agriculture

MDPI

Article

MSGV-YOLOv7: A Lightweight Pineapple Detection Method

Rihong Zhang [1], Zejun Huang [1], Yuling Zhang [2], Zhong Xue [3,*] and Xiaomin Li [1]

[1] College of Mechanical and Electrical Engineering, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China; zhangrihong@zhku.edu.cn (R.Z.); huangzejun@zhku.edu.cn (Z.H.); lixiaomin@zhku.edu.cn (X.L.)
[2] Shantou Agricultural Product Quality and Safety Center, Shantou 515071, China; gmc932@163.com
[3] South Subtropical Crops Research Institute, Chinese Academy of Tropical Agricultural Sciences, Zhanjiang 524091, China
* Correspondence: hbj46@catas.cn

Abstract: In order to optimize the efficiency of pineapple harvesting robots in recognition and target detection, this paper introduces a lightweight pineapple detection model, namely MSGV-YOLOv7. This model adopts MobileOne as the innovative backbone network and uses thin neck as the neck network. The enhancements in these architectures have significantly improved the ability of feature extraction and fusion, thereby speeding up the detection rate. Empirical results indicated that MSGV-YOLOv7 surpassed the original YOLOv7 with a 1.98% increase in precision, 1.35% increase in recall rate, and 3.03% increase in *mAP*, while the real-time detection speed reached 17.52 frames per second. Compared with Faster R-CNN and YOLOv5n, the *mAP* of this model increased by 14.89% and 5.22%, respectively, while the real-time detection speed increased by approximately 2.18 times and 1.58 times, respectively. The application of image visualization testing has verified the results, confirming that the MSGV-YOLOv7 model successfully and precisely identified the unique features of pineapples. The proposed pineapple detection method presents significant potential for broad-scale implementation. It is expected to notably reduce both the time and economic costs associated with pineapple harvesting operations.

Keywords: MSGV-YOLOv7; pineapple; deep learning; thin neck; target detection

Citation: Zhang, R.; Huang, Z.; Zhang, Y.; Xue, Z.; Li, X. MSGV-YOLOv7: A Lightweight Pineapple Detection Method. *Agriculture* **2024**, *14*, 29. https://doi.org/10.3390/agriculture14010029

Academic Editors: Heye Bogena, Cosimo Brogi, Christof Huebner and Andreas Panagopoulos

Received: 3 November 2023
Revised: 21 December 2023
Accepted: 21 December 2023
Published: 23 December 2023

1. Introduction

China, a prominent agricultural country with expansive farmlands, necessitates a more proficient and large-scale approach to agricultural production [1]. Under this context, the advent of smart agriculture is an inevitable trend for the future. Pineapples, a nutritious fruit extensively cultivated in the southern regions of China [2], are typically harvested manually. This process requires a significant labor force and faces issues such as labor shortages and concerns about worker safety. Therefore, the development of pineapple harvesting robots is becoming increasingly important [3]. These robots have the potential to improve both yield and harvesting efficiency while also ensuring the safety of laborers. This highlights the pressing need for research in the field of pineapple harvesting robotics [4,5]. For these robots to be effective and of high quality, real-time detection of pineapples is crucial in the effectiveness and quality of robot-assisted harvesting.

In recent years, there have been rapid advancements in deep learning technologies. This development has led to the widespread application of machine vision technologies. Now, they have become the primary tools for detecting agricultural products. These technologies have found extensive applications in remote intelligent monitoring [6,7], agricultural harvesting robots [8–11], ripeness detection [12], variety selection [13,14], yield prediction [15], and other related fields. As the field of object recognition technology continues to mature, experts from both domestic and international communities have increasingly integrated computer vision technology into the agricultural sector. Researchers

have dedicated their efforts to exploring precise fruit recognition algorithms, including traditional feature-based recognition methods as well as deep learning methods such as convolutional neural networks, which have exhibited promising results in agricultural applications. These algorithms can generally be categorized into single-stage and two-stage detection methods. Single-stage detection algorithms necessitate only one round of feature extraction for object detection. Examples of such algorithms include the Single Shot Multibox Detector (SSD) [16,17] and You Only Look Once (YOLO) [18–21] series, renowned for their swift detection speed and exceptional computational performance. Therefore, these algorithms are well suited for the real-time monitoring of agricultural targets in complex environments. Researchers have successfully enhanced these algorithms to achieve high-precision fruit detection in complex agricultural settings. For instance, Angelo Cardellicchio and his team [22] evaluated the YOLOv5 model using challenging datasets of tomato plants, demonstrating commendable performance in the recognition of leaf nodes, fruits, and flowers. Sekharamantry et al. [23] improved the precision of YOLOv5 by introducing adaptive pooling, attribute enhancement, and loss functions, resulting in proficient apple detection. Gai et al. [24] proposed a YOLOV4-DenseNet model with dense connections for cherry detection, effectively addressing the challenge of recognizing small target fruits. Tang et al. [25] developed a tea oil fruit detection algorithm based on YOLOv4-Tiny, mitigating issues related to fruit detection caused by lighting and leaf shadow, thereby achieving a 4.86% enhancement in average precision (AP) and a 12% reduction in model size, with further potential for reductions. Huang et al. [26] devised a GCS-YOLOv4-Tiny multi-stage fruit detection algorithm based on YOLOv4-Tiny, which improved the *mAP* by 17.45% and F1 by 13.8% compared with the original network. In contrast, two-stage object detection algorithms, such as Faster R-CNN, involve generating candidate regions and subsequently employing convolutional neural networks for object detection. Zhong et al. [27] enhanced the Faster R-CNN model by incorporating depth information to accurately locate clustered peppers, achieving an AP of 87.30%. Fangfang Gao et al. [28] employed Faster R-CNN for apple detection, attaining an average precision of 0.879 for apples under various obstructions, including leaves, branches, wires, and other fruits. In summary, single-stage object detection methods are well suited for tasks that necessitate real-time performance, while two-stage object detection methods often enhance detection precision at the expense of increased computational resources and time consumption.

In recent years, the emergence of lightweight convolutional models has been observed as a result of the continuous advancement of convolutional neural networks [29]. These models aim to achieve lightweight effects by reducing the computational load during the convolution process. For instance, Zhang et al. [30] devised a model that enhanced YOLOv4 by integrating MobileNetV3 with lightweight attention mechanisms to detect individual potatoes in diverse environments. This model accomplished a detection time of 43 milliseconds and an average precision of 91.4%. Additionally, Zhang et al. [31] introduced a pineapple detection approach based on SSD, wherein MobileNet was substituted with VGG16 for feature extraction to optimize detection speed. In addition, Liu et al. [32] successfully identified pineapples by incorporating improved YOLOv3 models, such as DenseNet.

Despite the robust detection precision offered by current fruit detection methods utilizing deep learning technologies, they still encounter several challenges, including complex network structures, a multitude of parameters, and slower system operation [33–35]. Therefore, this study seeks to design a real-time pineapple detection model for agricultural harvesting robots, proposing a lightweight deep learning model known as MSGV-YOLOv7. The objective is to strike a balance between detection precision and inference speed, thereby addressing the complexities of pineapple farming environments and providing significant support for the advancement of automated pineapple harvesting technology. The primary enhancements of this study are outlined as follows:

1. To enhance adaptability to mobile devices, the MobileOne lightweight network was introduced as a replacement for the YOLOv7 backbone network with the objective of diminishing parameter count and expediting model training speed.

2. The original network architecture was substituted with the lightweight GSConv and VoVGSCSP backbone designs, resulting in reduced computational complexity and network structure complexity while still maintaining satisfactory object detection precision.
3. The SimSPPF module was incorporated to enhance and optimize the SPP structure. Its uniqueness lies in the gradual pooling approach as opposed to simultaneous pooling at three scales. This modification effectively improves the efficiency of object feature extraction and candidate box selection.

2. Materials and Methods

2.1. Data Acquisition

The pineapple images assessed in this study were mainly obtained from a pineapple plantation in Xuwen County, Zhanjiang City, Guangdong Province, China. The geographical coordinates of the plantation are approximately 110.4 degrees longitude and 20.46 degrees latitude. These images were captured between 12 April and 18 April 2023 using the Intel RealSense D435i depth camera. The technical features of this camera include high-precision depth sensing and the capability to capture color images with a wide dynamic range. The captured raw images have a resolution of 640 × 640 pixels, aiding in the detailed capture of pineapple features. To reduce the risk of overfitting in the neural network model due to limited diversity in training samples, a variety of pineapple types were collected. During the image collection process, we conducted photography at various time intervals and angles. Specifically, the shooting spanned from 8 AM to 6 PM, and the angles covered a complete range from 0 to 360 degrees. Through this method, we collected a total of 3210 pineapple images. To minimize the impact of duplicate images during model training, a manual preprocessing step was performed on the raw image data. This included identifying and removing images irrelevant to the study, such as excessively blurred images, those containing other crops or non-agricultural scenes, and those with poor lighting conditions. As a result, data redundancy was reduced and image interference detrimental to the targeted analysis was eliminated. The objectives included enhancing the precision of pineapple detection under field conditions and covering a wide range of growth stages and environmental variations in the pineapple dataset. Accordingly, 1720 pineapple images were selected for the final dataset. To ensure diversity within the image data, the collected images include scenes with multiple pineapples, single pineapples, occlusions, and backlight. Figure 1 presents examples from the pineapple dataset.

(**a**) (**b**) (**c**) (**d**)

Figure 1. Pineapple dataset examples. (**a**) Multiple pineapples; (**b**) single pineapple; (**c**) occlusions; (**d**) backlight.

2.2. Data Augmentation

Data augmentation is an effective method that expands the dataset and enhances sample diversity by transforming and augmenting existing data, thereby improving the generalization ability of the model and reducing the risk of overfitting. In this study, 265 images were randomly selected as a test set from a total of 1720 images. To better balance sample diversity, computational resources, and training time while adhering to the 9:1 ratio principle, 930 out of the remaining 1455 images were randomly chosen for data augmentation. The augmentation process involves diverse methods, including mixed

augmentation, rotation adjustment, scaling transformation, noise injection, and brightness adjustment, all of which are detailed in Figure 2. These steps not only expanded the dataset but also ensured a balanced distribution and robustness of samples across all datasets. This method aids in a comprehensive and impartial assessment of model performance and effectively optimizes and improves the model during the training phase. Ultimately, the training set, validation set, and test set contained 2147, 238, and 265 samples, respectively; meanwhile, a strict 9:1 ratio was maintained between the training set and validation set, as well as between the combined training and validation sets and the test set. The description of the pineapple data expansion parameters is detailed in Table 1.

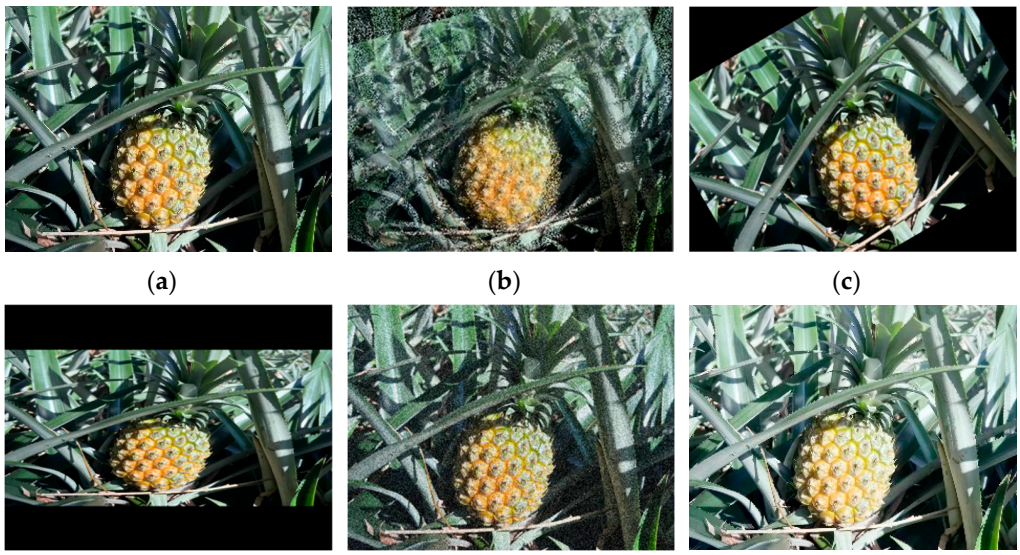

Figure 2. Pineapple image data augmentation techniques. (**a**) Original image; (**b**) mixed augmentation; (**c**) image rotation; (**d**) image scaling; (**e**) addition of Gaussian noise; (**f**) brightness enhancement.

Table 1. Pineapple data expansion parameter description.

Data	Original			Data Enhancement		
	Training	Validation	Testing	Training	Validation	Testing
Single pineapple	314	55	69	554	55	69
Multiple pineapples	305	62	73	538	62	73
Occlusion	302	57	58	531	57	58
Exposure/Backlight	296	64	65	524	64	65
Total	1217	238	265	2147	238	265

2.3. Experimental Platform and Parameter Configuration

This article discusses the training and testing procedures employed in the development of a pineapple identification model. Both phases were conducted in an identical environment, as presented in the table. The model was trained on a Windows 11 operating system, utilizing the Pytorch 1.8.2 framework for training and testing purposes. The computer for this study was equipped with an Intel(R) Core(TM) i5-12400 processor, 64 GB of memory, and an NVIDIA GeForce RTX 3060 12GB graphics card. In addition, the CUDA 11.1 parallel computing framework and CUDNN 8.2 deep neural network acceleration library were installed. The input image dimensions were set at 640 × 640 pixels with a batch size of 16. The training process encompassed 300 steps, employing a learning rate of 0.01 and a momentum value of 0.937. Stochastic gradient descent (SGD) optimization incorporating a weight decay of 0.005 was employed. The environmental configuration of the experimental platform is presented in Table 2.

Table 2. Test platform environment configuration.

Hardware Configuration or Software Environment	Model or Name	Reference or Version
CPU	Intel(R) Core(TM) i5-12400	Clock Speed of 2.5 GHz
GPU	NVIDIA GeForce RTX 3060	VRAM 12GB
Computing system	Windows	11
Network framework	Pytorch	1.8.2
Computing Architecture	CUDA	11.1
Compiler	Pycharm	2022.1.3
Compiled language	Python	3.9

2.4. Methodology

2.4.1. Standard YOLOv7 Model

The selection of the YOLOv7 algorithm in this study was motivated by its ability to strike a harmonious balance between speed and precision, which aligns with the primary objective of accurately detecting the real-time conditions of field pineapples. YOLOv7 introduces the innovative extended ELAN architecture, which enhances the self-learning capabilities of the network without disrupting the original gradient path. The ELAN architecture primarily comprises VoVNet and CSPNet, which optimize the gradient length of the entire network through a stacked structure in the computational block. This optimization facilitates more effective learning and convergence in deeper networks. Moreover, YOLOv7 incorporates a cascade-based model scaling method, which dynamically adjusts the model size to cater to specific detection requirements. The primary purpose of model scaling is to modify certain attributes of the model and generate models at different scales to accommodate varying inference speeds. The network structure of YOLOv7 is visually depicted in Figure 3.

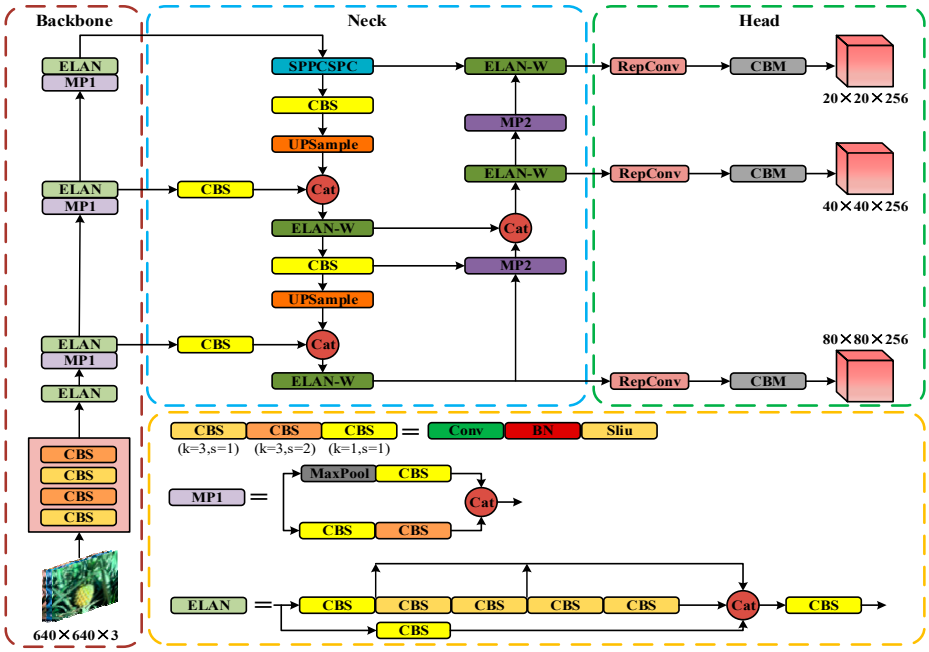

Figure 3. The network model structure of YOLOv7. ELAN, SPPCSPC, CBS, MP1, etc., represent different types of layers or operations, each with a specific function, such as spatial pyramid pooling (SPPCSPC) or max pooling (MP1).

2.4.2. MSGV-YOLOv7

To enhance the precision and efficiency of real-time pineapple detection, this study opted to substitute the YOLOv7 backbone network with MobileOne, thereby reducing the parameter count of the network. Moreover, a lightweight neck model, namely thin

neck, was devised. In this model, the E-ELAN module was replaced by the VoVGSCSP module, thereby reducing superfluous computational and memory usage. The integration of the GSConv module aimed to alleviate the computational burden while upholding a high level of precision. Additionally, to expedite the inference speed, the original SPPCSPC module was substituted with SimSPPF with the objective of reducing model complexity and augmenting real-time detection precision. The framework of the MSGV-YOLOv7 model is depicted in Figure 4.

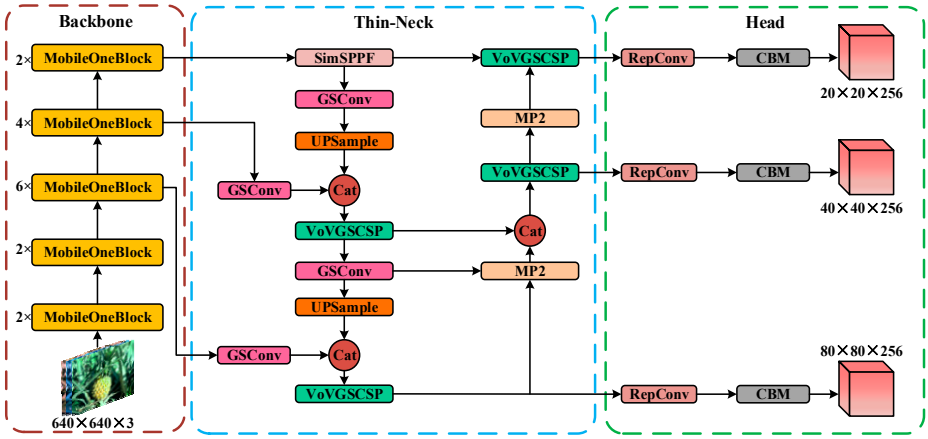

Figure 4. The network model structure of MSGV-YOLOv7.

2.4.3. MobileOne Network

MobileOne is constructed utilizing the MobileNet architecture and RepVGG convolutional neural network. While numerous conventional network backbones heavily rely on residual structures and SE (squeeze and excitation) modules for convolutional operations, MobileOne mitigates the additional computational load associated with residual structures by implementing the reparameterization technique of RepVGG. The utilization of SE modules is confined to the largest model structure, MobileOne-s4. In practical detection scenarios characterized by limited computational resources, there is a necessity to streamline the model, which involves the creation of a lightweight backbone network. The MobileOne model is designed based on MobileNetV1 and comprises a multitude of MobileOneBlocks. These MobileOneBlocks incorporate depthwise convolution and pointwise convolution, drawing inspiration from the reparameterization concepts of RepVGG.

The structure of the MobileOneBlock is illustrated in Figure 5 of this study. The visual representation on the left highlights a comprehensive building module, which consists of deep convolution and point convolution. 'Deep convolution' refers to the use of group convolutions, where the number of groups is equal to the number of input channels, ensuring that each channel is processed separately. 'Point convolution' is the term given to 1×1 convolutions, which are mainly employed to alter the number of output channels and merge channels in the resulting feature maps from deep convolutions. The deep convolution module includes layers of a 1×1 convolution, an over-parameterized 3×3 convolution, and batch normalization (BN). It is important to clarify that both 1×1 and 3×3 convolutions are types of group convolutions, with '1×1' meaning that each convolution filter processes one pixel at a time and '3×3' meaning that each filter covers a three-pixel by three-pixel area to capture more complex features. During the training phase of the network model, the MobileOne network incorporates these building blocks at the end of the training process through a parameterized technique. Owing to its lightweight structure and the over-parameterized design, the MobileOne network is recognized for its efficiency, making the MobileOne module particularly suitable for feature extraction tasks.

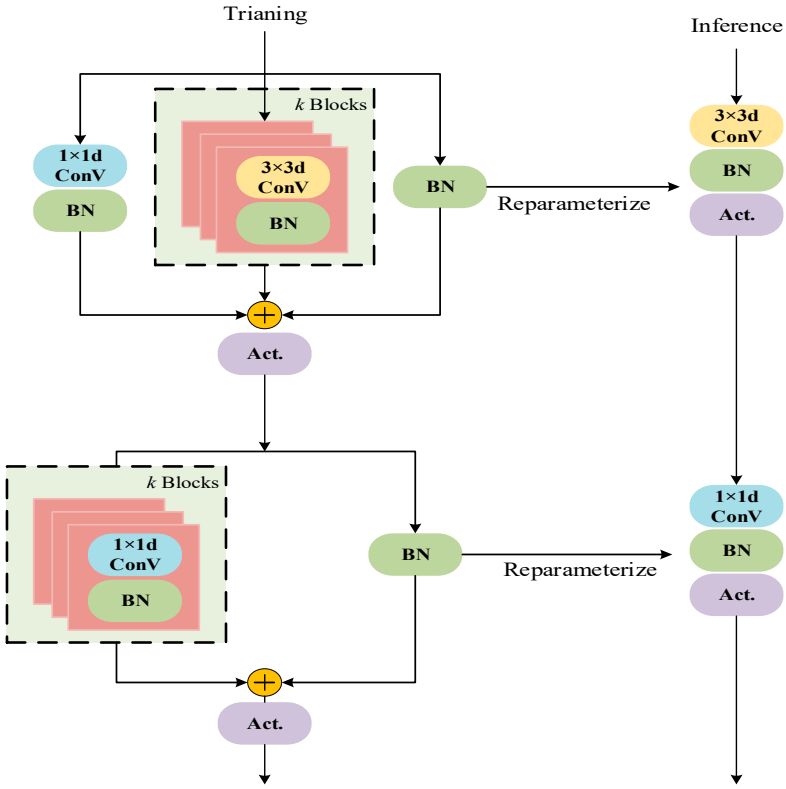

Figure 5. The structure of MobileOneblock.

2.4.4. SimSPPF

In this research, the SimSPPF module was deliberately selected for implementation based on its exceptional performance in comparison with SPPCSPC, particularly in terms of reducing computational complexity and enhancing the frame per second (Fps), all while maintaining a commendable level of precision in object detection. The SimSPPF module employs a unique approach by utilizing a series of small-sized pooling kernels arranged sequentially as opposed to a single large-sized pooling kernel, thereby resulting in an expanded perceptive field. Specifically, it sequentially processes the input through multiple max pooling layers with the dimensions of 5×5, replaces a 9×9 convolution operation with two 5×5 convolution operations, and substitutes a 13×13 convolution operation with three 5×5 convolution operations. This innovative design not only preserves the original functionalities but also reduces the computational burden, accelerates operational speed, and thus enhances the efficiency of the SimSPPF structure. A comprehensive illustration of the intricate architecture of the SimSPPF module is presented in Figure 6.

The relevant equations are shown in Equations (1)–(5).

$$F1 = CBR(F) \tag{1}$$

$$F2 = \text{Maxpooling}(F1) \tag{2}$$

$$F3 = \text{Maxpooling}(F2) \tag{3}$$

$$F4 = \text{Maxpooling}(F3) \tag{4}$$

$$F5 = CBR([F; F2; F3; F4]) \tag{5}$$

The SimSPPF module has exhibited a commendable performance in the identification of image features from the pineapple dataset, effectively reducing superfluous parameter data while reinforcing the essential textural features inherent in the images. Moreover, it has surpassed the SPPCSPC module in terms of computational speed during forward

propagation. The integration of the SimSPPF module significantly enhances the capability of the YOLOv7 model in extracting image features.

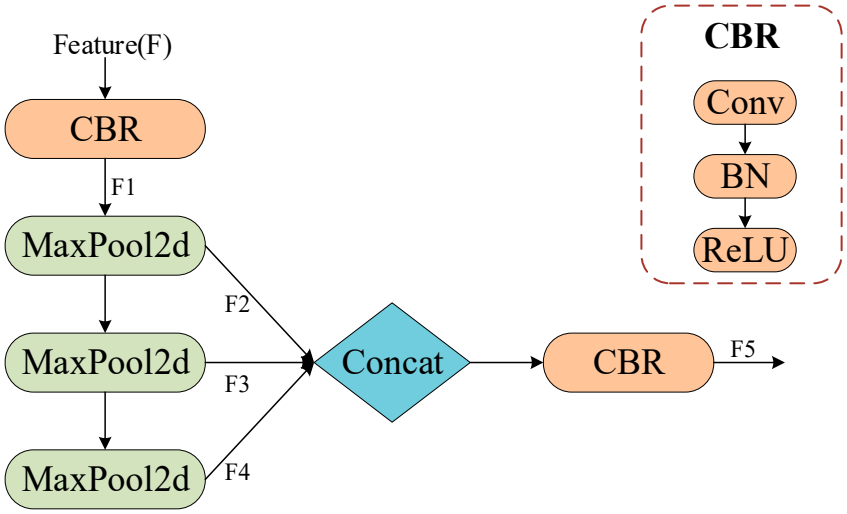

Figure 6. Structure diagram of the SimSPPF module.

2.4.5. GSConv

To expedite the speed of pineapple detection, it is necessary to minimize the time required for information processing. However, excessively reducing the model parameters may compromise the precision in recognizing the surface features of pineapples. Hence, it is crucial to strike a balance between the detection speed and the model precision. In this study, we have employed a lightweight approach to optimize the intermediate layers of the model and introduced the GSConv lightweight convolution to supplant the conventional convolution for processing the feature maps obtained from the backbone of the model. Following the processing by the GSConv convolution, the VoVGSCSP lightweight structure is employed for the upsampling and downsampling of image features. The design of the GSConv convolution structure, as illustrated in Figure 7, integrates the advantages of standard convolution and depthwise separable convolution. When processing pineapple images with complex backgrounds, the GSConv convolution structure simultaneously employs standard convolution (SConv) and depthwise separable convolution (DWConv). In contrast to DWConv, it refrains from severing the connections between channels and instead seeks to preserve them to ensure high precision for the model. The results are subsequently combined and rearranged to enhance the nonlinear expressive capabilities of the model. For the task of pineapple detection in complex agricultural fields, such a nonlinear feature extraction can more effectively capture the characteristic information of pineapples, thereby endowing the model with a plethora of learning material, hence enhancing its generalization ability and robustness. The specific mathematical expressions are calculated as follows:

$$X_c = \sigma(bn(Conv2d(X_{input}))) \tag{6}$$

$$X_{out} = \delta(X_c \oplus DWConv(X_c)) \tag{7}$$

where $Conv2d$ represents the two-dimensional convolution of the input image X_{input}, bn represents the normalization operation, σ represents the activation function, $\oplus$ represents the concating of the two kinds of convolution, and the final δ represents shuffle, aiming to obtain the final output X_{out} by shuffling this result.

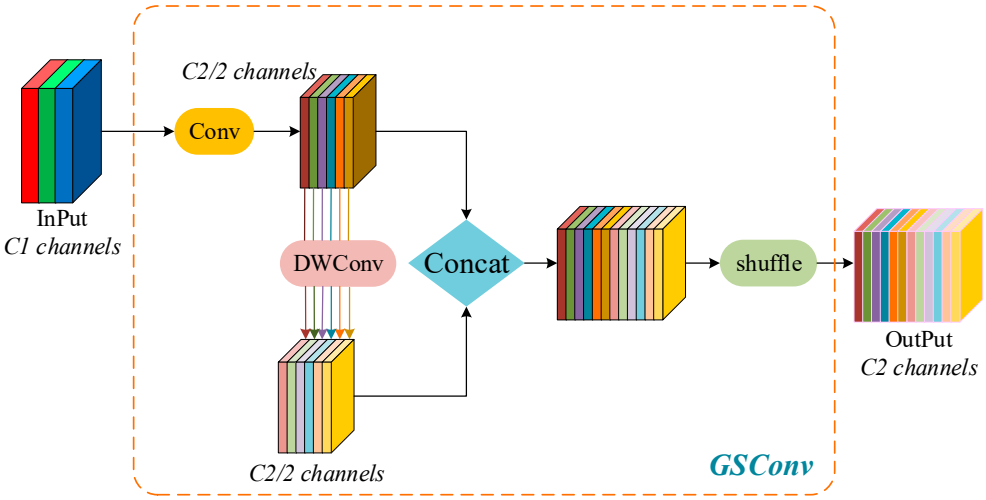

Figure 7. Structure of the GSConv module.

2.4.6. VoVGSCSP

The VoVGSCSP module, an advanced network module based on GSConv, is characterized by the continuous integration of a GS bottleneck design. The structure of this module is depicted in Figure 8. The fundamental concept behind this design is to achieve a high degree of feature reuse through a one-time aggregation method employed by the cross-stage partial network (GSCSP) module. This ensures a balance between the model precision and speed. Therefore, not only does it enhance the model precision, but it also significantly reduces the computational complexity and network structure.

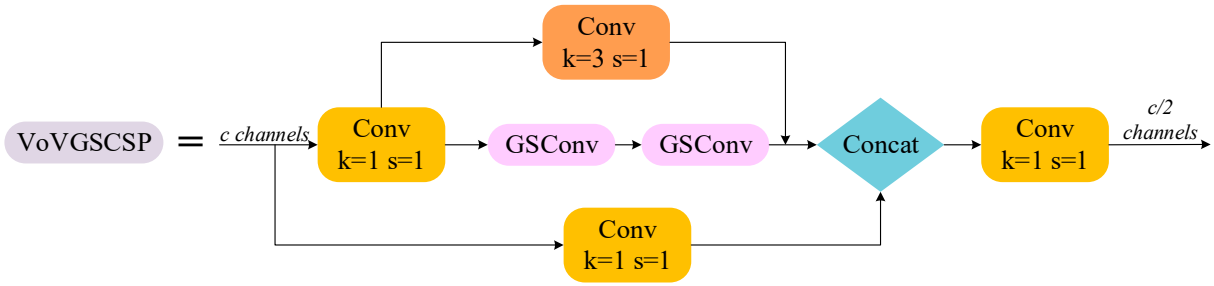

Figure 8. Architecture of VoVGSCSP.

Upon undergoing processing by GSConv, the feature maps are fed into the main body of VoVGSCSP, resulting in their channel dimensions reaching their maximum values; meanwhile, their width and height reach their minimum values. This design effectively reduces the presence of redundant repetitive information. In addition, two GSConv convolutions are employed in the main body, enabling the rapid transmission of strong semantic features and facilitating the upsampling and downsampling of feature maps. As a result, the information processing time of the entire model is shortened.

2.5. Evaluation Metrics

To evaluate the performance of the model, various metrics were employed, including precision (P), recall (R), and mean average precision (mAP). The calculations for these metrics are as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$AP = \int_0^1 P(R)dR \tag{10}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{11}$$

During the execution of the pineapple detection task, this study adopts a comprehensive array of evaluation metrics to holistically evaluate the model performance. In terms of positive sample prediction, this study classifies them into three categories: true positive (*TP*), false positive (*FP*), and false negative (*FN*). *TP* denotes the number of samples correctly identified by the model, intersecting with the true bounding box. *FP* represents the number of samples erroneously classified as positive samples that fail to intersect with the true bounding box. *FN* signifies the number of actual positive samples that the model fails to detect.

To quantify the predictive precision of the model in various pineapple-related scenarios, this paper introduces the concept of precision (*P*) as an evaluative metric. Precision is defined as the ratio of correctly predicted positive samples to the total number of predicted positive samples. Conversely, the ability of the model to detect actual positive samples is captured by the metric of recall (*R*), which is calculated as the ratio of true positives to the total number of actual positive samples.

In addition, this study also employs *AP* and *mAP* as pivotal indicators to assess the ability of the model to detect specific categories. *AP* represents the area under the precision–recall curve, while *mAP* is derived by averaging the *AP* values across all categories. This study specifically concentrates on the detection task pertaining to a singular category, denoted as $n = 1$.

To ensure efficacy of the model in real-world scenarios, particularly in the context of pineapple harvesting, it is necessary to consider its real-time performance. This comprises an analysis of the model inference time (i.e., the speed at which predictions are made) and its complexity (i.e., the number of parameters involved). This paper evaluates the model effectiveness, complexity, precision, recall, and overall performance in pineapple detection by analyzing these aforementioned factors.

3. Experimental Results and Analysis

3.1. Comparison of Experimental Results from Different Backbone Networks

To further verify the performance of the enhanced YOLOv7 model, a strategy was implemented in this study, involving the replacement of the backbone network of the YOLOv7 model with MobileOne, ShuffleNetV2, GhostNet, and MobilenetV3. Pineapple object detection experiments were conducted on the same dataset, and an in-depth comparative analysis was carried out with the improved model. The experimental data for these four lightweight models, as well as the original YOLOv7 model, are presented in Table 3.

Table 3. Comparison of experimental results with different backbone networks.

Networks	Mode Size (MB)	P (%)	R (%)	mAP (%)	Fps (f/s)
YOLOv7	37.65	93.68	91.48	93.62	42.11
YOLOv7+ShuffleNetV2	29.12	96.85	89.42	91.08	46.95
YOLOv7+GhostNet	26.08	95.60	90.91	93.46	38.82
YOLOv7+MobileNetV3	28.44	94.92	91.17	92.71	32.57
YOLOv7+MobileOne	24.71	94.86	91.11	94.61	46.70

Model size (MB): the memory size of the model file; P (%): precision—the percentage of correct positive predictions; R (%): recall—the percentage of actual positive cases correctly identified; *mAP* (%): mean average precision—a metric that combines precision and recall, where higher values indicate better performance; Fps (f/s): frames per second—the number of frames (images) the model can process in one second. Higher values indicate faster processing.

First and foremost, a key concern in this study was the number of parameters in the model, particularly when considering the deployment of the model on devices with limited computing power. In this regard, the YOLOv7+MobileOne model exhibited a commendable performance with a size of merely 24.71 MB, which is significantly smaller than the other reference models. Compared with the original YOLOv7 model, which had a size of 37.65 MB, YOLOv7+MobileOne successfully achieved a nearly 35% reduction in size. This reduction holds immense value for applications that necessitate the deployment of the model on automated pineapple harvesting robots.

With respect to object detection performance, *mAP* was selected as one of the fundamental evaluation metrics in this study. In this aspect, YOLOv7+MobileOne achieved a performance of 94.61%, second only to YOLOv7+GhostNet at 93.46%, and it achieved an increase of nearly 0.99% compared with the *mAP* of the original YOLOv7 model. This result unequivocally validates that YOLOv7+MobileOne can still deliver exceptional detection performance even when the model size is significantly reduced.

This study was centered on evaluating the real-time processing capability of the model. In this context, YOLOv7+MobileOne achieved a processing speed of 46.70 frames per second (f/s). While slightly lower than YOLOv7+ShuffleNetV2 (46.95 f/s), this figure is significantly higher than the original YOLOv7 model (42.11 f/s). This result unequivocally demonstrates the exceptional detection performance of YOLOv7+MobileOne.

Based on the aforementioned comparative results, it is evident that the overall performance of the YOLOv7+MobileOne model surpasses that of its counterparts. It is particularly well suited for detecting pineapple targets in complex field environments. Therefore, the MSGV-YOLOv7 model employed MobileOne as its lightweight backbone network. However, considering the complex nature of real-world pineapple target detection scenarios, the model necessitates an even greater performance. To this end, additional techniques such as GsConv and VoVGSCSP were introduced to enhance the model's detection precision.

3.2. Ablation Experiment

In this study, a series of ablation experiments were systematically conducted to evaluate the effectiveness of MobileOne, GSConv, VoVGSCSP, and SimSPPF modules on performance improvements of the YOLOv7 network. The experimental results are presented in Table 4. Under the same conditions, the baseline model without any added modules demonstrated certain recognition capabilities, with a precision of 93.68%, recall rate of 91.48%, and *mAP* of 93.62% when operating at 42.11 f/s. After the initial introduction of the MobileOne feature extraction network, the model precision increased to 94.86%, demonstrating a clear benefit of MobileOne in feature extraction, while the frame rate was also slightly improved to 46.70 f/s. Continuing from MobileOne, the addition of the GSConv module led to a further increase in precision to 95.96%, recall to 92.45%, and *mAP* to 95.34%. Although the frame rate slightly dropped to 45.52 f/s, in general, it indicates that the GSConv module significantly enhanced the ability of the model to capture detailed features. After the subsequent integration of the VoVGSCSP module, we observed a slight decrease in precision to 94.72%, despite the recall rate and *mAP* improving to 93.05% and 95.49%, respectively. Importantly, this enhancement was achieved while maintaining a frame rate of 46.08 f/s, which highlights the effectiveness of the VoVGSCSP module in processing complex visual information. The integration of the SimSPPF module to the model, which included MobileOne, GSConv, and VoVGSCSP, led to further improvements: the precision increased to 95.66%, the recall reached 92.83%, the *mAP* rose to 96.65%, and there was a significant improvement in the frame rate to 59.63 f/s. These results highlight the significant contribution of the SimSPPF module, particularly in enhancing the overall performance of the model in terms of the detection speed for pineapples.

Table 4. Comparison of ablation experimental results.

MobileOne	GSConv	VoVGSCSP	SimSPPF	P (%)	R (%)	mAP (%)	FPS (f/s)
-	-	-	-	93.68	91.48	93.62	42.11
√	-	-	-	94.86	91.11	94.61	46.70
√	√	-	-	95.96	92.45	95.34	45.52
√	√	√	-	94.72	93.05	95.49	46.08
√	√	√	√	95.66	92.83	96.65	59.63

'-': This symbol signifies that the corresponding technical module has not been adopted in the current model structure. '√': This symbol indicates that the corresponding technical module has been integrated into the current model structure.

3.3. Comparison of Different Object Detection Models

Through an analysis of the data presented in Table 5, a comprehensive evaluation of the performance of five object detection models—Faster R-CNN, SSD, YOLOv5n, YOLOv8l, and YOLOv7—can be conducted. Notably, the MSGV-YOLOv7 model exhibited a significantly smaller size of 13.17 MB compared with other detection models such as Faster R-CNN. Moreover, in terms of precision and recall, MSGV-YOLOv7 outperformed the aforementioned models, achieving scores of 95.66% and 92.83%, respectively. Most remarkably, the *mAP* of MSGV-YOLOv7 reached an impressive 96.65%, surpassing the performance of other detection models. Additionally, the frame rate achieved by MSGV-YOLOv7 is the highest at 59.63 f/s.

Table 5. Comparison of experimental results for different object detection models.

Experiments	Mode Size (MB)	P (%)	R (%)	mAP (%)	Fps (f/s)
Faster R-CNN	139.21	85.89	78.19	81.76	27.34
SSD	28.39	91.62	86.34	90.28	30.02
YOLOv5n	48.91	93.19	87.07	91.43	37.59
YOLOv8l	42.74	94.61	90.86	93.96	48.25
YOLOv7	37.65	93.68	91.48	93.62	42.11
MSGV-YOLOv7	13.17	95.66	92.83	96.65	59.63

The model introduced in this paper not only exhibits a significant reduction in computation time but also demonstrates remarkable enhancements in detection performance. It outperforms other models in terms of detection precision while maintaining a reasonable detection speed. Therefore, as an efficient and advantageous solution, it distinguishes itself in real-time pineapple recognition applications.

To thoroughly evaluate the performance of the MSGV-YOLOv7 model in real-time pineapple recognition, this study undertook a meticulous design of comparative experiments. These experiments involved a detailed comparison with three prominent detection models widely acknowledged in the industry: Faster R-CNN, YOLOv5n, and YOLOv7. The experimental setup comprised diverse pineapple varieties with the objective of unveiling the comprehensive performance and advantages of MSGV-YOLOv7 in real-time pineapple detection tasks.

The detection results of Faster R-CNN, YOLOv5n, YOLOv7, and MSGV-YOLOv7 on the same dataset are presented in Figure 9. Through a comparative analysis, the superior performance of MSGV-YOLOv7 in terms of precision and stability becomes evident. The figure illustrates that MSGV-YOLOv7 exhibits higher precision when confronted with complex backgrounds and small targets, such as the detection of pineapples amidst cluttered leaf backgrounds. In contrast, Faster R-CNN and YOLOv5n exhibit some instances of false positives or false negatives under similar conditions. MSGV-YOLOv7, on the other hand, successfully detects the target and provides a higher confidence level. Additionally, MSGV-YOLOv7 demonstrates a significant advantage in terms of bounding box precision. Compared with the other three models, its bounding boxes are more compact and closely aligned with the actual target, indicating a higher level of localization precision. Moreover,

the stability of MSGV-YOLOv7 is noteworthy, as depicted in the figure. It consistently maintains accurate target detection across different scenes and conditions, which is crucial for ensuring robustness in practical applications.

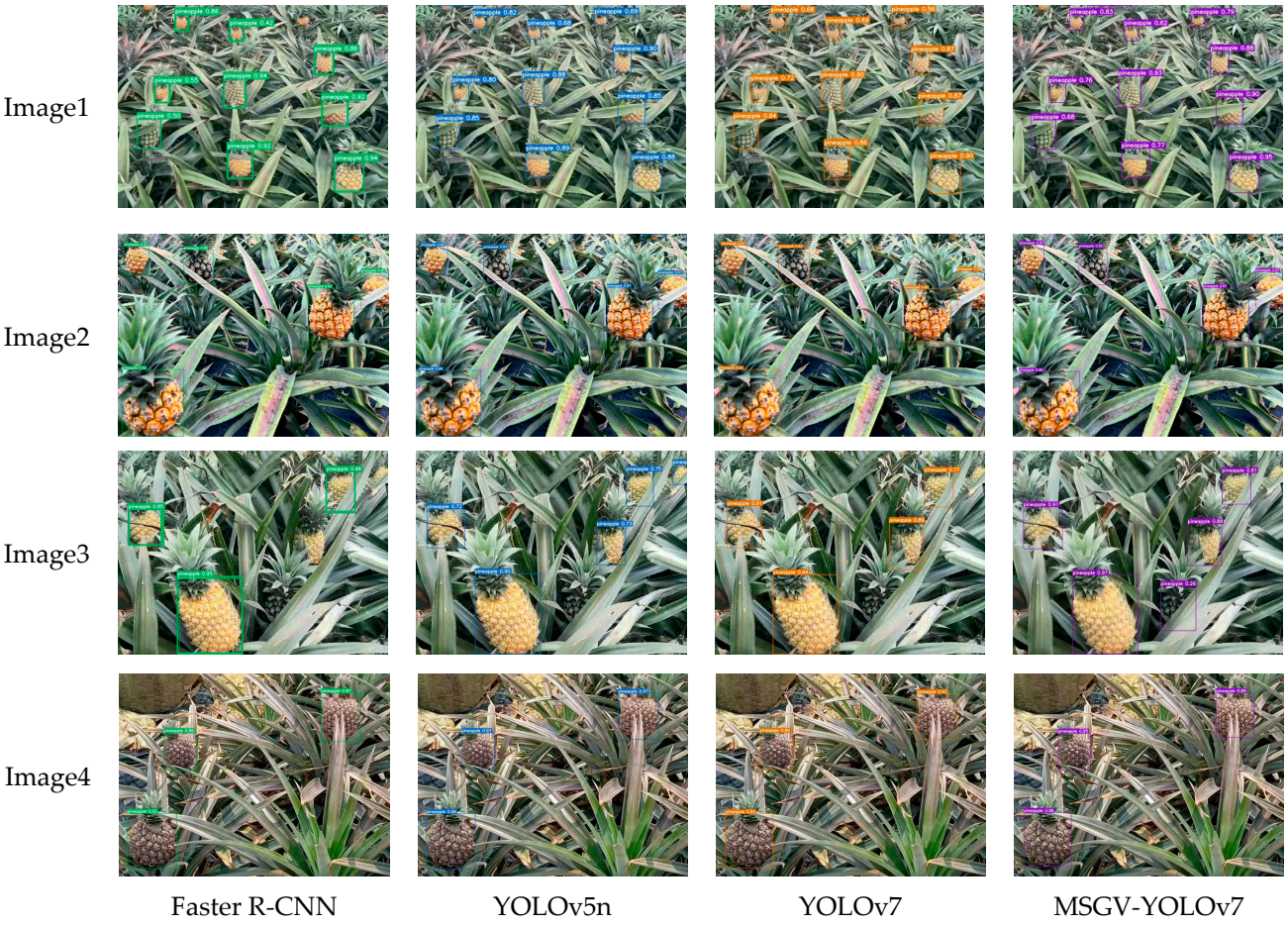

Figure 9. Comparison of detection results between MSGV-YOLOv7 and other network models.

The remarkable advantages of MSGV-YOLOv7 in terms of precision, localization precision, and stability can be attributed to its unique network architecture and optimization strategies. These enhancements provide a more dependable option for object detection tasks in practical applications. Future research is expected to further refine MSGV-YOLOv7, leading to additional breakthroughs in the field of automated pineapple harvesting.

3.4. Analysis Experiment on Feature Attributes

In this study, the utilization of Grad-CAM was implemented to produce object detection heat maps, which serve the purpose of visually highlighting significant regions in an image. These heat maps facilitate the visualization of model detection results and unveil the focal areas of interest for the model in this study. Grad-CAM, a visualization technique grounded in gradient computation, was employed to derive the weights assigned to each channel in the ultimate convolutional layer. These weights were subsequently applied to the feature map, thereby leading to a heat map superimposed upon the original image. In this heat map, the pixel values' magnitude serves as an indicator of the relative importance attributed to the corresponding region in relation to the classification decision at hand.

Figure 10 presents a comparative analysis of a heatmap, showcasing the performance of MSGV-YOLOv7 in contrast with other mainstream network models across various scenarios. The heatmap can intuitively show the focus of attention of the model when extracting features. The warmer the color, the more attention of the model, and the red part

(the warmest part) represents the focus of the model. The visual evidence clearly demonstrates the exceptional object detection capabilities of MSGV-YOLOv7. Notably, in complex situations involving multiple objects, or when objects are obscured by vegetation, MSGV-YOLOv7 displays precise target localization. Additionally, compared with other models, MSGV-YOLOv7 exhibits more concentrated and clear detection hotspots. Even in complex scenes with numerous distracting background elements, MSGV-YOLOv7 proficiently identifies the target objects. In contrast, other models, such as YOLOv5n, may sometimes produce false positives or overlook certain objects in specific situations. However, MSGV-YOLOv7 demonstrates greater stability in these aspects. Through the comparative analysis of the visualization results, the enhanced MSGV-YOLOv7 model effectively showcases its capability to extract object features, particularly in scenes with insufficient semantic information. This attests to its stronger resilience and broader applicability.

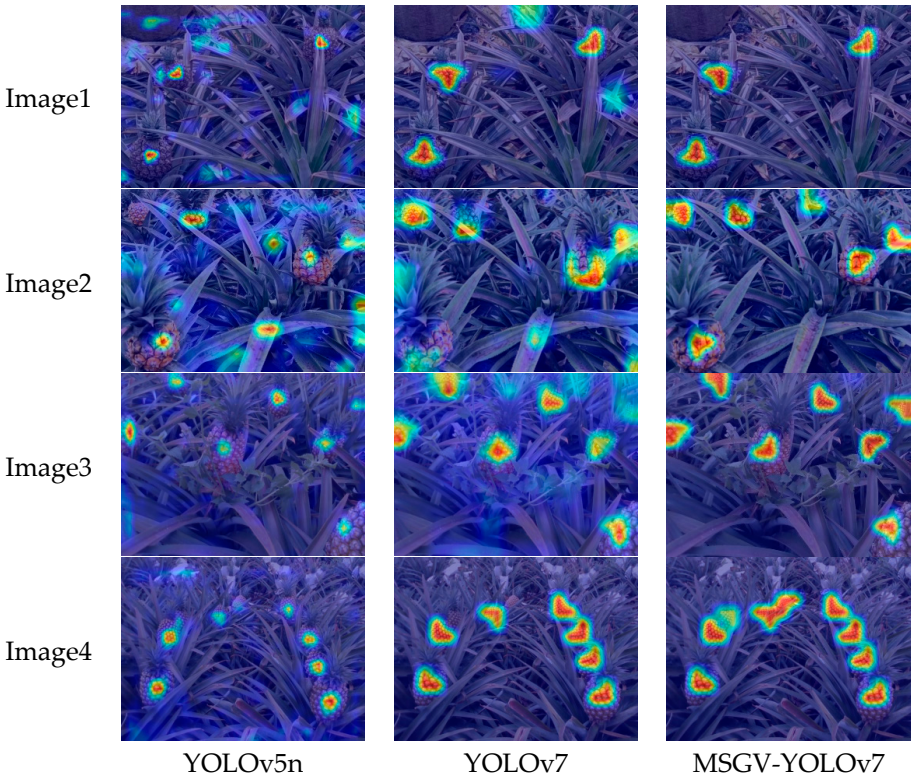

Figure 10. Heatmap for object detection in close-up images of MSGV-YOLOv7 and other network models.

4. Conclusions

To optimize the efficiency of pineapple harvesting robots in recognition and target detection, this paper proposed a lightweight pineapple detection algorithm based on YOLOv7. By integrating the innovative MobileOne backbone network and the thin neck network into the YOLOv7 algorithm, the model not only enhances its capability to capture pineapple features but also attains high computational speed and low memory consumption. The parameter size of the model is merely 13.17 MB, approximately one-third of the original model. Experiments demonstrate that the model achieves a mean average precision (*mAP*) of 96.65% on the dataset, with significant improvements in the precision (P), recall (R), and frames per second (Fps). The Grad-CAM heatmap visualization experiments confirm that the model significantly improves accuracy by focusing on pineapples, effectively resisting interference from background elements; furthermore, it is suitable for a variety of pineapple image types. Compared with other object detection algorithms, the model proposed in this article shows significant advantages in overall performance, meeting the needs for pineapple recognition in complex farm environments. Therefore, the MSGV-YOLOv7

model can provide a valuable technical reference for future applications in mobile or embedded devices, laying the foundation for the development of target detection for pineapple harvesting robots and offering valuable research solutions for similar fruit detection tasks.

Author Contributions: Conceptualization, R.Z. and Z.X.; methodology, Z.H. and R.Z.; software, Z.H.; validation, R.Z., Z.X., and Y.Z.; investigation, Z.H.; data curation, Z.H.; writing—original draft preparation, Z.H.; writing—review and editing, R.Z. and X.L.; visualization, Z.H.; supervision, R.Z. and X.L.; funding acquisition, R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Open Competition Program of Top Ten Critical Priorities of Agricultural Science and Technology Innovation for the 14th Five-Year Plan of Guangdong Province (2022SDZG03), Hainan Province Science and Technology Special Fund (ZDYF2023XDNY058), and the Characteristic Innovation Project of Guangdong University in 2022 (2022KTSCX057).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data supporting this study's findings are available from the corresponding author upon reasonable request.

Acknowledgments: The author would like to thank the editors and reviewers for their comments on improving the quality of this work.

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

References

1. Huang, Z.; Liang, Q. Agricultural organizations and the role of farmer cooperatives in China since 1978: Past and future. *China Agric. Econ. Rev.* **2018**, *10*, 48–64. [CrossRef]
2. Li, D.; Jing, M.; Dai, X.; Chen, Z.; Ma, C.; Chen, J. Current status of pineapple breeding, industrial development, and genetics in China. *Euphytica* **2022**, *218*, 85. [CrossRef]
3. Brainy, J.R.V.J.; Suganthi, K.D.N.; Narayanamoorthy, S.; Ilakiya, U.; Innab, N.; Alshammari, A.; Ahmadian, A.; Jeon, J. A perspective study for the assessment of field robots in agriculture: An enhanced fuzzy MADM approach. *Comput. Electron. Agric.* **2023**, *214*, 108296. [CrossRef]
4. Meng, F.; Li, J.; Zhang, Y.; Qi, S.; Tang, Y. Transforming unmanned pineapple picking with spatio-temporal convolutional neural networks. *Comput. Electron. Agric.* **2023**, *214*, 108298. [CrossRef]
5. Guo, A.F.; Li, J.; Guo, L.Q.; Jiang, T.; Zhao, Y.P. Structural design and analysis of an automatic pineapple picking and collecting straddle machine. *J. Phys. Conf. Series.* **2021**, *1777*, 012029. [CrossRef]
6. Wang, S.; Jiang, H.; Qiao, Y.; Jiang, S.; Lin, H.; Sun, Q. The Research Progress of Vision-Based Artificial Intelligence in Smart Pig Farming. *Sensors* **2022**, *22*, 6541. [CrossRef] [PubMed]
7. Dhanya, V.G.; Subeesh, A.; Kushwaha, N.L.; Vishwakarma, D.K.; Kumar, T.N.; Ritika, G.; Singh, A.N. Deep learning based computer vision approaches for smart agricultural applications. *Artif. Intell. Agric.* **2022**, *6*, 211–229. [CrossRef]
8. Miao, Z.; Yu, X.; Li, N.; Zhang, Z.; He, C.; Li, Z.; Deng, C.; Sun, T. Efficient tomato harvesting robot based on image processing and deep learning. *Precis. Agric.* **2023**, *24*, 254–287. [CrossRef]
9. Yang, Y.; Han, Y.; Li, S.; Yang, Y.; Zhang, M.; Li, H. Vision based fruit recognition and positioning technology for harvesting robots. *Comput. Electron. Agric.* **2023**, *213*, 108258. [CrossRef]
10. Liu, J.; Liu, Z. The Vision-Based Target Recognition, Localization, and Control for Harvesting Robots: A Review. *Int. J. Precis. Eng. Manuf.* **2023**, 1–20. [CrossRef]
11. Zhu, C.; Wu, C.; Li, Y.; Hu, S.; Gong, H. Spatial location of sugarcane node for binocular vision-based harvesting robots based on improved YOLOv4. *Appl. Sci.* **2022**, *12*, 3088. [CrossRef]
12. 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]
13. Luo, T.; Zhao, J.; Gu, Y.; Zhang, S.; Qiao, X.; Tian, W.; Han, Y. Classification of weed seeds based on visual images and deep learning. *Inf. Process. Agric.* **2023**, *10*, 40–51. [CrossRef]
14. Taheri-Garavand, A.; Nasiri, A.; Fanourakis, D.; Fatahi, S.; Omid, M.; Nikoloudakis, N. Automated in situ seed variety identification via deep learning: A case study in chickpea. *Plants* **2021**, *10*, 1406. [CrossRef] [PubMed]
15. Vignesh, K.; Askarunisa, A.; Abirami, A.M. Optimized Deep Learning Methods for Crop Yield Prediction. *Comput. Syst. Sci. Eng.* **2023**, *44*, 1051–1067. [CrossRef]
16. Wang, L.; Shi, W.; Tang, Y.; Liu, Z.; He, X.; Xiao, H.; Yang, Y. Transfer Learning-Based Lightweight SSD Model for Detection of Pests in Citrus. *Agronomy* **2023**, *13*, 1710. [CrossRef]
17. Ding, F.; Zhuang, Z.; Liu, Y.; Jiang, D.; Yan, X.; Wang, Z. Detecting defects on solid wood panels based on an improved SSD algorithm. *Sensors* **2020**, *20*, 5315. [CrossRef]

18. Wang, Y.; Yang, L.; Chen, H.; Hussain, A.; Ma, C.; Al-gabri, M. Mushroom-YOLO: A deep learning algorithm for mushroom growth recognition based on improved YOLOv5 in agriculture 4.0. In Proceedings of the 2022 IEEE 20th International Conference on Industrial Informatics (INDIN), Perth, Australia, 25–28 July 2022; pp. 239–244.
19. Lippi, M.; Bonucci, N.; Carpio, R.F.; Contarini, M.; Speranza, S.; Gasparri, A. A yolo-based pest detection system for precision agriculture. In Proceedings of the 2021 29th Mediterranean Conference on Control and Automation (MED), Bari, Italy, 22–25 June 2021; pp. 342–347.
20. Niu, Y.; Lu, M.; Liang, X.; Wu, Q.; Mu, J. YOLO-plum: A high precision and real-time improved algorithm for plum recognition. *PLoS ONE* **2023**, *18*, e0287778. [CrossRef]
21. Junos, M.H.; Mohd Khairuddin, A.S.; Thannirmalai, S.; Dahari, M. Automatic detection of oil palm fruits from UAV images using an improved YOLO model. *Vis. Comput.* **2022**, *38*, 2341–2355. [CrossRef]
22. Cardellicchio, A.; Solimani, F.; Dimauro, G.; Petrozza, A.; Summerer, S.; Cellini, F.; Renò, V. Detection of tomato plant phenotyping traits using YOLOv5-based single stage detectors. *Comput. Electron. Agric.* **2023**, *207*, 107757. [CrossRef]
23. Sekharamantry, 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]
24. Gai, R.; Chen, N.; Yuan, H. A detection algorithm for cherry fruits based on the improved YOLO-v4 model. *Neural Comput. Appl.* **2023**, *35*, 13895–13906. [CrossRef]
25. Tang, Y.; Zhou, H.; Wang, H.; Zhang, Y. Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. *Expert Syst. Appl.* **2023**, *211*, 118573. [CrossRef]
26. Huang, M.L.; Wu, Y.S. GCS-YOLOV4-Tiny: A lightweight group convolution network for multi-stage fruit detection. *Math. Biosci. Eng.* **2023**, *20*, 241–268. [CrossRef] [PubMed]
27. Zhong, S.; Xu, W.; Zhang, T.; Chen, H. Identification and depth localization of clustered pod pepper based on improved Faster R-CNN. *IEEE Access* **2022**, *10*, 93615–93625. [CrossRef]
28. Gao, F.; Fu, L.; Zhang, X.; Majeed, Y.; Li, R.; Karkee, M.; Zhang, Q. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. *Comput. Electron. Agric.* **2020**, *176*, 105634. [CrossRef]
29. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
30. Zhang, Z.G.; Zhang, Z.D.; Li, J.N.; Wang, H.; Li, Y.; Li, D. Potato detection in complex environment based on improved YOLOv4 model. *Trans. Chin. Soc. Agric. Eng. (Trans. CSAE)* **2021**, *37*, 170–178, (In Chinese with English Abstract).
31. Zhang, X.; Gao, Q.; Pan, D.; Cao, P.C.; Huang, D.H. Research on spatial positioning system of fruits to be picked in field based on binocular vision and SSD model. *J. Phys. Conf. Ser.* **2021**, *1748*, 042011. [CrossRef]
32. Liu, T.H.; Nie, X.N.; Wu, J.M.; Zhang, D.; Liu, W.; Cheng, Y.F.; Zheng, Y.; Qiu, J.; Qi, L. Pineapple (Ananas comosus) fruit detection and localization in natural environment based on binocular stereo vision and improved YOLOv3 model. *Precis. Agric.* **2023**, *24*, 139–160. [CrossRef]
33. Su, L.; Sun, H.; Zhang, S.; Lu, X.; Wang, R.; Wang, L.; Wang, N. Cucumber Picking Recognition in Near-Color Background Based on Improved YOLOv5. *Agronomy* **2023**, *13*, 2062. [CrossRef]
34. Yang, H.; Liu, Y.; Wang, S.; Qu, H.; Li, N.; Wu, J.; Yan, Y.; Zhang, H.; Wang, J.; Qiu, J. Improved Apple Fruit Target Recognition Method Based on YOLOv7 Model. *Agriculture* **2023**, *13*, 1278. [CrossRef]
35. Chen, C.; Wang, F.; Cai, Y.; Yi, S.; Zhang, B. An Improved YOLOv5s-Based Agaricus bisporus Detection Algorithm. *Agronomy* **2023**, *13*, 1871. [CrossRef]