agriculture

MDPI

*Article*

# Grape Cluster Real-Time Detection in Complex Natural Scenes Based on YOLOv5s Deep Learning Network

**Chuandong Zhang [1,2,*], Huali Ding [1], Qinfeng Shi [1] and Yunfei Wang [2]**

[1] School of Mathematics and Computer Application Technology, Jining University, Qufu 273100, China
[2] College of Mechanical and Electronic Engineering, Northwest A&F University, Xianyang 712100, China
* Correspondence: jnzcd@jnxy.edu.cn

**Abstract:** Due to differences in planting environment, color, shape, size, and compactness, accurate detection of grape clusters is very difficult. Herein, a real-time detection method for grape clusters based on the YOLOv5s deep learning algorithm was proposed. More specifically, a novel dataset called Grape-internet was constructed, which consisted of 8657 grape images and corresponding annotation files in complex scenes. By training and adjusting the parameters of the YOLOv5s model on the data set, and by reducing the depth and width of the network, the lightweight processing of the network was completed, losing only a small amount of accuracy. As a result, the fast and accurate detection of grape clusters was finally realized. The test results showed that the precision, recall, mAP and F1 of the grape cluster detection network were 99.40%, 99.40%, 99.40% and 99.40%, respectively, and the average detection speed per image was 344.83 fps, with a model size of 13.67 MB. Compared with the YOLOv5x, ScaledYOLOv4-CSP and YOLOv3 models, the precision of YOLOv5s was 1.84% higher than that of ScaledYOLOv4-CSP, and the recall rate and mAP were slightly lower than three networks by 0.1–0.3%. The speed was the fastest (4.6 times, 2.83 times and 6.7 times of YOLOv3, ScaledYOLOv4-CSP and YOLOv5x network, respectively) and the network scale was the smallest (1.61%, 6.81% and 8.28% of YOLOv3, ScaledYOLOv4-CSP YOLOv5x, respectively) for YOLOv5s. Moreover, the detection precision and recall rate of YOLOv5s was 26.14% and 30.96% higher, respectively, than those of Mask R-CNN. Further, it exhibited more lightweight and better real-time performance. In short, the detection network can not only meet the requirements of being a high precision, high speed and lightweight solution for grape cluster detection, but also it can adapt to differences between products and complex environmental interference, possessing strong robustness, generalization, and real-time adaptability.

**Keywords:** grape cluster detection; YOLOv5s; object detection; real-time detection; lightweight



**Citation:** Zhang, C.; Ding, H.; Shi, Q.; Wang, Y. Grape Cluster Real-Time Detection in Complex Natural Scenes Based on YOLOv5s Deep Learning Network. *Agriculture* **2022**, *12*, 1242. https://doi.org/10.3390/agriculture12081242

Academic Editors: Nen Fu Huang and Ho-Hsien Chen

Received: 3 July 2022
Accepted: 13 August 2022
Published: 17 August 2022



## 1. Introduction

Accurate prediction of grape yield is key to the scientific management and production scheduling of large-scale grape planting enterprises, and is also an important basis for adjusting their marketing strategies [1–3]. Traditional yield estimation relies on manual experience, and its efficiency and accuracy are low, meaning it cannot meet the rapid and accurate prediction requirements of large-scale planting enterprises. The intelligent perception and acquisition of grape cluster information is one of the key techniques in grape yield estimation [4–7]; namely, accurate yield prediction can be achieved by using the results of identifying and counting grape clusters in complex natural scenes as the input for the application of yield estimation. However, due to the differences in color, shape, size and compactness of grape clusters, as well as the influence of light and shading, accurate grape cluster detection is challenging. Up to now, the existing grape cluster target detection algorithms have been mainly used in research on precision, but there is much less real-time detection research. It is of great significance to detect grape clusters in real-time with high

precision in order to improve crop yield estimation efficiency, and to economically benefit planting enterprises.

To date, lots of research on grape target detection and segmentation has been carried out by traditional machine learning- and deep learning-based methods. For traditional machine learning-based methods, grape detection is mostly based on feature vectors and related machine learning models, which can achieve high-precision in detecting targets. However, the robustness and adaptability of the detection algorithms need to be improved, as it is difficult to extract effective features for the recognition of shading, light and targets of different varieties in complex planting environments. For example, Liu et al. [4] employed a Support Vector Machine (SVM) that combined color and texture information to detect grape clusters in images, with an accuracy rate of 88.00%. Aquino [8] et al. used a three-layer neural network to segment each cluster of grape berries in the CIELAB color space with an average precision of 87.05%. Nuske et al. [9] detected and estimated the yield of grape berries under various conditions by analyzing texture, color and shape, where the average error was in the range of 3.00–11.00%. Badeka et al. [10] detected red and white grapes based on the K-nearest neighbor (KNN) model, combining local binary patterns (LBPs) to extract texture and color features; the detection accuracy of red grapes was 94.00%, and that of white grapes was 83.00%, indicating that the model had insufficient generalization ability to different varieties of grapes. Luo et al. [11] proposed an overlapping grape cluster segmentation algorithm based on K-means clustering and color features; the recognition accuracy of double overlapping grape clusters was 88.33%. Rodrigo et al. [12] detected grape clusters with different degrees of overlapping and different colors under natural lighting conditions by using a berry identification and grape cluster detection strategy based on the histogram of oriented gradient (HOG) and LBP information; the average precision was 88.61%. By comparison, deep learning-based methods possess powerful feature extraction abilities, and can automatically learn information about various features from low-level to high-level images. They can then cope with factor changes including pose, color and illumination [13]. It is widely used in the detection and segmentation of grapes in natural scenes [14], and has become a promising direction for grape cluster detection research. Grimm et al. [15] used a fully convolutional neural network (FCN) with VGG-Net 16 as the backbone to segment and detect many organs, including grape berries, and the precision of berry detection was 86.60% with an F1 value of 87.60%. Cecotti et al. [6] compared eleven existing pre-trained (VGG16, VGG19, Resnet50, GoogLeNet, etc.) networks to semantically segment red and white grape clusters with a transfer learning-based method, and results showed that the ResNet structure had the best performance, with an accuracy of 99.00%. Marani et al. [3] adopted transfer learning, and used four pre-trained networks (AlexNet, GoogLeNet, VGG16, VGG19) to automatically segment grape cluster images. Results showed that VGG19 has the best performance, where the average accuracy rate of segmentation on clusters was 80.58%. Santos et al. [13] used a Mask R-CNN-based network for instance segmentation and object detection for grape clusters, with an F1 value of 91.00%. Zabawa [16] introduced the "edge" class in pixel segmentation to separate adjacent berries, and reconstructed instance segmentation as semantic segmentation. They proposed a fully convolutional semantic segmentation with the lightweight backbone MobileNetV2 as the encoder, and DeepLabV3+ as the decoder. The network achieved accurate and fast counting of vineyard berries, where the recognition accuracy of berries was in the range of 85.00%–94.00%. Aguiar [17] detected objects of grape clusters at different growth stages in images by fine-tuning the pre-trained SSD MobileNet-V1 model, and the mAP value was 66.96%. Ghiani [7] proposed a grape cluster detection method based on Mask R-CNN frameworks with a mAP of 91.00%. Yin et al. [18] used the Mask R-CNN network to detect grape images, where the average precision of grape cluster detection was 89.53%.

Grape detection methods based on deep convolutional neural networks are mostly used for field grape yield prediction. Considering the huge differences in shape, size and compactness among grape varieties [13], the phenotypic characteristics of the shape and size of berries or grape clusters, and the desire to achieve higher detection results, semantic

segmentation and instance segmentation of berries [15,16] or grape clusters [3,6,13] have become the preferred choices for most researchers. Compared with object detection, semantic segmentation and instance segmentation require a large amount of high-quality annotation data, and the cost of obtaining accurate pixel-level annotation information is huge [6,7,15], without an obvious improvement in precision [13]. Although the above deep learning-based methods exhibit high precision, some shortcomings limit the applications in small mobile terminal devices, such as their high complexity, large number of parameters, large network scale, high computational cost and the insufficient real-time performance. Recently, end-to-end target detection has become the main means of grape detection, so it is necessary to design an end-to-end lightweight grape target detection algorithm that can meet the real-time requirements of the model, while ensuring the precision of grape cluster detection. The YOLOv5s network has high detection precision, fast reasoning speed and strong real-time performance, which has been successfully used in the detection of apples [19], mangos [20] and kiwifruits [21]. In the current work, a lightweight grape cluster real-time detection algorithm, based on YOLOv5s, was proposed by adopting grape clusters as the research object for fast and accurate detection of grape clusters in complex natural scenes.

## 2. Materials and Methods

### *2.1. Materials*

Grape images in the Grape-internet dataset used in this study were all collected from the internet, including Kyoho, Summer Black, Cabernet Sauvignon, Midknight Beauty, Manicure Finger, Fujiminori, Syrah and other red grape varieties, for a total of 787 grape images after data cleaning. They were then randomly cropped to different resolutions (from 514 × 460 pixels to 4160 × 3120 pixels) and manually annotated using LabelImg [22].

To improve the generalization ability of the model, the 787 images were augmented into 8657 images through transformations such as rotation, horizontal mirroring, scaling, translation and changing brightness, and 70.00% (6059 images) of the enhanced dataset was randomly selected as training data to train the YOLOv5s target detection algorithm. For the remaining dataset, 10.00% (866 images) was used as the validation dataset, and the rest (1732 images) was used as the testing dataset. The information of the grape dataset, Grape-internet, is shown in Table 1. The Grape-internet dataset contains seven red varieties of grapes, with great differences in color, compactness, shape, outline, berry particle size and texture differences [13]. In addition, there are differences among the data set images in shooting angle, shooting distance, shooting equipment, imaging size, resolution and shooting light, as well as different degrees of occlusion and overlap. All differences mentioned above reflect the complexity and diversity of the dataset, which greatly increases the difficulty of detection, and brings greater challenges to the detection network, which can then better verify the robustness, generalizability and adaptability of the detection algorithm. Moreover, other differences are listed in Table 1 as follows: the color differences for different growth stages after coloring for the same species (Figure 1), the color, shape, compactness and texture differences for different species after ripening (Figure 2), the color and shape distortion differences due to different shadow and occlusion (Figure 3), and the image differences due to the shooting angle and diversity of light (Figure 4). These characteristics of the dataset further enhance the difficulty of grape cluster detection.


**Figure 1.** Color of Kyoho grapes in different growth periods: (**a**) Two grape clusters of different colors; (**b**) Different ripe grapes of different colors; (**c**) Different color from grapes in (**a**,**b**).

**Table 1.** Grape-internet dataset information.

| | Variety | Number of Original Images | Number of Images after Amplification | Interfering Factors |
|---|---|---|---|---|
| Grape | Kyoho | 128 | 1408 | scene differences, image size, quality differences, shooting techniques, occlusion and overlap, light changes |
| | Summer Black | 119 | 1309 | scene differences, image size, quality differences, shooting techniques, occlusion and overlap, light changes |
| | Cabernet Sauvignon | 108 | 1188 | scene differences, image size, quality differences, shooting techniques, occlusion and overlap, light changes |
| | Midknight Beauty | 100 | 1100 | scene differences, image size, quality differences, shooting techniques, occlusion and overlap, light changes |
| | Manicure Finger | 98 | 1078 | scene differences, image size, quality differences, shooting techniques, occlusion and overlap, light changes |
| | Fujiminori | 120 | 1320 | scene differences, image size, quality differences, shooting techniques, occlusion and overlap, light changes |
| | Syrah | 114 | 1254 | scene differences, image size, quality differences, shooting techniques, occlusion and overlap, light changes |


**Figure 2.** Different species of grapes vary in color, shape, compactness and texture: (**a**) Kyoho ripe grapes; (**b**) Midknight Beauty ripe grapes; (**c**) Syrah ripe grapes.

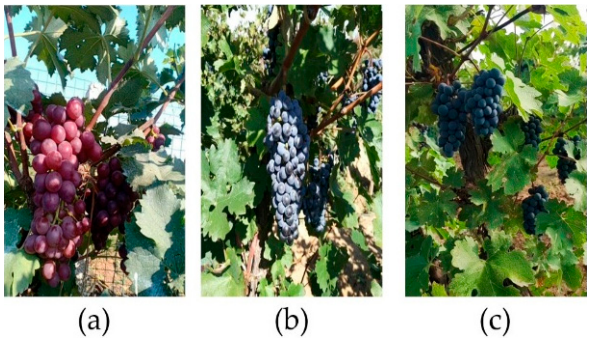


**Figure 3.** Distortion of colors and shapes under shadow or occlusion: (**a**) Grape color distortion in shadow; (**b**) Different color of grapes in the sun and in the shadow; (**c**) Obstructed grape shape distortion.

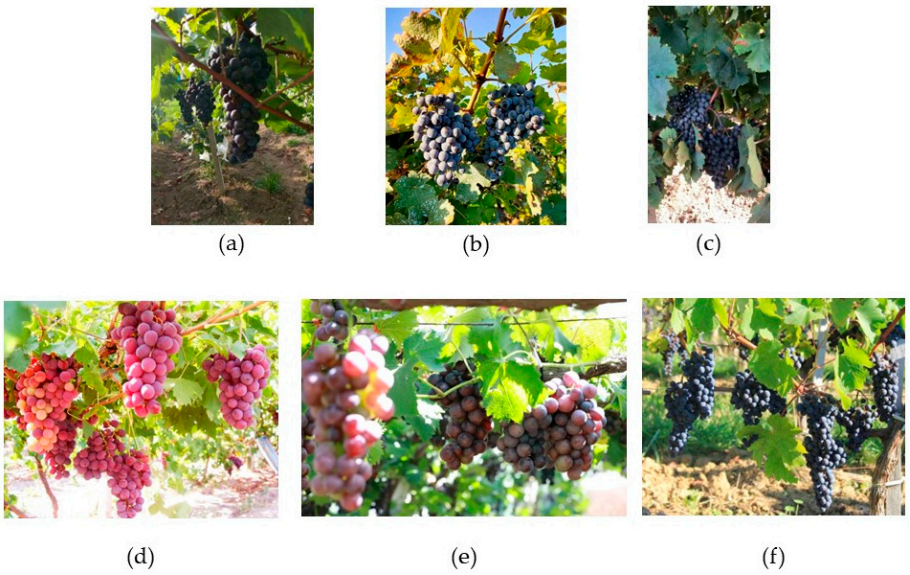


**Figure 4.** Diversity of shooting angles and light: (**a**) Flat angle; (**b**) Elevation angle; (**c**) Depression angle; (**d**) Sidelight; (**e**) Backlight; (**f**) Front light.

### *2.2. Grape Cluster Detection Based on YOLOv5s Deep Learning Algorithm*

#### 2.2.1. YOLOv5s Network Frame

Since the experimental images in this study were all derived from the internet, there were significant differences between the images in variety, shooting background, resolution, illumination and occlusion. To avoid the adverse effects of these differences on the target detection algorithm, and to improve the robustness of the grape detection algorithm, the YOLOv5s deep learning algorithm was used for grape target detection in this work.

The YOLO (You Only Look Once) network [23] is a classic one-stage structure target detection algorithm. The location and classification of detected objects can be predicted in one stage, which can significantly improve detection speed [24]. YOLOv5 is a member of the YOLO family, and the algorithm structure has inherited, optimized and improved the previous modules, such as the Center and Scale Prediction (CSP) structure, the Spatial Pyramid Pooling (SPP) structure and the Feature Pyramid Networks (FPN) + Path Aggregation Network (PAN) structure in YOLOv4. It has also added an adaptive image scaling operation at the input end. Additionally, the Focus structure was added to the baseline network, which greatly improved both the detection speed and precision.

YOLOv5s, one of the four network structures of YOLOv5 [25], possesses a pretrained speed of 156.25 fps and a mAP of 56.0%, which endows it with fast network inference, high detection precision and good real-time performance. Thus, YOLOv5s is suitable for deployment to small mobile devices, and meets the real-time, lightweight and high-precision requirements in this line of work.

The structure of the grape cluster detection model based on YOLOv5s consists of four parts: Input, Backbone, Neck, and Output. These are shown in Figure 5. The components and functions of each module are as follows:

1. Mosaic data augmentation can make the network more robust.
2. Convolution, Batch Normalization and Leaky-ReLU (CBL): A module composed of the activation of functions within the Convolution layer, the Batch Normalization layer and the Leaky-ReLU.
3. Res unit: By drawing on the residual structure in the Resnet network, the network can be built deeper.
4. CSP structure: There are two CSP structures in the network. The CSP1_*X* structure is applied to the Backbone network, which can reduce the amount of computation while

ensuring detection precision. The CSP2_X structure is applied to the Neck, which can strengthen the network feature fusion abilities [26].

5. FPN+ PAN structure [27,28]: The extracted semantic features and localization information are fused to improve the ability of feature extraction.
6. SPP: A spatial pyramid pooling layer, which mainly converts convolutional features of different sizes into pooled features of the same length [29].

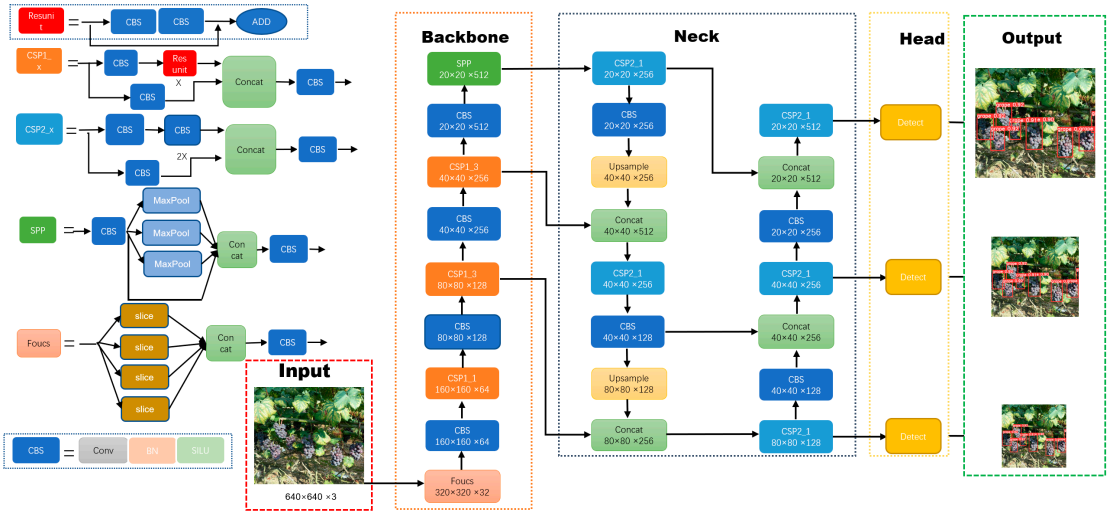


**Figure 5.** Structure of the grape cluster detection model based on YOLOv5s.

### 2.2.2. Fine-Tuning and Training of YOLOv5s Grape Detection Model

The machine configuration used in the experiment is Intel Core E5-1620 processor, 3.50 GHz, 32 GB memory, 500 GB hard disk, and the 11 GB NVIDIA RTX 2080Ti GPU. The PyTorch deep learning framework (PyTorch 1.8, Python 3.8) under the Windows 10 operating system was built to realize grape cluster detection based on YOLOv5s. The main steps are as follows:

1. Organization of data. After the grape image data was downloaded, preprocessing operations such as cleaning, screening and resolution adjustment were performed. The grape clusters in all images were manually labeled, and the data set was divided into a training set (6059 images), a validation set (866), and a testing set 1732) in a ratio of 7:1:2.
2. Fine-tuning of model parameters. To obtain a better grape cluster detection effect based on YOLOv5s, the model parameters were fine-tuned, mainly including network input size, batch size, classes, epoch, learning rate, "conf-thres" and "Iou-thres." Parameters of YOLOv5s in the work are shown in Table 2.
3. Network training and testing. The YOLOv5s grape cluster detection model was trained by using the training set and the validation set. After the training, the weight file of the detection model was obtained, and the performance of the model was evaluated by using the test set. The network produced the location box and probability of the identified grape target.

**Table 2.** Grape cluster detection parameters based on YOLOv5s.

| Parameters | Values |
|---|---|
| Input size | 640 × 640 |
| Batch_size | 8 |
| Classes | 1 |
| Epoch | 300 |

**Table 2.** *Cont.*

| Parameters | Values |
|---|---|
| Learning rate | $1.0 \times 10^{-2}$ |
| Conf-thres | 0.001 |
| Iou-thres | 0.6 |

2.2.3. Training Results of Grape Detection Model Based on YOLOv5s

The loss curves during training and validation, including the localization loss (box_loss), confidence loss (obj_loss), and classification loss (cls_loss), are shown in Figure 5. The positioning loss measures the error between the prediction frame of the grape target and the GroundTruth; the smaller the loss function value, the smaller the error, and the more accurate the prediction. Confidence loss is a measure of the probability that the grape target exists in the region of interest, and the smaller the value of the loss function, the higher the precision. The classification loss represents the ability of the algorithm to correctly predict the grape category; the smaller the loss value, the more accurate the classification. In this study, there was only one grape category and no other categories, so the value was 1. As shown in Figure 6, the loss value decreased rapidly in the first 50 epochs of grape cluster detection model training, and the training curve converged faster. During this period, the precision, recall and mAP increased rapidly, indicating that the model learning efficiency was high. With the deepening of training, the slope of the training curve gradually decreased, and basically stabilized after 280 epochs of training. The loss fluctuated around 0.011. Therefore, the model output after 300 epochs of training was determined as the target recognition model for this study.

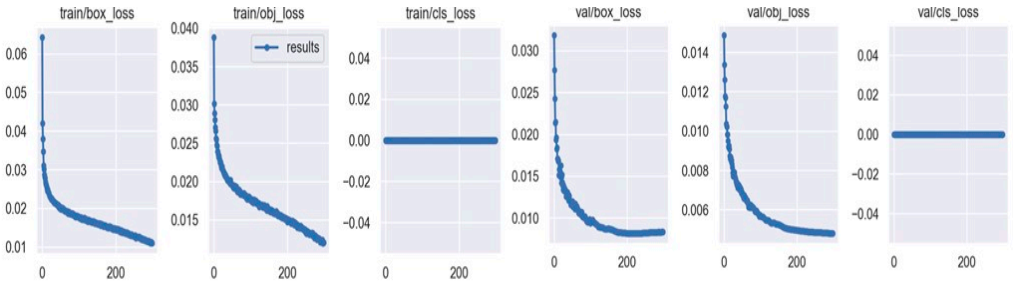


**Figure 6.** Loss curve during training and validation.

The detection effect of the trained YOLOv5s algorithm on grape clusters is shown in Figure 7. It is clear that the algorithm could completely detect the grape clusters in the test set images, indicating that the algorithm effectively detects grape clusters.

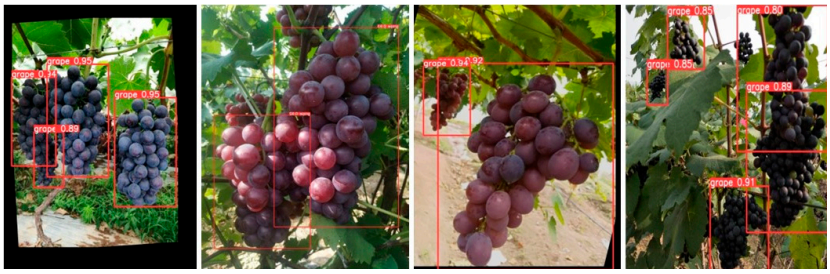

**Figure 7.** Grape cluster detection results based on YOLOv5s algorithm.

### *2.3. Model Performance Ealuation*

In this study, five indicators of precision, recall, mAP (mean average precision), F1, detection speed and model scale were used to evaluate the performance of the models. The calculations for precision, recall, mAP and F1 are shown in Equations (1)–(4).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

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

$$\text{mAP} = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Here, $TP$ is the number of correctly identified grapes, $FP$ is the number of incorrectly identified grapes, $FN$ is the number of unidentified grapes, and $C$ is the number of grape categories. Since only grape clusters are detected, and there is only one category, $C = 1$ was used in this study.

## 3. Results

### *3.1. Grape Cluster Target Detection Results and Analysis*

To verify the performance of the proposed algorithm, the grape clusters in the 1732 test set images were tested; there were 4144 grape cluster targets in the 1732 test set images. As shown in Table 3, the precision, recall, mAP and F1 of the proposed method were 99.40%, 99.40%, 99.40% and 99.40%, respectively. The detection speed was 344.83 fps, and the size of the detection network was only 13.67 MB. The results showed that the proposed method possesses high precision, a small network scale and real-time detection speed, which can provide a technical reference for the deployment of the grape cluster detection model in mobile terminals.

**Table 3.** Detection results of grape clusters.

| Evaluation Indicator | Precision (%) | Recall (%) | mAP (%) | F1 (%) | Detection Speed (fps) | Size (MB) |
|---|---|---|---|---|---|---|
| Results | 99.40 | 99.40 | 99.40 | 99.40 | 344.83 | 13.67 |

To evaluate the prediction performance of the model, especially on new data, we conducted a 5-fold cross validation by dividing 8657 grape images into 5 non-repeated subsets, and selecting a different subset as the validation images, and the remaining 4 subsets as the training images. As a result, the precision, recall, mAP and F1 values of the 5-fold cross-validation were 99.54%, 99.42%, 99.44% and 99.48% respectively, consistent with Table 3, which indicated that YOLOv5s is a reliable and stable network.

### *3.2. Comparison of Different Target Detection Algorithms*

To further verify the performance of the YOLOv5s algorithm, the YOLOv5s network was compared with the YOLOv5x [25], the ScaledYOLOv4-CSP [30] and the YOLOv3 [24] networks on the test set of 1732 images. Precision, recall, mAP, F1, detection speed, size and parameters were used as evaluation indicators. Performance comparison of four end-to-end object detection networks for grape cluster detection are shown in Figure 8.

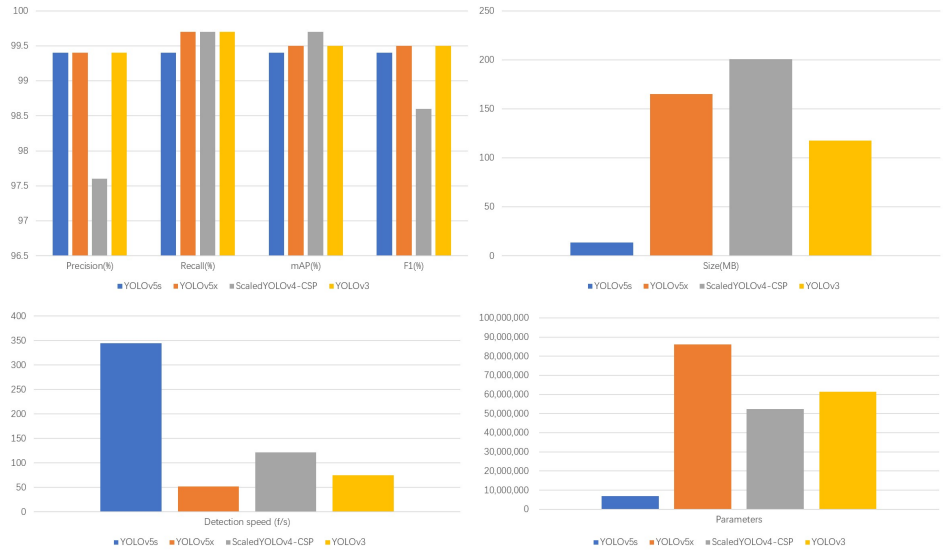


**Figure 8.** Performance comparison of four end-to-end object detection networks for grape cluster detection.

Figure 8 shows that the precision, recall, mAP and F1 values of YOLOv5s were 99.40%, 99.40%, 99.40% and 99.40%, respectively, which were close to the detection performance of the other three networks. The precision was 1.84% higher than that of ScaledYOLOv4-CSP. The mAP was slightly lower than the other three networks by 0.1–0.3%, indicating that the four networks perform similarly in the recognition of grape clusters in the dataset Gape-internet. For the recognition speed of the model, the inference time per image of YOLOv5s was 2.9 ms (344.83 fps), which was the fastest among the four networks (4.6 times, 2.83 times and 6.7 times that of the YOLOv3, ScaledYOLOv4-CSP and YOLOv5x networks, respectively). Although YOLOv5s and ScaledYOLOv4-CSP met the real-time requirements for grape cluster detection, YOLOv5s had more advantages. The data showed that the speed of the network was improved by reducing the depth and width of the network [30]. Moreover, in terms of network model scale, the size of the YOLOv5s network model was only 13.67 MB, and the number of parameters was $0.7 \times 10^7$, which are much smaller than those of other three network models; the size of the Scaled YOLOv4-CSP network model was 200.74 MB and the number of parameters was $5.25 \times 10^7$, the size of the YOLOv3 network model was 117.72MB, and the number of parameters was $6.15 \times 10^7$, and the YOLOv5x network model size was 165.01 MB, and the number of parameters was $8.62 \times 10^7$. It is obvious that the YOLOv5s network had the smallest scale and met the lightweight requirements for grape cluster detection while ensuring the precision of the network.

Compared with the other three network algorithms, the precision, recall, mAP and F1 values of the YOLOv5s network were not the highest, but they were almost the same. There was a loss of precision caused by the reduction of network depth and width, in exchange for an increase in network speed and a reduction in network size. Compared with the speed increase and scale reduction, the precision loss appears negligible. In general, the greater the depth and width of the network model, the stronger the expression and learning ability of the network, and the better the performance. When the performance reaches a certain level, the performance increase caused by the continuous increase of network depth and width is not obvious, while the number of calculations and number of parameters will increase significantly. Greatly increasing the depth and width of the network will not make the network more practical. Balancing the depth and width of the network is an effective way to optimize learning ability while still considering speed and precision. Considering the advantage of YOLOv5s, the high precision, high speed and lightweight detection for grape cluster in small channels can be realized while reducing the depth and width of the network.

## 4. Discussion

Different from apples [19], mangos [20], kiwifruits [21] and other fruits, grape clusters exhibit differences not only in color, size, shape and compactness of different species, but also within different growth periods of the same species [13]. In addition, grapes are grown in open environments, meaning that there are many interference factors such as light changes, overlaps, shadows and occlusions. These inter-grape differences, intra-grape differences, image quality differences, and complex environmental changes have great impact on the detection effects of grape detection networks. Due to these factors, accurate grape cluster detection has always been a challenging task. As the experimental images were all derived from the internet, the differences between images are more obvious in terms of inter-product, intra-product, shooting background, resolution, illumination and occlusion, which greatly increases the difficulty of detection. From the results in Table 3, the detection precision, recall rate and mAP by employing the YOLOv5s network model were all 99.4%, greater than 99%, and close to 100%, indicating that the grape cluster detection network based on YOLOv5s can be successfully adapted to the inter-product and intra-product differences of grapes. The network also successfully adapted to the changes of image quality and complex environments, which proves its strong robustness.

To verify the detection performance of the proposed algorithm in the study, the detection results of the proposed algorithm were compared with those of the algorithm proposed by Santos et al. As shown in Table 4, when the IOU threshold was 60%, the precision of YOLOv2 and YOLOv3 for grape cluster detection on the grape dataset WGISD was 55.90% and 58.70%, respectively, which was 25.51–29.06% lower than that of Mask R-CNN. Additionally, the recall rates were 45.50% and 38.90%, respectively, which was 40.05–48.75% lower than that of Mask R-CNN. The work does not reflect the advantages of the end-to-end detection network. Compared with the test results of Mask R-CNN, the detection accuracy of YOLOv5s was 26.14% higher, and the recall rate was 30.96% higher. In addition, the detection speed of the network proposed in this study reached 344.83 fps, and the size was only 13.67 MB, which met the real-time and lightweight performance requirements, and was conducive to the deployment of the network to mobile devices.

**Table 4.** Performance Comparison of End-to-End Detection Algorithms for Grape Clusters.

| Network Model | Precision (%) | Recall (%) |
|---|---|---|
| Mask R-CNN | 78.80 | 75.90 * |
| YOLOv2 | 55.90 | 45.50 |
| YOLOv3 | 58.70 | 38.90 |
| YOLOv5s (our) | 99.40 | 99.40 |

* The detection result is obtained when the IOU threshold was 60%.

The study of Santos et al. did not take into account the detection accuracy, or the lightweight and real-time performance of the grape cluster detection network at the same time. When the precision is high, the lightweight performance requirement cannot be achieved, and the pursuit of lightweight performance will come at the expense of reduced precision. In this study, YOLOv5s was used as the detection network, and a large number of data samples were used as the data set, which demonstrated that the network was able to meet the high precision, high speed and lightweight performance requirements for grape cluster detection.

## 5. Conclusions

To achieve high-precision and high-speed intelligent detection of grape clusters in complex natural environments, in this study, a real-time detection method for grape clusters based on the YOLOv5s algorithm was proposed. The main conclusions were as follows:

1. The precision, recall, mAP and F1 of the proposed method were 99.40%, 99.40%, 99.40% and 99.40%, respectively, and the detection speed reached 344.83 fps. The network

model size was 13.67 MB. This shows that the grape cluster detection network based on YOLOv5s can be adapted to the inter-product and intra-product differences of grapes, as well as proving its strong robustness by adapting to the changes of image quality and complex environments.

2. Compared with the YOLOv5x, ScaledYOLOv4-CSP and YOLOv3 networks, the precision of the YOLOv5s network was 1.84% higher than that of the ScaledYOLOv4-CSP. While the recall rate and mAP were slightly lower than the other three networks by 0.1–0.3%, the YOLOv5s network was the fastest and the smallest among the four networks. This shows that the YOLOv5s network can meet the precision and real-time detection performance requirements for grape cluster detection by properly balancing the depth and width of the network.

In short, YOLOv5s can achieve high precision, high speed and lightweight grape cluster detection in small channels, and can provide certain technical reference for the application of grape cluster detection models in small mobile terminals. Because the network was trained based on red grape varieties, it is still difficult to accurately identify grape varieties with the similar-background color grapes or early-growing red grapes, which is the limitation of the proposed network. In future, the similar-background color grapes will be included in the tested varieties to meet the production estimation needs of multiple grape varieties in the same vineyard.

**Author Contributions:** Conceptualization, C.Z. and Y.W.; methodology, C.Z., H.D. and Y.W.; software, C.Z.; validation, C.Z. and Q.S.; formal analysis, C.Z., H.D. and Y.W.; investigation, C.Z. and H.D.; resources, Y.W.; data curation, C.Z. and Q.S.; writing—original draft preparation, C.Z.; writing—review and editing, C.Z., H.D. and Y.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** Key research and development plan project OF Jining city (No. 2021ZDZP025).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We thank all of the founders and all of the reviewers.

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

## References

1. Torres-Sánchez, J.; Mesas-Carrascosa, F.J.; Santesteban, L.-G.; Jiménez-Brenes, F.M.; Oneka, O.; Villa-Llop, A.; Loidi, M.; López-Granados, F. Grape cluster detection using UAV photogrammetric point clouds as a low-cost tool for yield forecasting in vineyards. *Sensors* **2021**, *21*, 3083. [CrossRef]
2. Gennaro, S.F.D.; Toscano, P.; Cinat, P.; Berton, A.; Matese, A. A low-cost and unsupervised image recognition methodology for yield estimation in a vineyard. *Front. Plant Sci.* **2019**, *10*, 559. [CrossRef] [PubMed]
3. Marani, R.; Milella, A.; Petitti, A.; Reina, G. Deep neural networks for grape bunch segmentation in natural images from a consumer-grade camera. *Precis. Agric.* **2021**, *22*, 387–413. [CrossRef]
4. Liu, S.; Whitty, M. Automatic grape bunch detection in vineyards with an SVM classifier. *J. Appl. Log.* **2015**, *13*, 643–653. [CrossRef]
5. Liu, S.; Cossell, S.; Tang, J.; Dunn, G.; Whitty, M. A computer vision system for early stage grape yield estimation based on shoot detection. *Comput. Electron. Agric.* **2017**, *137*, 88–101. [CrossRef]
6. Cecotti, H.; Rivera, A.; Farhadloo, M.; Pedroza, M.A. Grape detection with convolutional neural networks. *Expert Syst. Appl.* **2020**, *159*, 113588. [CrossRef]
7. Ghiani, L.; Sassu, A.; Palumbo, F.; Mercenaro, L.; Gambella, F. In-Field automatic detection of grape bunches under a totally uncontrolled environment. *Sensors* **2021**, *21*, 3908. [CrossRef]
8. Aquino, A.; Diago, M.P.; Millán, B.; Tardáguila, J. A new methodology for estimating the grapevine-berry number per cluster using image analysis. *Biosyst. Eng.* **2017**, *156*, 80–95. [CrossRef]
9. Nuske, S.; Wilshusen, K.; Achar, S.; Yoder, L.; Narasimhan, S.; Singh, S. Automated visual yield estimation in vineyards. *J. Field Robot.* **2014**, *31*, 837–860. [CrossRef]
10. Badeka, E.; Kalabokas, T.; Tziridis, K.; Nicolaou, A.; Vrochidou, E.; Mavridou, E.; Papakostas, G.A.; Pachidis, T. Grapes Visual Segmentation for Harvesting Robots Using Local Texture Descriptors. *Comput. Vis. Syst.* **2019**, *11754*, 98–109. [CrossRef]

11. Luo, L.; Tang, Y.; Lu, Q.; Chen, X.; Zhang, P.; Zou, X. A vision methodology for harvesting robot to detect cutting points on peduncles of double overlapping grape clusters in a vineyard. *Comput. Ind.* **2018**, *99*, 130–139. [CrossRef]
12. Pérez-Zavala, R.; Torres-Torriti, M.; Cheein, F.A.; Troni, G. A pattern recognition strategy for visual grape bunch detection in vineyards. *Comput. Electron. Agric.* **2018**, *151*, 136–149. [CrossRef]
13. Santos, T.T.; de Souza, L.L.; dos Santos, A.A.; Avila, S. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Comput. Electron. Agric.* **2020**, *170*, 105247. [CrossRef]
14. Milella, A.; Marani, R.; Petitti, A.; Reina, G. In-field high throughput grapevine phenotyping with a consumer-grade depth camera. *Comput. Electron. Agric.* **2019**, *156*, 293–306. [CrossRef]
15. Grimm, J.; Herzog, K.; Rist, F.; Kicherer, A.; Töpfer, R.; Steinhage, V. An adaptable approach to automated visual detection of plant organs with applications in grapevine breeding. *Biosyst. Eng.* **2019**, *183*, 170–183. [CrossRef]
16. Zabawa, L.; Kicherer, A.; Klingbeil, L.; Töpferc, R.; Kuhlmanna, H.; Roscher, R. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2020**, *164*, 73–83. [CrossRef]
17. Aguiar, A.S.; Magalhães, S.A.; dos Santos, F.N.; Castro, L.; Pinho, T.; Valente, J.; Martins, R.; Boaventura-Cunha, J. Grape bunch detection at different growth stages using deep learning quantized models. *Agronomy* **2021**, *11*, 1890. [CrossRef]
18. Yin, W.; Wen, H.; Ning, Z.; Ye, J.; Dong, Z.; Luo, L. Fruit detection and pose estimation for grape cluster–harvesting robot using binocular imagery based on deep neural networks. *Front. Robot. AI* **2021**, *8*, 626989. [CrossRef]
19. Yan, B.; Fan, P.; Lei, X.; Liu, Z.; Yang, F. A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5. *Remote Sens.* **2021**, *13*, 1619. [CrossRef]
20. Bargoti, S.; Underwood, J. Deep fruit detection in orchards. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3626–3633. [CrossRef]
21. Yao, J.; Qi, J.; Zhang, J.; Shao, H.; Yang, J.; Li, X. A Real-Time Detection Algorithm for Kiwifruit Defects Based on YOLOv5. *Electronics* **2021**, *10*, 1711. [CrossRef]
22. Tzutalin. LabelImg. Available online: https://github.com/tzutalin/labelImg (accessed on 15 June 2022).
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, Real-time ObjectDetection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Ultralytics. Yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 15 June 2022).
26. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
27. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
28. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
30. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Scaled-YOLOv4: Scaling Cross Stage Partial Network. *arXiv* **2011**, arXiv:2011.08036.