated an increase in recall rate by 2.2% and 0.4%. Meanwhile, in comparison to YOLO-G, the utilization of the enhanced loss function WIoUv3 in the model resulted in a 2.59% improvement in F1 score. This signifies that without augmenting the model’s size, the improved loss function can enhance bounding box fitting capability and elevate model performance. In contrast to YOLOv8s, the upgraded model YOLO-GEW exhibited enhancements across all evaluation metrics, with a 2.32% increase in AP score and a reduction of 7.41 MB in model size. Figure 6 illustrates the P-R curve of the enhanced model, demonstrating a larger area under its curve compared to other models and indicating superior overall performance for YOLO-GEW while further validating the effectiveness of this approach.

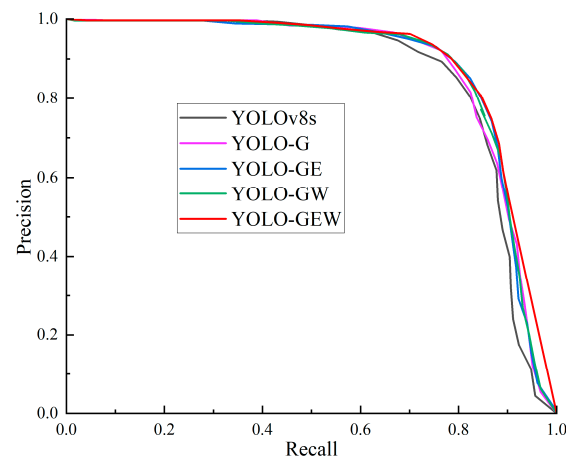


Figure 6. P-R curve of the improved model.

4. Discussion

Due to the similarity in color between the fruit of “Yuluxiang” pear and its leaves, tree canopy, and other background elements, as well as challenges in distinguishing it under various factors such as bagging, lighting conditions, and weather in natural environments, efficient and intelligent identification of “Yuluxiang” pears becomes a formidable task. Therefore, reducing the model size is advantageous for deployment and real-time picking. To further validate the enhanced model performance, this study compared YOLO-GEW with lightweight detection models YOLOv3-YOLOv8. The experimental results are shown in Table 6.

Table 6. Test results of “Yuluxiang” pear by different lightweight test models.

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)	Model Size (MB)	Parameters (M)	FLOPs (G)	Inference (ms)
YOLOv8s	86.59	77.12	81.58	86.51	21.48	11.13	28.5	5.8
YOLO-GEW	89.93	79.64	84.47	88.83	14.07	7.19	18.6	6.5
YOLOv7-Tiny	88.15	79.27	83.47	87.32	11.72	6.01	13.1	3.0
YOLOv6s	87.12	78.01	82.31	85.88	31.32	16.30	44.1	6.3
YOLOv5s	88.38	77.72	82.71	86.77	13.78	7.01	15.8	4.5
YOLOv4-Tiny	83.59	79.71	81.60	85.91	21.21	10.86	23.8	7.4
YOLOv3-Tiny	89.22	76.20	82.20	83.45	16.63	8.67	12.9	3.1

According to Table 6, the YOLO-GEW model proposed in this study is superior to other models in terms of precision, F1, and AP. Specifically, compared with YOLOv8s, YOLOv7-Tiny, YOLOv6s, YOLOv5s, YOLOv4-Tiny, and YOLOv3-Tiny, its AP is 2.32%, 1.51%, 2.95%, 2.06%, 2.92%, and 5.38% higher, respectively. It shows the excellence of the model in precision. In addition, YOLOv7-Tiny has the smallest model size and parameters, 2.35 MB and 1.18 M smaller than YOLO-GEW, respectively. The model size, parameters, and FLOPs of YOLO-GEW are all lower compared to those of YOLOv8s, YOLOv6s, and YOLOv4-Tiny. The model size and parameters of YOLOv5s are slightly smaller than those of YOLO-GEW at 0.29 MB and 0.18 M respectively. Among the seven lightweight models, YOLOv7-Tiny has the smallest model size and parameters, being 2.35 MB and 1.18 M smaller than YOLO-GEW, respectively; YOLOv3-Tiny has the lowest FLOPs at 12.9 G. Compared to YOLOv7-Tiny and YOLOv3-Tiny, YOLO-GEW shows an improvement of 1.78% and 0.71% in precision, a boost of 0.37% and 3.44% in recall, as well as an increase of 1.00% and 2.27% in F1 score. The inference time of YOLO-GEW is 6.5 ms, with the longest inference time for YOLOv4-Tiny being 7.4 ms, and the shortest for YOLOv7-Tiny being 3.0 ms. According to Figure 7, the test set demonstrates satisfactory detection performance for the seven lightweight object detection models. However, YOLOv4-Tiny exhibits a certain level of missed detections, while YOLOv7-Tiny and YOLOv3-Tiny display false detections in Figure 7d, potentially attributed to the similarity in colors between fruits and leaves as well as distant distances. After considering various factors, the YOLO-GEW proposed in this study effectively achieves a harmonious balance among model size, detection accuracy, and speed. This lays a solid foundation for deployment on embedded devices and provides technical support for the development of the “Yuluxiang” pear-picking robot.

In the field of pear fruit detection, Li et al. [31] utilized a ground tripod and cameras mounted on a drone platform to capture high-resolution images of pears for monitoring their growth status. They proposed an advanced multi-scale collaborative perception network known as YOLOv5sFP specifically designed for accurate pear detection. This model achieved an impressive average precision (AP) value of 96.12% while maintaining a compact model size of 50.01 MB. The introduction of Ma et al. [32] enhanced the performance of the YOLOv4 model in recognizing ‘Hongxiangsu’ pear fruit in natural environments by addressing challenges such as color similarity. This resulted in an improved detection mAP of 90.18% and a reduced model size of 136 MB. Inspired by YOLOv5, Sun et al. [33] proposed a model called YOLO-P for efficient and accurate detection of Akidzuki pears. The model achieved an AP of 97.6%, representing a 1.8% improvement over the original YOLOv5s. Moreover, the model also demonstrated volume compression, reducing its size from 13.7 MB to merely 8.3 MB, resulting in a compression rate of 39.4%. The aforementioned research primarily focused on conducting experimental studies on various pear varieties and yielded significant outcomes. It is worth noting that compared to “Yuluxiang” pears, the research samples consisted of untreated pears without bagging treatment, thereby posing less difficulty.

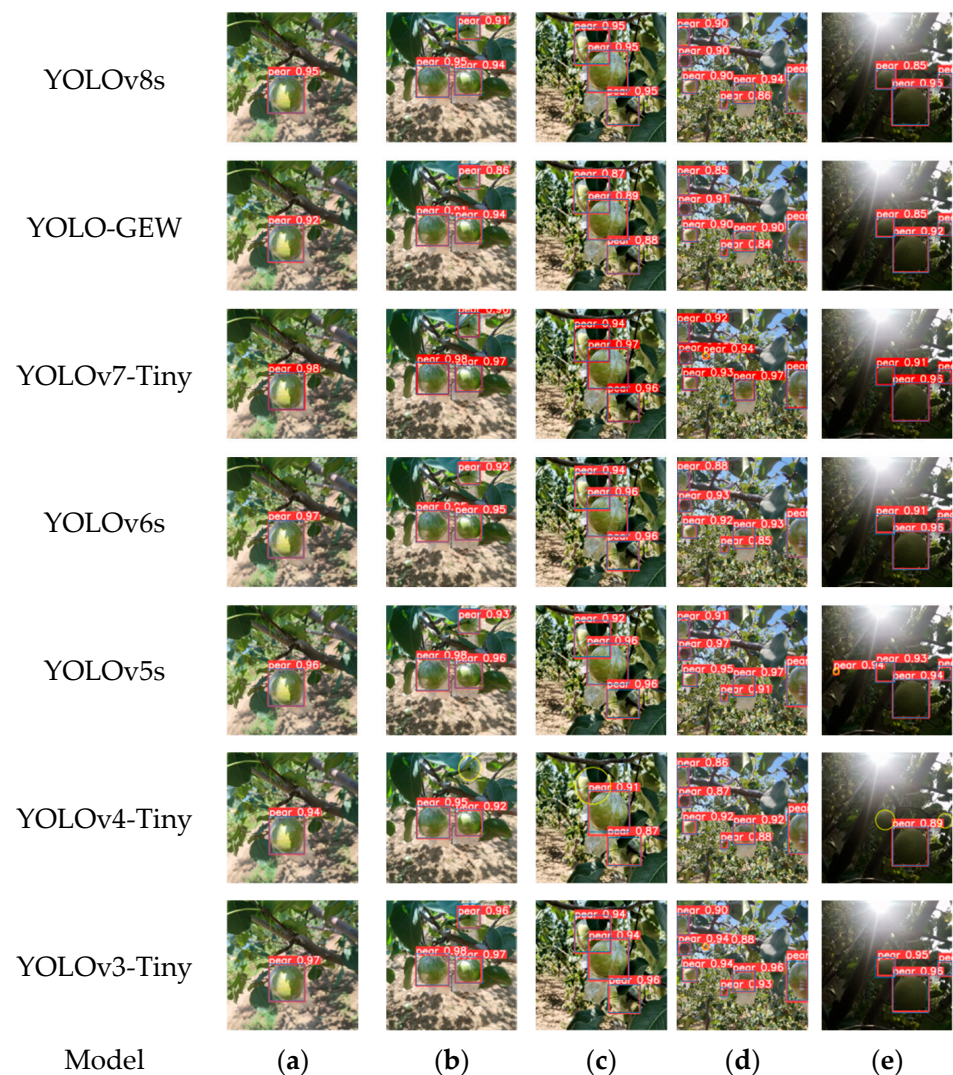


Figure 7. Detection results of different lightweight detection models in some test sets. (a) single fruit; (b) multiple fruits; (c) overlapping of fruits; (d) intensive fruits; (e) backlight. The red box represents predictive boxes, and the yellow circle represents missed or false detection.

The present study conducted research on the detection of “Yuluxiang” pears in a non-structural environment and successfully created an image dataset for “Yuluxiang” pears. The proposed lightweight detection model, YOLO-GEW, has a small size of only 14.07 MB but achieves an impressive average precision (AP) score of 88.83%, fully satisfying the current requirements for detection. In future research, it is imperative to employ more efficient cameras for image capture and collection of large-scale datasets encompassing diverse scenarios such as varying occlusions, lighting conditions, weather patterns, and fruit quantities. This will significantly enhance the robustness of the model validation. The primary focus should be on augmenting the detection speed of lightweight models on embedded devices and effectively deploying them in harvesting robots specifically tailored for “Yuluxiang” pear orchards.

5. Conclusions

In order to detect rapidly and accurately “Yuluxiang” pear fruits in non-structural environments, enhancements were made to the YOLOv8s model based on the “Yuluxiang” pear dataset. These enhancements addressed challenges such as fruit color similarity to leaves, fruit bagging, and complex surroundings. First, we substituted the original model’s backbone network with GhostNet to reduce parameters and enhance feature

extraction capabilities. Second, four EMA attention mechanisms were introduced in the neck section to emphasize the features of “Yuluxiang” pears and optimize feature fusion while effectively preventing overfitting. Finally, by replacing CIoU Loss with WIoUv3 Loss, we could strengthen the adaptability of bounding box loss function and improve model performance. In summary, by integrating these three improvement strategies, a lightweight “Yuluxiang” pear detection model YOLO-GEW, suitable for non-structural environments, was successfully constructed. The main conclusions of this study are as follows.

First, the performance of the YOLOv8 model was tested considering both real-time detection and accuracy. Based on this, further research was conducted using the YOLOv8s model. Then, in order to reduce network complexity and enhance feature extraction capability, experiments were carried out by replacing the backbone networks of YOLOv8s with lightweight models such as GhostNet, MobileNetv3, and FasterNet. Among these models, GhostNet achieved a maximum AP of 87.84% and had the smallest model size of 13.93 MB. It is named the YOLO-G model for subsequent research. Next, the effects of attention mechanisms CBAM, SE, CA, and EMA on the neck feature fusion ability were compared, and it was concluded that the addition of EMA could effectively improve the detection performance of the network model without significantly increasing model size and parameters. Finally, ablation experiments were designed to verify that the improved WIoUv3 loss function can strengthen the fitting ability of boundary frame loss and improve the performance of the model without increasing the model size. The precision, recall, F1, AP, and model size of YOLO-GEW are 89.93%, 79.64%, 84.47%, 88.83%, and 14.07 MB, respectively, which provides a foundation for deployment in embedded devices.

The improved algorithm’s performance was further validated by comparing the YOLO-GEW model with seven other models, namely YOLOv8s, YOLOv7-Tiny, YOLOv6s, YOLOv5s, YOLOv4-Tiny, and YOLOv3-Tiny. Results demonstrated that compared to the alternative models, the proposed method exhibited superior precision, F1 score, and AP. However, it had a slightly larger model size and more parameters than YOLOv7-Tiny. Moreover, in comparison to YOLOv3-Tiny, YOLO-GEW only increased the FLOPs by 6.9 G. Considering both accuracy and size considerations together suggests that utilizing our proposed method for real-time detection of “Yuluxiang” pear fruit in non-structured environments is advantageous.

Author Contributions: Conceptualization, R.R.; methodology, R.R. and H.S.; software, R.R., X.L., J.J. and M.X.; writing—original draft, R.R. and T.C.; writing—review and editing, R.R., N.W. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Science and Technology Innovation Fund Project of Shanxi Agricultural University (Project No: 2020BQ02). Research and Innovation Projects for Graduate Students in Shanxi Province (Project No: 2023KY303).

Data Availability Statement: The data are available from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mu, L.; Liu, H.; Cui, Y.; Fu, L.; Gejima, Y. Mechanized technologies for scaffolding cultivation in the kiwifruit industry: A review. *Inf. Process. Agric.* **2018**, *5*, 401–410. [[CrossRef](#)]
2. Montoya Cavero Luis, E.; Díaz de León Torres, R.; Gómez Espinosa, A.; Escobedo Cabello Jesús, A. Vision systems for harvesting robots: Produce detection and localization. *Comput. Electron. Agric.* **2021**, *192*, 106562. [[CrossRef](#)]
3. Sun, H.; Zhang, S.; Ren, R.; Su, L. Surface Defect Detection of “Yuluxiang” Pear Using Convolutional Neural Network with Class-Balance Loss. *Agronomy* **2022**, *12*, 2076. [[CrossRef](#)]
4. Yang, S.; Bai, M.; Hao, G.; Zhang, X.; Guo, H.; Fu, B. Transcriptome survey and expression analysis reveals the adaptive mechanism of ‘Yulu Xiang’ Pear in response to long-term drought stress. *PLoS ONE* **2021**, *16*, e0246070. [[CrossRef](#)] [[PubMed](#)]
5. Wu, X.; Shi, X.; Bai, M.; Chen, Y.; Li, X.; Qi, K.; Cao, P.; Li, M.; Yin, H.; Zhang, S. Transcriptomic and Gas Chromatography-Mass Spectrometry Metabolomic Profiling Analysis of the Epidermis Provides Insights into Cuticular Wax Regulation in Developing ‘Yuluxiang’ Pear Fruit. *J. Agric. Food Chem.* **2019**, *67*, 8319–8331. [[CrossRef](#)]

6. Joseph, R.; Santosh Kumar, D.; Ross, B.G.; Ali, F. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640.
7. Farhadi, A.; Redmon, J. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 1 July 2017; pp. 6517–6525. [[CrossRef](#)]
8. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
9. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *Comput. Vis. Pattern Recognit.* **2020**, *27*, 198–215. [[CrossRef](#)]
10. Fu, L.; Feng, Y.; Wu, J.; Liu, Z.; Gao, F.; Majeed, Y.; Al-Mallahi, A.; Zhang, Q.; Li, L.; Cui, Y. Fast and accurate detection of kiwifruit in orchard using improved YOLOv3-tiny model. *Precis. Agric.* **2020**, *22*, 754–776. [[CrossRef](#)]
11. Li, S.; Zhang, S.; Xue, J.; Sun, H.; Ren, R. A Fast Neural Network Based on Attention Mechanisms for Detecting Field Flat Jujube. *Agriculture* **2022**, *12*, 717. [[CrossRef](#)]
12. Luo, Q.; Rao, Y.; Jin, X.; Jiang, Z.; Wang, T.; Wang, F.; Zhang, W. Multi-Class on-Tree Peach Detection Using Improved YOLOv5s and Multi-Modal Images. *Smart Agric.* **2022**, *4*, 84–104. [[CrossRef](#)]
13. Liu, X.; Fan, C.; Li, J.; Gao, Y.; Zhang, Y.; Yang, Q. Identification Method of Strawberry Based on Convolutional Neural Network. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 237–244. [[CrossRef](#)]
14. Wang, L.; Qin, M.; Lei, J.; Wang, X.; Tan, K. Blueberry maturity recognition method based on improved YOLOv4-Tiny. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 170–178. [[CrossRef](#)]
15. Li, S.; Zhang, S.; Xue, J.; Sun, H. Lightweight target detection for the field flat jujube based on improved YOLOv5. *Comput. Electron. Agric.* **2022**, *202*, 107391. [[CrossRef](#)]
16. Chen, S.; Xiong, J.; Jiao, J.; Xie, Z.; Huo, Z.; Hu, W. Citrus fruits maturity detection in natural environments based on convolutional neural networks and visual saliency map. *Precis. Agric.* **2022**, *23*, 1515–1531. [[CrossRef](#)]
17. Yang, G.; Wang, J.; Nie, Z.; Yang, H.; Yu, S. A Lightweight YOLOv8 Tomato Detection Algorithm Combining Feature Enhancement and Attention. *Agronomy* **2023**, *13*, 1824. [[CrossRef](#)]
18. Sekharamanthy, P.K.; Melgani, F.; Malacarne, J. Deep Learning-Based Apple Detection with Attention Module and Improved Loss Function in YOLO. *Remote Sens.* **2023**, *15*, 1516. [[CrossRef](#)]
19. Lou, Z.; Li, P.; Song, F.; Sun, Q.; Ding, H. Lightweight Passion Fruit Detection Model Based on Embedded Device. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 262–322. [[CrossRef](#)]
20. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589. [[CrossRef](#)]
21. Zhang, C.; Zhu, L.; Yu, L. Review of Attention Mechanism in Convolutional Neural Networks. *Comput. Eng. Appl.* **2021**, *57*, 64–72. [[CrossRef](#)]
22. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, 8 June 2018; pp. 421–429. [[CrossRef](#)]
23. Sun, L.; Hu, G.; Chen, C.; Cai, H.; Li, C.; Zhang, S.; Chen, J. Lightweight Apple Detection in Complex Orchards Using YOLOv5-PRE. *Horticulturae* **2022**, *8*, 1169. [[CrossRef](#)]
24. Ouyang, D.; He, S.; Zhan, J.; Guo, H.; Huang, Z.; Luo, M.; Zhang, G. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023. [[CrossRef](#)]
25. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IOU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. In Proceedings of the Computer Vision and Pattern Recognition, Oxford, UK, 24 January 2023. [[CrossRef](#)]
26. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324. [[CrossRef](#)]
27. Chen, J.; Kao, S.H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, DC, USA, 10 November 2023. [[CrossRef](#)]
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [[CrossRef](#)]
29. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 4 March 2021. [[CrossRef](#)]
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018. [[CrossRef](#)]
31. Li, Y.; Rao, Y.; Jin, X.; Jiang, Z.; Wang, Y.; Wang, T.; Wang, F.; Luo, Q.; Liu, L. YOLOv5s-FP: A Novel Method for In-Field Pear Detection Using a Transformer Encoder and Multi-Scale Collaboration Perception. *Sensors* **2022**, *23*, 30. [[CrossRef](#)] [[PubMed](#)]

32. Ma, S.; Zhang, Y.; Zhou, G.; Liu, B. Recognition of pear fruit under natural environment using an improved YOLOv4 model. *J. Hebei Agric. Univ.* **2022**, *45*, 105–111. [[CrossRef](#)]
33. Sun, H.; Wang, B.; Xue, J. YOLO-P: An efficient method for pear fast detection in complex orchard picking environment. *Front Plant Sci* **2022**, *13*, 1089454. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.