IMPROVED YOLOv8N-BASED DETECTION OF GRAPES IN ORCHARDS
/
基于改进 YOLOv8n 的果园葡萄检测方法

Shan TAO, Shiwei WEN, Guangrui HU, Yahao GE, Jingming WEN, Xiaoming CAO, Jun CHEN*)
College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling 712100, China
Tel: +86-13572191773; E-mail: chenjun_jdxy@nwsuaf.edu.cn
DOI: https://doi.org/10.35633/inmateh-74-42

Keywords: *Fresh table grapes, Automatic detection, Algorithm lightweight design, YOLOv8n*

ABSTRACT

To address the issues of low detection accuracy, slow speed, and large parameter size in detecting fresh table grapes in natural orchard environments, this study proposes an improved grape detection model based on YOLOv8n, termed YOLOGPnet. The model replaces the C2f module with a Squeeze-and-Excitation Network V2 (SENetV2) to enhance gradient flow through more branched cross-layer connections, thereby improving detection accuracy. Additionally, the Spatial Pyramid Pooling with Enhanced Local Attention Network (SPPELAN) substitutes the SPPF module, enhancing its ability to capture multi-scale information of the target fruits. The introduction of the Focaler-IoU loss function, along with different weight adjustment mechanisms, further improves the precision of bounding box regression in object detection. After comparing with multiple algorithms, the experimental results show that YOLOGPnet achieves an accuracy of 93.6% and mAP@0.5 of 96.8%, which represents an improvement of 3.5 and 1.6 percentage points over the baseline model YOLOv8n, respectively. The model's computational load, parameter count, and weight file size are 6.8 Gflops, 2.1 M, and 4.36 MB, respectively. The detection time per image is 12.5 ms, showing reductions of 21.84%, 33.13%, 30.79%, and 25.60% compared to YOLOv8n. Additionally, comparisons with YOLOv5n and YOLOv7-tiny in the same parameters reveal accuracy improvements of 0.7% and 1.9%, respectively, with other parameters also showing varying degrees of enhancement. This study offers a solution for accurate and rapid detection of table grapes in natural orchard environments for intelligent grape harvesting equipment.

摘要

针对自然果园环境下鲜食葡萄的检测精度低、速度慢、参数量较大等问题，本研究提出了一种基于改进 YOLOv8n 的葡萄检测模型（YOLOGPnet）。该模型使用压缩与激励网络（Squeeze-and-Excitation Network V2，SENetV2）替换了 C2f 模块，通过更多的分支跨层连接使梯度流更加丰富，提高模型的检测精度；并将 SPPF 模块替换为增强局部注意力的空间金字塔池化网络（Spatial Pyramid Pooling with Enhanced Local Attention Network，SPPELAN），提升了网络捕捉目标果实的多尺度信息的能力；通过使用 Focaler-IoU 损失函数，和引入不同的权重调整机制提高了目标检测中的边界框回归精度问题。试验结果表明，YOLOGPnet 的精确度和 mAP@0.5 分别为 93.6%、96.8%，相较于 YOLOv8n，分别提高了 3.5 和 1.6 个百分点。该模型的计算量、参数量和权重文件大小分别为 6.8 Gflops、2.1 M 和 4.36 MB，单幅图像检测耗时为 12.5 ms，相较于 YOLOv8n，分别降低了 21.84%、33.13%、30.79% 和 25.60%。该研究为智能化葡萄采摘装备在自然果园环境下准确且快速地检测鲜食葡萄提供了一种解决方案。

INTRODUCTION

According to the Food and Agriculture Organisation of the United Nations (in full English, FAO), the global production of grapes reached about 80.1 × 10^8 kg in 2022 *(Khan N et al., 2021).* Grape harvesting, as a labour-intensive operation, is challenged by global labour resource constraints, and harvesting robotics is becoming increasingly important in grape growing *(Zhao et al., 2023)*. Traditional target detection in dense berry class mainly relies on colour, shape and texture features, and with the development of deep learning technology, a large number of deep learning-based target detection methods with high accuracy and robustness have emerged *(Ying et al., 2023).*

Over the years, researchers worldwide have extensively studied machine vision technology for fruit and vegetable target recognition and picking point localization *(Song et al., 2023)*. *Lu et al. (2021)* proposed the Swin-T-YOLOv5 model to detect grape clusters at different growth stages. *Zhao et al. (2022)* introduced an improved YOLOv4 method for predicting grape cluster picking points.

Wu et al. (2023) developed the Ghost-HRNet model, integrating object detection and key point localization to focus on grape peduncle positioning. *Ning et al. (2021)* innovatively used an improved Mask R-CNN to select optimal picking points at the horizontal central positions near the critical centroid of the peduncle area. *Zhang et al. (2023)* employed YOLOv5 GAP to detect green grape clusters effectively in densely grown and shaded environments. *Su et al. (2022)* proposed a lightweight grape detection method by integrating feature maps of different resolutions. *Wang et al. (2020)* introduced the SwinGD model for visual recognition of grape clusters. *Cha et al. (2021)* replaced Faster R-CNN's backbone with VGG16 to achieve accurate detection of Red Globe grapes in natural environments. *Zhu et al. (2021)* improved the YOLOX-Tiny model to detect red and green grape clusters. *Sun et al. (2023)* proposed the MRWYOLOv5s model, achieving a mAP of 97.74%, an improvement of 2.32% over the original model. *Li et al. (2021)* developed the YOLO grape model for detecting grape clusters of various colors, achieving an F1-score of 90.93% for green grapes and an average F1-score of 91.42%. *Zhao et al. (2022)* designed a lightweight end-to-end YOLO-GP model with integrated picking point prediction. *Cha et al. (2023)* employed transfer learning for Red Globe grape detection in natural settings. *Zhang et al. (2023)* utilized YOLOv5 GAP for accurate detection of densely grown grape clusters. *Liu et al. (2024)* proposed the YOLOX-RA model for fast and precise detection of densely grown and occluded grape clusters. *Lu et al. (2022)* developed the Swin-Transformer-YOLOv5 model, achieving 97% detection accuracy under cloudy conditions. *Guo et al. (2023)* introduced the YOLO y4+ model, which enhanced robustness in unstructured environments using a parameter-free attention mechanism. *Zhang et al. (2022)* developed the Grape-Internet dataset, improving detection efficiency through lightweight processing. *Qiu et al. (2022)* enhanced detection speed with an improved SM-YOLOv4 algorithm, achieving a detection time of 10.82 ms. *Yang et al. (2024)* proposed the YOLOv8s-grape detection method, significantly improving mAP and detection efficiency. *Jiang et al. (2024)* introduced the YOLOv8n-GP model, effectively enhancing feature extraction for grape stems.

In addition to its extensive application in grape harvesting, the YOLO series of algorithms has been widely used for target recognition and pest detection in other fruits and vegetables, providing valuable insights for improving recognition algorithms in this study. *Wen et al. (2024)* proposed a lightweight detection model based on an improved YOLOv8 network, incorporating partial convolution (Pconv) blocks to enhance apple detection under occlusion and varying lighting conditions. *Chen et al. (2024)* modified the YOLOv5 backbone by adding Transformer modules with attention mechanisms, replacing the original PAFPN Neck with a bidirectional weighted fusion BiFPN structure, and integrating a P2 shallow downsampling module in the Head structure. These modifications improved the accuracy of apple detection in natural environments by 3.7%. *Zhao et al. (2022)* conducted detection experiments on melon fruits using YOLOv3, YOLOv4, YOLOv5s, and an improved ResNet_YOLO model, finding YOLOv5 to perform best and demonstrating the feasibility of mixed detection for images of four Cucurbitaceae fruits. Ren et al. *(2024)* introduced MSCI-YOLOv8s, which enhanced the model's ability to capture multi-scale disease features in grape leaf images and achieved a real-time detection efficiency of 37.2 ms.

These studies have made significant progress in grape cluster target detection and picking point localisation, providing strong technical support to further improve fruit and vegetable recognition accuracy, but in the actual grape picking work it is necessary to ensure the recognition accuracy of fruit targets as well as to achieve the lightweight of the visual model, so this study improves and compares the existing models from these two aspects.

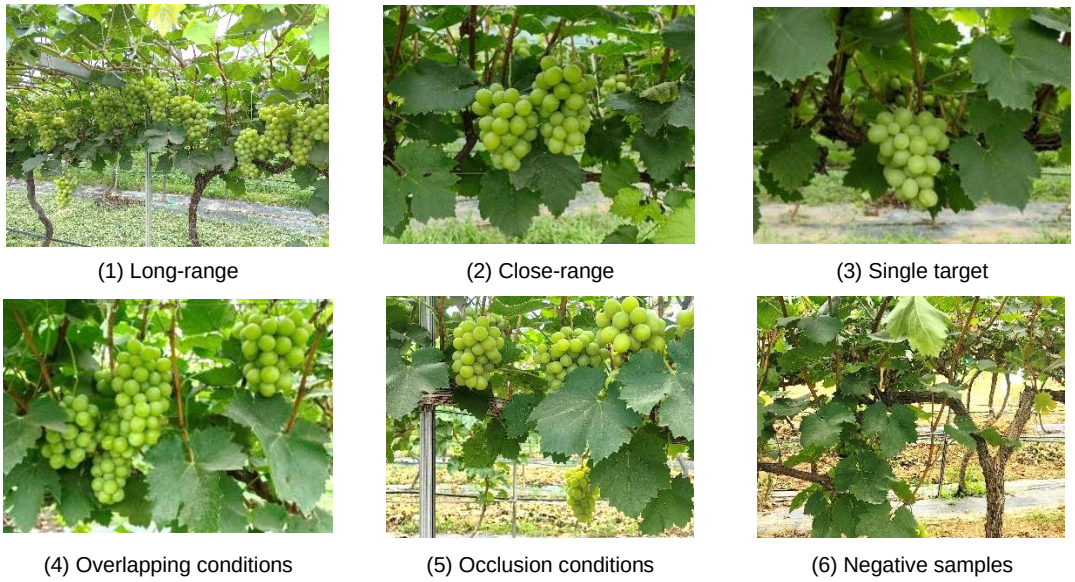
MATERIALS AND METHODS

Construction of grape image data set in orchard environment

Image acquisition

The grape fruit images in this dataset were captured between August 20th and 23rd, 2023, at the Caoxin Research Base in Yangling District, Xianyang City, Shaanxi Province. Due to the challenge of distinguishing green grapes (such as Shine Muscat, Bijou, and Zuijinxiang) from the similarly colored leaves, green grape varieties were selected as the experimental subjects. Based on varying shooting distances, the images were categorized into close-range (0.5m), mid-range (1.2m), and long-range (2m) shots, and were captured using a HUAWEI P60 smartphone. A total of 1,536 images were collected under different lighting conditions and in complex environments, of which 1,100 images were used for training, 200 for validation, and 236 for testing. The training set was used for model training and parameter tuning, the validation set for optimizing the network structure, and the test set for evaluating the model's generalization capability.

The images were uniformly cropped to a resolution of 1240×1240, and resized to 640×640 pixels during training to ensure efficient inference and accuracy. The dataset includes grape clusters of various forms, with up to 20 clusters appearing in long-range images, and close-range images presenting cases of overlapping and occlusion.



(1) Long-range (2) Close-range (3) Single target

(4) Overlapping conditions (5) Occlusion conditions (6) Negative samples

Fig. 1 - Grape fruit collection image example

The sample images were manually annotated using the LabelImg software, where the grape cluster regions were labeled with minimum enclosing bounding boxes. This process generated XML files in VOC format, extracting information such as the coordinates of the center point, bounding box width, and height, which were then saved as TXT label files. The entire image dataset was divided into training, validation, and test sets. A diverse dataset enhances the model's generalization ability and robustness, while also improving the model's adaptability to different scales. On the self-constructed dataset, YOLOv8 employed various data augmentation techniques, including mosaic augmentation, mixup augmentation, random perturbation, and color distortion, effectively expanding the dataset size.

Detection model based on improved YOLOv8

YOLOGpnet

YOLOv8 is categorized into five models—n, s, m, l, and x—designed for different application scenarios. As the model depth increases, detection accuracy improves. YOLOv8n, with the smallest number of parameters, offers the fastest detection speed. To ensure real-time performance, this study focuses on enhancing the YOLOv8n model. The architecture consists of four components: the Input, Backbone, Neck, and Head.

As shown in Figure 2, YOLOGPnet replaces the C2f structure of the baseline model with SENetV2 in the Backbone. C2f is an improved version of the C3 structure in YOLOv5, whereas SENetV2 enriches the gradient flow through more branched cross-layer connections, enhancing feature representation capabilities. It aims to improve recognition accuracy by optimizing spatial feature extraction and channel-level representation. Additionally, the SPPF module is replaced by the SPPELAN module, allowing the network to better adapt to input images of varying sizes, capture multi-scale information, and improve feature map expression and object detection performance. In the Neck network, a Path Aggregation Network (PAN) is employed to enhance feature fusion for objects at different scales. The Head network decouples the classification and detection processes and is mainly responsible for loss calculation and bounding box selection. Loss computation includes positive and negative sample assignment strategies and the calculation of the loss function, with the regression branch incorporating Distribution Focal Loss (DFLoss) and Complete Intersection over Union Loss (CIOULoss). YOLOGPnet replaces CIOULoss with the Focaler-IoU loss function, improving the accuracy of bounding box prediction by optimizing class imbalance and bounding box regression through a weight adjustment mechanism.

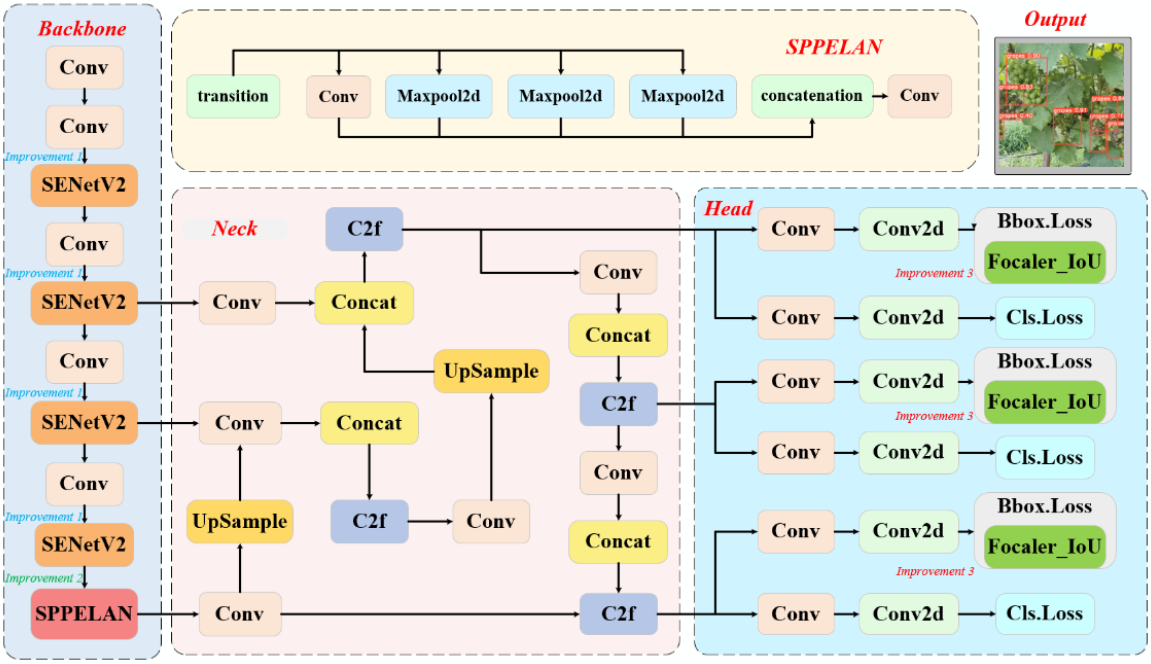
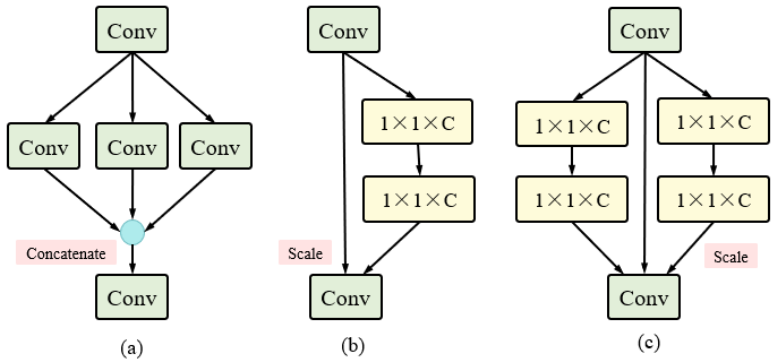


Fig. 2 - YOLOGpnet model structure

SENetV2 module

The multi-layer structure of deep neural networks makes it difficult to effectively propagate learned features, which can lead to performance degradation. This issue can be mitigated by enhancing feature propagation through shortcut connections in residual modules. The Squeeze-and-Excitation Network V2 (SENetV2) is an image classification model based on convolutional neural networks (CNNs), which improves recognition accuracy by extracting spatial features and optimizing channel representations. Figure 3 compares three network modules: ResNeXt merges features through a multi-branch CNN structure; SENet applies global average pooling, fully connected layers, and Sigmoid activation after standard convolution to obtain channel weights and scale the features; SENetV2 combines the characteristics of both, employing a multi-branch fully connected layer to squeeze and excite the features before scaling them.



(a) ResNeXt; (b)SENet; (c)SENetV2

Note: Concatenate refers to the merging operation, Scale refers to the scaling operation, and 1×1×C denotes a fully connected layer with a size of 1×1 and C channels.

Fig. 3 - Comparison of neural network modules

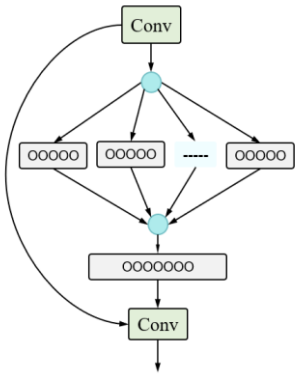


Fig. 4 - Compression and excitation module structure diagram

The design of SENetV2 enhances feature representation granularity and the ability to integrate global information through a multi-branch structure. The proposed SaE module (as shown in Figure 4) dynamically adjusts channel weights through the squeeze-and-excitation process, either enhancing or suppressing specific channel features. The output of the squeeze operation is passed through a multi-branch fully connected layer for excitation and then restored to its original shape. By incorporating multi-branch dense layer design, SENetV2 significantly improves prediction accuracy while maintaining nearly the same number of parameters. It enhances the network's ability to capture both intra-channel and inter-channel patterns, effectively accounting for dependencies between channels.

The training process is illustrated in Fig.5. Extensive experiments demonstrate that SENetV2 surpasses existing architectures in terms of accuracy.

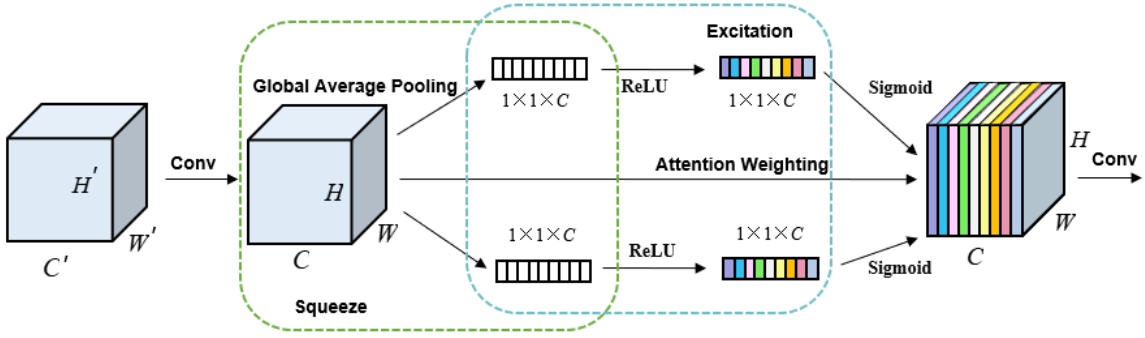
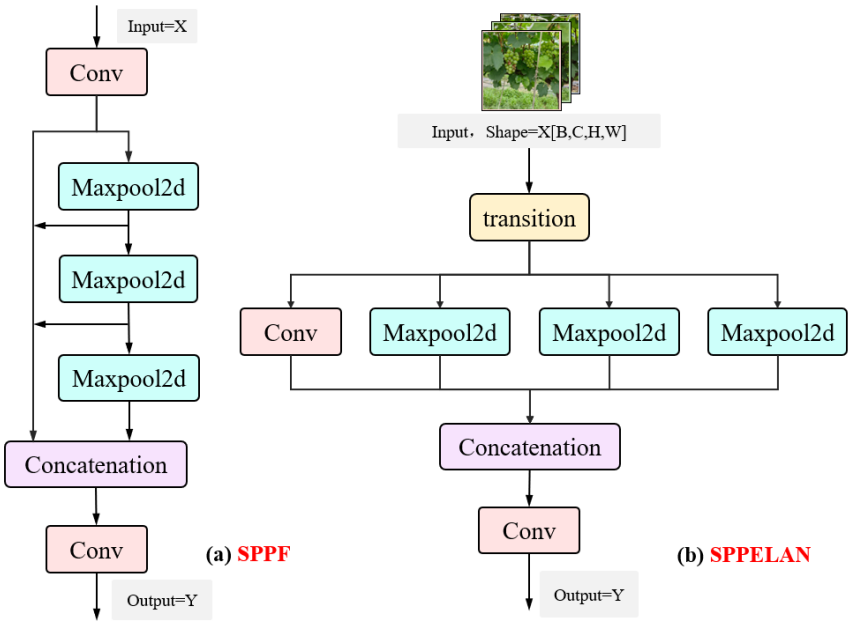


Fig. 5 - Compression and excitation module

SPPELAN module

The Spatial Pyramid Pooling with Enhanced Local Attention Network (SPPELAN) module is designed to improve model performance. It consists of two key components: Spatial Pyramid Pooling (SPP) and the Enhanced Local Attention Network (ELAN). Figure 6 compares the processing flows of SPPF and SPPELAN. SPP generates fixed-size features through multi-scale pooling, enhancing the network's ability to capture multi-scale information and improving its adaptability to input images of varying sizes and object detection performance. ELAN, utilizing a local attention mechanism, dynamically adjusts feature weights, allowing the network to focus more on critical regions. This makes it particularly effective in complex backgrounds and for detecting small objects, which is especially useful in fruit recognition tasks within orchard environments. The feature maps processed by SPP and ELAN are then fused, further enhancing the model's representational capacity.



Maxpool2d refers to the application of 2D max pooling on an input signal composed of multiple input planes.

Fig. 6 - (a)SPPF Module (b)SPPELAN Module

Focaler-IoU loss function

The loss function is a critical tool for measuring the difference between model predictions and actual results. In the regression branch of YOLOv8's head network, Distribution Focal Loss (DFLoss) and Complete Intersection over Union Loss (CIOULoss) are combined. YOLOGPnet replaces CIOULoss with Focaler-IoU. The CIOU loss function (Formula 1) primarily considers the overlap degree and size differences between bounding boxes, making it susceptible to scale variations and reducing detection accuracy. IoU represents the intersection over union between the predicted and ground truth boxes, α is a weighting coefficient, v reflects the aspect ratio difference between the predicted and ground truth boxes, $\rho^2(b^{gt}, b)$ denotes the Euclidean

distance between their center points, and c is the diagonal length of the smallest enclosing box of the two. To improve the accuracy of grape picking point recognition, optimizing the choice of loss functions is crucial.

To improve the accuracy of grape picking point identification, optimizing the choice of the loss function is especially important.

$$CIoU\_Loss = 1 - IoU + \alpha v + \frac{\rho^2(b^{gt},b)}{c^2} \quad (1)$$

The Focaler-IoU loss function combines Focal Loss and IoU Loss to address issues of class imbalance and bounding box regression accuracy in object detection. Focal Loss, as defined in Equation (2), balances the weights of positive and negative samples through the parameter α and adjusts the importance of hard and easy samples using the parameterγ; $\hat{p}$ represents the predicted probability of the model. IoU evaluates the overlap between the predicted and ground truth bounding boxes, with Equation (3) representing the ratio of the intersection area to the union area of the predicted and true boxes. IoU Loss, as defined in Equation (4), optimizes the position of the predicted box by maximizing its IoU with the ground truth box, thereby improving regression accuracy, where B_p is the predicted bounding box and B_t is the ground truth bounding box.

Focaler-IoU focuses on different regression samples and reconstructs the IoU loss using linear mapping, emphasizing the impact of hard and easy samples in bounding box regression. Equation (5) defines the mechanism by which the loss is adjusted according to the IoU value: when IoU is below the lower threshold d, the loss is set to 0; when IoU exceeds the upper threshold u, the loss is set to 1; and when IoU falls between d and u, the loss increases linearly with the IoU value. This design focuses on samples with moderate overlap, enhancing the model's feature extraction capability.

$$Focal\ Loss=-\alpha(1-\hat{p})^{\gamma} log\ (\hat{p}) \quad (2)$$

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} = \frac{|B_p \cap B_t|}{|B_p \cup B_t|} \quad (3)$$

$$IoU\ Loss=1-IoU(B_p, B_t) \quad (4)$$

$$IoU^{focaler\_Loss} = \begin{cases} 0, & IoU < d \\ \frac{IoU-d}{u-d}, & d \ll IoU \ll u \\ 1, & IoU > u \end{cases} \quad (5)$$

RESULTS

EXPERIMENTAL DESIGN AND ANSLYSIS

Evaluation indicators

To validate the detection capability of the YOLOGPnet algorithm, the specific calculation formulas for accuracy and mean Average Precision (mAP) used in this study are provided in Equations (6) and (7) as evaluation metrics for detection performance, TP refers to the number of samples correctly predicted as positive by the model, while FP represents the number of samples incorrectly predicted as positive. N denotes the total number of target categories to be detected or classified by the model, and AP_i is the average precision of the i category. The model's performance is assessed using mAP@0.5%, parameter count, and GFLOPs. Additionally, for evaluating the real-time performance of the grape orchard detection and harvesting task, the model's inference speed is measured using the frames per second (FPS) for single-image inference.

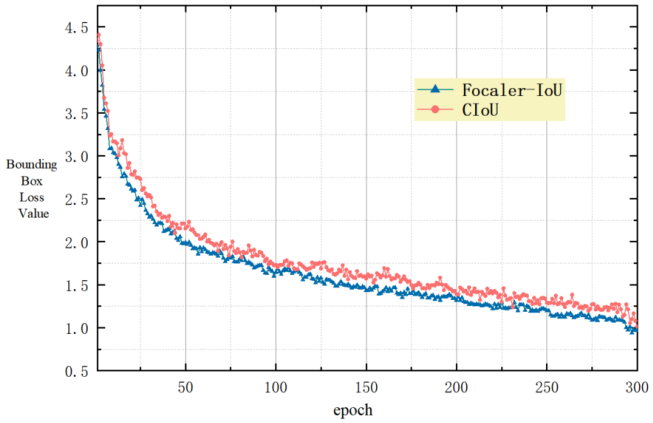
$$P = \frac{TP}{TP + FP} \quad (6)$$

$$mAP = \frac{1}{N}\sum_{i=1}^{n} AP_i \quad (7)$$

Performance Validation of the Loss Function

The optimization of the loss function allows the predicted results to better approximate the true values. Figure 7 compares the CIoU loss function used in the original YOLOv8n model with the Focaler-IoU loss function introduced in this study. Their convergence speed and loss values are similar; however, when using the CIoU loss function for the bounding box, the regression is slower, and the final loss value after convergence is higher. The Focaler-IoU loss function introduced in this study demonstrates a faster and more stable convergence, gradually converging after 280 epochs with the lowest final loss value. This not only accelerates the convergence speed of the model but also improves its accuracy.

As shown in Figure 8, the blue bounding boxes represent the detection results using Focaler-IoU Loss, while the green bounding boxes correspond to the results using CIoU Loss. Focaler-IoU Loss enables the bounding boxes to more closely approximate the minimum enclosing rectangles of the target grapes, resulting in superior detection performance.



CIoU indicates that the model uses CIoU Loss as the bounding box loss function;Focaler-IoU indicates that the model uses Focaler-IoU Loss as the bounding box loss function.

Fig. 7 - Comparison of Bounding Box Loss Convergence with Improved Methods

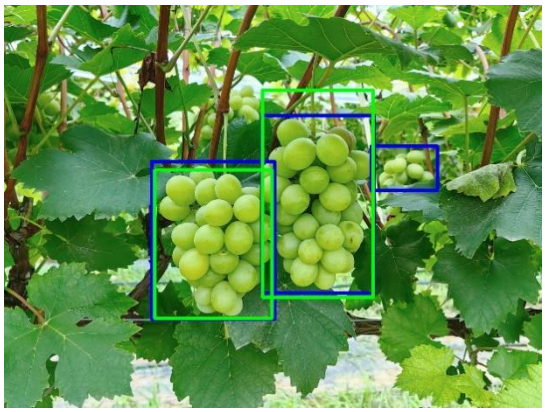


Fig. 8 - Comparison of Detection Performance between CIoU Loss and Focaler-IoU Loss

Comparison of Detection Results from Different Models

To evaluate the detection performance of different models on the grape cluster dataset, this study selected YOLOv5s, YOLOv7-tiny, YOLOv8n, and the improved model YOLOGPnet for experiments on the same test set. The detection results of each model were compared and analyzed using metrics such as accuracy. Two-stage object detection algorithms were excluded from the comparison due to their large computational load and weight file size, which do not meet the requirements for lightweight real-time detection and are unsuitable for use in orchard environments. The YOLO series is more appropriate for the lightweight real-time detection demands of this dataset. The experimental results of different models are presented in Table 2.

Table 2

Comparative Experimental Results of Different Models on the Test Set

Model	Precision (%)	mAP @0.5 (%)	Gflops	Parameter（M）	Weight File（MB）	Inference Time per Image（ms）
YOLOv5s	92.9	95.7	10.3	3.8	7.4	15.9
YOLOV7-tiny	91.7	94.6	7.8	2.8	4.5	12.1
YOLOV8n	90.1	95.2	8.7	3.2	6.3	16.8
Improved Model YOLOGpnet	93.6	96.8	6.8	2.14	4.36	12.5

The comparative results show that the proposed YOLOGPnet model outperforms YOLOv5n, and YOLOv7-tiny across all evaluation metrics. Compared to YOLOv5s, the proposed model improves accuracy and mAP@0.5 by 0.7 and 1.1 percentage points, respectively, while reducing computational load, parameter count, and weight file size by 33.98%, 43.68%, and 41.08%, respectively, with an increase in detection speed of 3.4 ms. The model significantly reduces computational resource consumption while maintaining high accuracy, making it suitable for deployment on mobile devices due to its smaller memory footprint and weight file size.

As shown in Table 3, the improved model achieves an accuracy of 93.6% and an mAP@0.5 of 96.8%, representing increases of 3.5 and 1.6 percentage points, respectively, compared to the original YOLOv8n model. Additionally, all other metrics show improvements across different models. The model's weight file size is reduced by 30.79%, while parameter count and detection time are reduced by 33.13% and 4.3 ms, respectively. Overall, the model demonstrates outstanding performance in grape cluster detection, with significantly enhanced recognition accuracy.

Ablation Study

To verify the effectiveness of the proposed improvements, ablation experiments were conducted under the same experimental conditions using the original YOLOv8n as the baseline. By testing different combinations of the SENetV2, SPPELAN module, and Focaler-IoU loss function, the accuracy, mAP@0.5, computational load, weight file size, and single-image detection time were evaluated on the same grape dataset. The results are shown in Table 3. In these experiments, A represents replacing the C2f module in YOLOv8n's backbone network with SENetV2, B represents replacing the SPPF module with SPPELAN, and C represents replacing the CIoU loss function with Focaler-IoU. The symbol "×" indicates that the improvement strategy was not applied, while "√" indicates that the improvement strategy was applied.

Table 3

Ablation Study of Different Improvement Methods

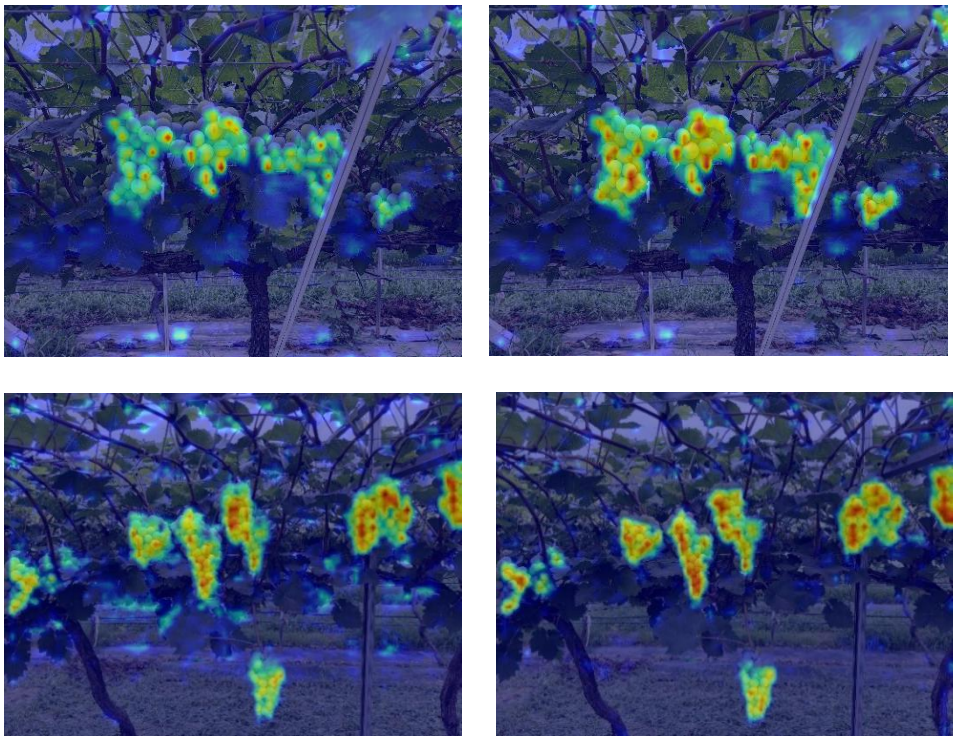
A	B	C	Precision (%)	mAP @0.5 (%)	Gflops	Parameter（M）	Weight File（MB）	Inference Time per Image（ms）
—	—	—	90.1	95.2	8.7	3.2	6.3	16.8
√	—	—	93.2	96.4	7.3	2.7	4.3	12.0
—	√	—	91.9	95.3	7.6	2.6	4.7	14.1
—	—	√	90.9	95.1	8.2	3.1	6.1	12.6
√	√	—	93.4	96.6	6.7	2.1	4.2	11.8
√	—	√	92.7	96.1	7.2	2.3	4.9	12.9
√	√	√	93.6	96.8	6.8	2.14	4.36	12.5

Compared to the baseline YOLOv8n model, the YOLOv8n+A model improved detection accuracy and mAP@0.5 by 3.1 and 1.2 percentage points, respectively, while reducing the computational load, parameter count, weight file size, and single-image detection time by 16.09%, 37.5%, 31.75%, and 28.57%, respectively. This improvement is attributed to the introduction of SENetV2, which enhances feature extraction capabilities through the squeeze-and-excitation operations while reducing model parameters, thus validating its effectiveness in both lightweight design and performance enhancement. The YOLOv8n+B model improved detection accuracy and mAP@0.5 by 1.8 and 0.1 percentage points, respectively, while reducing the computational load, parameter count, weight file size, and detection time by 12.64%, 18.75%, 25.4%, and 16.07%, respectively. The SPPELAN module reduced feature redundancy and enhanced the extraction of detailed features, further improving model performance. The YOLOv8n+C model increased detection accuracy by 0.8 percentage points, although mAP@0.5 decreased slightly by 0.1 percentage points. However, single-image detection time decreased by 4.2 ms, with parameter count and weight file size remaining almost unchanged.

Compared to YOLOv8n, the YOLOv8n+A+B+C model improved accuracy and mAP@0.5 by 3.5 and 1.6 percentage points, respectively, while reducing computational load, parameter count, weight file size, and detection time by 21.84%, 34.38%, 30.79%, and 25.60%, respectively. When compared to the YOLOv8n+A+B model, accuracy and mAP@0.5 improved by 0.2 percentage points, with computational load remaining nearly the same. Compared to the YOLOv8n+A+C model, the YOLOv8n+A+B+C model reduced detection time by 0.4 ms, while increasing accuracy and mAP@0.5 by 0.9 and 0.7 percentage points, respectively. Additionally, the computational load, parameter count, and model size decreased by 5.56%, 8.70%, and 11.02%, respectively, demonstrating significantly superior overall performance compared to the YOLOv8n+A+C model.

Heatmap Visualization

To intuitively evaluate the detection performance of the YOLOGPnet model, this study utilizes Grad-CAM to generate heatmaps for visualizing the target detection process. In the heatmaps, red and yellow regions represent areas that have a greater influence on the model's decision-making. Figure 9 displays the detection results for several grape images, indicating a high level of consistency between the YOLOGPnet model and the original images. Compared to the original model, YOLOGPnet more accurately identifies overlapping fruits and shows improved feature capture along the edges of the fruits. This demonstrates its enhanced ability to extract and focus on features in complex backgrounds and for weak semantic targets.



（a）YOLOv8n （b）YOLOGPnet

Fig. 9 - Heatmap of Grape Image Detection

Comparison of Detection Performance Across Models Under Different Conditions

To validate the performance and generalization ability of the improved model in real grape orchard environments, this study randomly selected 30 images under different shooting conditions, including close-range, long-range, cloudy, and sunny scenes. The images contain complex situations such as fruit overlap and leaf occlusion, without applying image augmentation, to simulate a realistic orchard environment. Comparative experiments were conducted using the YOLOv5s, YOLOv7-tiny, YOLOv8n, and YOLOGPnet models. Figures 10 and 11 display the prediction results of each model, with purple, green, red, and blue bounding boxes representing the predictions of YOLOv5s, YOLOv7-tiny, YOLOv8n, and YOLOGPnet, respectively. Missed and misdetections are marked with light blue and yellow circles.

Close-range grape images help analyze model detection performance in complex situations such as overlap and occlusion. Figure 10 shows that detecting green grapes is challenging due to their similar color to the orchard background. Under sufficient lighting, the distinction between grapes and the background is clearer, and all four models perform relatively well. However, YOLOv8n performs poorly in cases of severe occlusion, while YOLOv7-tiny tends to identify overlapping grapes as a single target. Additionally, both models mistakenly detect branches as grapes. In contrast, YOLOGPnet demonstrates excellent detection accuracy in complex scenes, with its predicted bounding boxes more closely approximating the minimum enclosing rectangles of the targets, reducing gaps and over-wrapping, and improving the accuracy of picking point prediction. YOLOGPnet only exhibited one instance of an inaccurate bounding box, whereas the other models showed more misdetections, and YOLOv8n even missed the target entirely.

Long-range grape images better reflect the performance of each model in multi-object detection, particularly under poor lighting conditions, which further test the models' generalization capabilities. On cloudy days, the brightness of the grapes and leaves decreases, making the boundaries less distinct and the detection of overlapping and occluded grapes more difficult. Figure 11 shows that YOLOv5s and YOLOv8n performed poorly in long-range detection under all weather conditions, with many missed and misdetections. YOLOv7-tiny performed better on sunny days compared to cloudy conditions. Meanwhile, YOLOGPnet had only two misdetections under all conditions, and the bounding box convergence remained excellent in long-range images, indicating good adaptability to lighting variations.

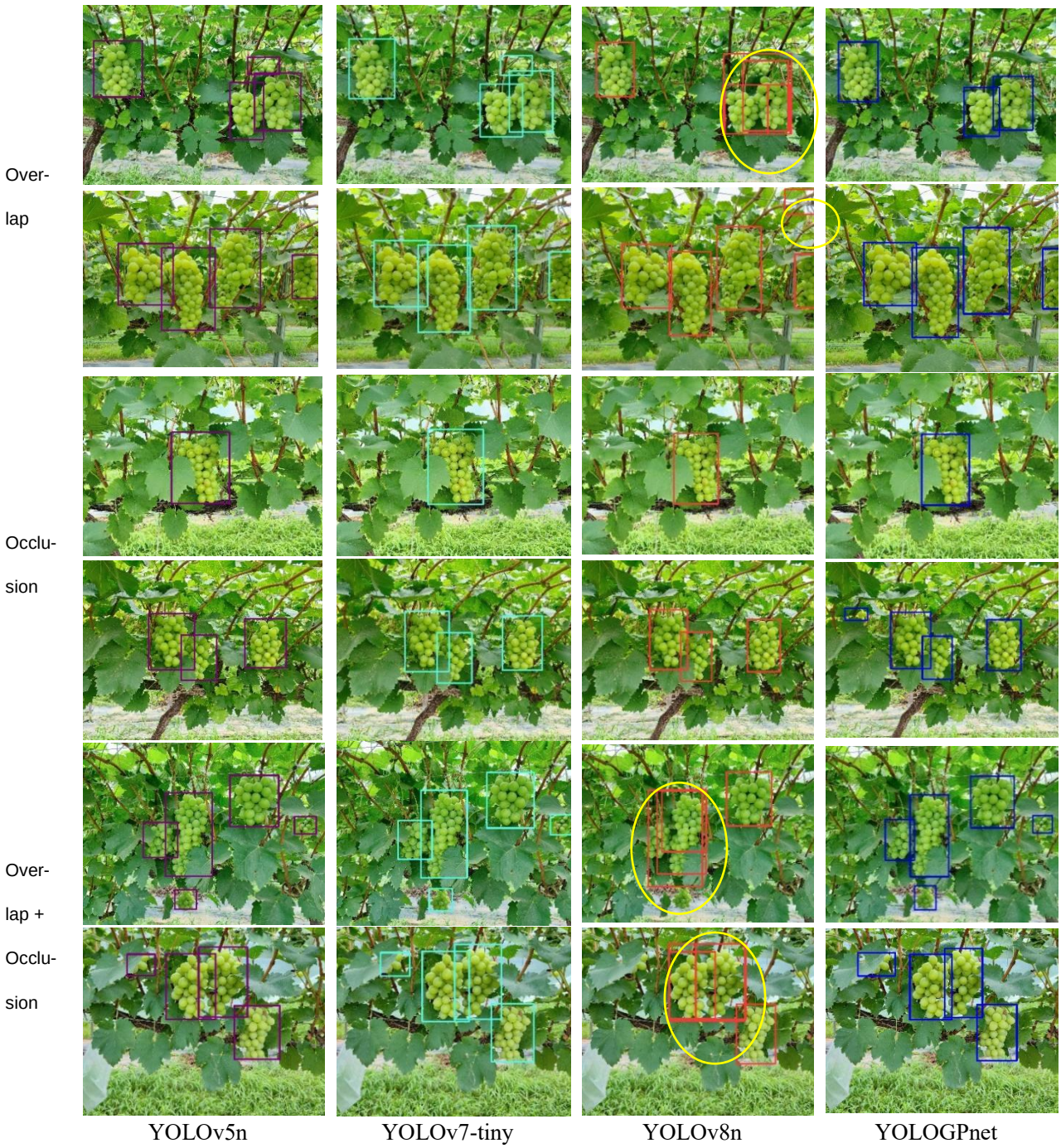
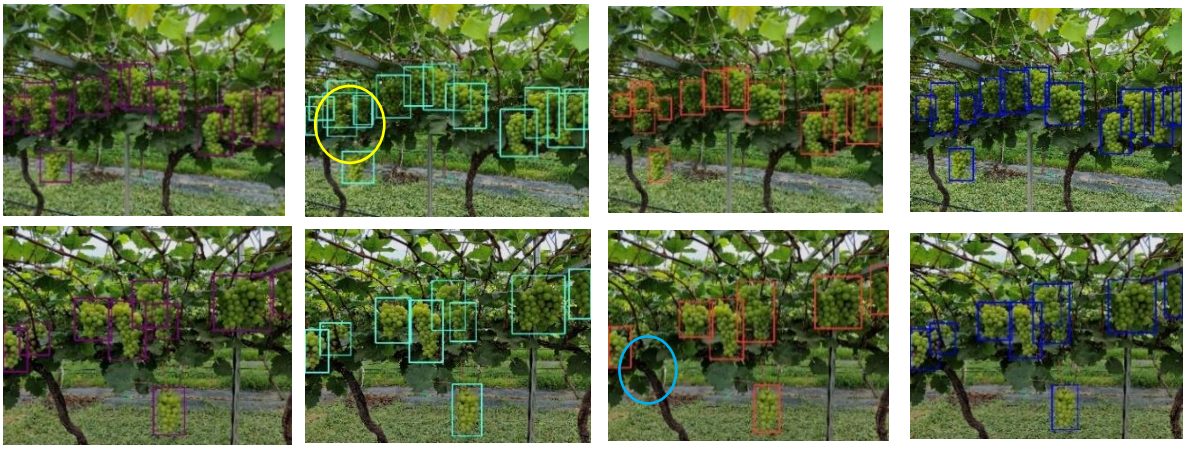


Fig.10 - Comparative Detection Results of Multiple Models under Different Conditions (Close-range)



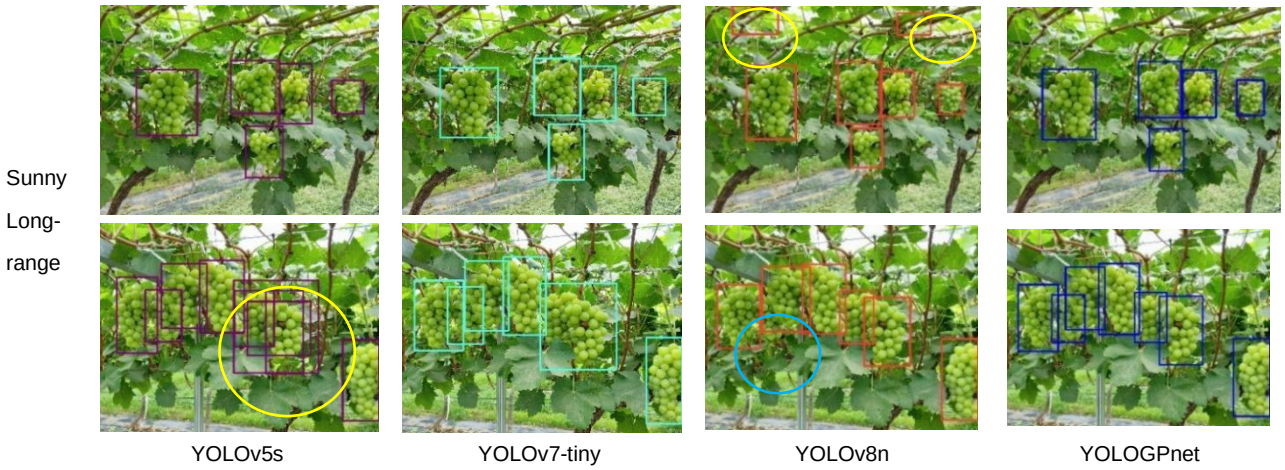


Fig.11 - Comparative Detection Results of Multiple Models under Different Weather Conditions (Long-range)

CONCLUSIONS

This study proposes the YOLOGPnet model, which demonstrates significant performance improvements. By incorporating SENetV2 and SPPELAN modules, the model effectively addresses the limitations in multi-scale feature extraction, while the Focaler-IoU loss function further enhances regression accuracy and predictive performance. Comparative experiments under varying lighting conditions validate the practicality and robustness of YOLOGPnet, showing fewer false positives and missed detections compared to other models, with predicted bounding boxes more closely aligning with target fruits. The improvements in evaluation metrics and detection performance indicate that YOLOGPnet maintains high accuracy and stability in complex environments, making it particularly suitable for resource-constrained applications.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China （No.32272001） and Qinchuangyuan “Scientist + Engineer” Team Development Program of the Shaanxi Provincial Department of Science and Technology （No.2023KXJ-016）.

REFERENCES

[1] Cha, Z., Zhou, W., Wu, J., (2021). Field-based recognition of Red Globe grape bunches using transfer learning and the Faster R-CNN model(基于迁移学习 Faster R-CNN 模型田间红提葡萄果穗的识别). *Journal of Shihezi University (Natural Science Edition)*, Vol. 1, pp.26-31, Xinjiang/China.

[2] Chen, Y., Niu, Y., Cheng, W., Zheng, L., Sun, D., (2024). Apple detection method in the natural environment based on improved YOLOv5. *Inmateh- Agricultural Engineering*, Vol. 72, pp.183-192, Romania.

[3] Guo, C., Zheng, S., Cheng, G., Zhang, Y., Ding, J., (2023). An improved YOLOv4 used for grape detection in unstructured environments. *Frontiers in Plant Science*, Vol.14, 1209910, Switzerland.

[4] Jiang, T., Li, Y., Feng, H., Wu, J., Sun, W., Ruan, Y., (2024). Research on a trellis grape stem recognition method based on YOLOv8n-GP. *Agriculture*, Vol.9, 1449, Switzerland.

[5] Khan, N., Fahad, S., Naushad, M., (2020). Grape production critical review in the world . https://ssrn.com/abstract=3595842.

[6] Li, H., Li, C., Li, G., Chen, L., (2021). A real-time table grape detection method based on an improved YOLOv4-tiny network in complex backgrounds. *Biosystems Engineering*, Vol.212, pp.347-359, , England.

[7] Liu, B., Zhang, Y., Wang, J., Luo, L., Lu, Q., Wei, H., Zhu W., (2023). An improved lightweight network based on deep learning for grape recognition in unstructured environments. *Information Processing in Agriculture*, Vol.2, pp.202-216, Beijing/China

[8] Liu, P., Zhu, Y., Zhang, T., Hou, J., (2020). Recognition and image segmentation algorithm for closely packed grape bunches under natural conditions (自然环境下贴叠葡萄串的识别与图像分割算法). *Transactions of the Chinese Society of Agricultural Engineering*, 2020, Vol. 6, pp.161-169, Beijing/China.

[9] Lu, J., Lei, W., (2021). Over lapping grape segmentation algorithm based on full convolutional network and concave point search(基于全卷积网络与凹点搜索的重叠葡萄分割算法). *Journal of Optoelectronics-Laser*, Vol. 03, pp. 231-240, Tianjin/China.

[10] Lu, S., Liu, X., He, Z., Zhang, X., Liu, W., Karkee, M., (2022). Swin-Transformer-YOLOv5 for real-time wine grape bunch detection. *Remote Sensing*, Vol.14, pp.53-58, Switzerland.

[11] Ning, Z., Luo, L., Liao, J., Wen, H., Wei, H., Lu, Q., (2021). Grape stem recognition and optimal picking point positioning based on deep learning (基于深度学习的葡萄果梗识别与最优采摘定位). *Transactions of the Chinese Society of Agricultural Engineering*, Vol. 9, pp. 222-229, Beijing/China.

[12] 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*, Vol.14, 2677, Switzerland.

[13] Ren, J., Zhang, H., Wang, G., Dai, C., Teng, F., Li, M., (2024). Real-time grape disease detection model based on improved YOLOv8s. *INMATEH-Agricultural Engineering*, *72*(1). *INMATEH - Agricultural Engineering*, Vol. 72, pp.96-105, Romania.

[14] Song, Y., Yang, S., Zheng, Z., Ning, J., (2023). Tea-picking point semantic segmentation algorithm based on multi-head self-attention mechanism (基于多头自注意力机制的茶叶采摘点语义分割算法). *Transactions of the Chinese Society for Agricultural Machinery*, Vol. 09, pp. 297-305, Beijing/China.

[15] Su, S., Chen, R., Fang, X., Zhu, Y., Zhang, T., Xu, Z., (2022). A novel lightweight grape detection method. *Agriculture*, Vol. 9, 1364, Switzerland.

[16] Sun, J., Wu, Z., Jia, Y., Gong, D., Wu, X., Shen, J., (2023). Grape detection in orchard environments based on improved YOLOv5s (基于改进 YOLOv5s 的果园环境葡萄检测). *Transactions of the Chinese Society of Agricultural Engineering*, 2023, Vol.18, pp.192-200, Beijing/China.

[17] Wang, J., Zhang, Z., Luo, L., Zhu, W., Chen, J., Wang, W., (2021). SwinGD: A robust grape bunch detection model based on Swin transformer in complex vineyard environments. *Horticulturae*, Vol. 7, 492, Switzerland.

[18] Wen, S., Zhou, J., Hu, G., Zhang, H., Tao, S., Wang, Z., & Chen, J., (2024). PcMNet: an efficient lightweight apple detection algorithm in natural orchards. *Smart Agricultural Technology*, Vol.9, 100623, Netherlands.

[19] Wu, Z., Xia, F., Zhou, S., Xu, D., (2023). A method for identifying grape stems using keypoints. *Computers and Electronics in Agriculture*, Vol. 209, 107825, England.

[20] Yang, W., Qiu, X., (2024). A lightweight and efficient model for grape bunch detection and biophysical anomaly assessment in complex environments based on YOLOv8s. *Frontiers in Plant Science*, 2024, Vol.15, 1395796, Switzerland.

[21] Yang, Y., Han, Y., Li, S., Yang, Y., Zhang, M., Li, H., (2023). Vision-based fruit recognition and positioning technology for harvesting robots. *Computers and Electronics in Agriculture*, Vol. 213, England.

[22] Zhang, C., Ding, H., Shi, Q., Wang, Y., (2022). Grape cluster real-Time detection in complex natural scenesbased on YOLOv5s deep learning network. *Agriculture*, Vol.14,1242, Switzerland.

[23] Zhang, T., Wu, F., Wang, M., Chen, Z., Li, L., Zou, X., (2023). Grape-bunch identification and picking point location on occluded fruit axes based on YOLOv5-GAP. *Horticulturae*, Vol. 4, 498, Switzerland

[24] Zhao, C., Fan, B., Li, J., Feng, Q., (2023). Advances, challenges, and trends in agricultural robotics technology (农业机器人技术进展、挑战与趋势). *Smart Agriculture (Bilingual)*, Vol. 04, pp. 1-15, Beijing/China.

[25] Zhao, R., Zhu, Y., Li, Y., (2022). An end-to-end lightweight model for grape and picking point simultaneous detection. *Biosystems Engineering*, Vol. 223, pp. 174-188 , England.

[26] Zhao, F., Zhang, J. W., Zhang, N., Zhang, N., Tan, Z., Xie, Y., Zhang, S., Han, Z., Li, M., (2022). Detection of cucurbits' fruits based on deep learning. *INMATEH- Agricultural Engineering*, Vol. 72, pp.321-330, Romania.