agriculture

MDPI

Article

MYOLO: A Lightweight Fresh Shiitake Mushroom Detection Model Based on YOLOv3

Peichao Cong *, Hao Feng, Kunfeng Lv, Jiachao Zhou and Shanda Li

School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China
* Correspondence: cpclzx2022@gxust.edu.cn

Abstract: Fruit and vegetable inspection aids robotic harvesting in modern agricultural production. For rapid and accurate detection of fresh shiitake mushrooms, picking robots must overcome the complex conditions of the growing environment, diverse morphology, dense shading, and changing field of view. The current work focuses on improving inspection accuracy at the expense of timeliness. This paper proposes a lightweight shiitake mushroom detection model called Mushroom You Only Look Once (MYOLO) based on You Only Look Once (YOLO) v3. To reduce the complexity of the network structure and computation and improve real-time detection, a lightweight GhostNet16 was built instead of DarkNet53 as the backbone network. Spatial pyramid pooling was introduced at the end of the backbone network to achieve multiscale local feature fusion and improve the detection accuracy. Furthermore, a neck network called shuffle adaptive spatial feature pyramid network (ASA-FPN) was designed to improve fresh shiitake mushroom detection, including that of densely shaded mushrooms, as well as the localization accuracy. Finally, the Complete Intersection over Union (CIoU) loss function was used to optimize the model and improve its convergence efficiency. MYOLO achieved a mean average precision (*mAP*) of 97.03%, 29.8M parameters, and a detection speed of 19.78 ms, showing excellent timeliness and detectability with a 2.04% higher *mAP* and 2.08 times fewer parameters than the original model. Thus, it provides an important theoretical basis for automatic picking of fresh shiitake mushrooms.

Keywords: picking robot; fresh mushroom sorting; YOLOv3; detection model; lightweight

Citation: Cong, P.; Feng, H.; Lv, K.; Zhou, J.; Li, S. MYOLO: A Lightweight Fresh Shiitake Mushroom Detection Model Based on YOLOv3. *Agriculture* **2023**, *13*, 392. https://doi.org/10.3390/agriculture13020392

Academic Editors: Wen-Hao Su and Zhou Zhang

Received: 13 November 2022
Revised: 3 February 2023
Accepted: 4 February 2023
Published: 7 February 2023

1. Introduction

Mushrooms are an important foodstuff for humans; their cultivation is a significant part of the agricultural development of many countries and is growing rapidly [1]. In Europe, the average mushroom consumption per person is about 3.5 kg per year. The European mushroom market is expected to expand at the highest compound annual growth rate (CAGR) of 8.07% during the period 2017–2023. According to the analysis report of Zion Market Research, the total capital of the global mushroom market exceeded $59.48 billion in 2021, with a CAGR of more than 9.2% from 2016 to 2020. Global mushroom consumption is expected to reach 20.84 million tons in 2026, with a CAGR of 6.41%. Therefore, if manual sorting of mushrooms is still used, a lot of labour will be wasted [2,3]. Fresh shiitake mushroom picking is a key aspect of mushroom production, which is still predominantly manual, labour intensive, inefficient, and costly. The use of picking robots to replace manual labour can improve production efficiency and reduce costs. The prerequisite for automated robotic picking is the rapid and accurate detection of fresh shiitake mushrooms. Visual detection is one of the optimal means of achieving these goals, and related issues have become popular research topics in recent years [4]. Therefore, it is of great research value and relevance to investigate a visual detection algorithm that can accurately detect fresh shiitake mushrooms in mushroom sheds.

For fruit and vegetable testing, numerous studies based on traditional machine vision theory have emerged. Arefi et al. [5] extracted foreground information in Red-Green-Blue

(RGB) space and combined the Intensity-Hue-Saturation (IHS) and Luminance In-phase Quadrature (YIQ) spaces to obtain combined features to detect ripe tomatoes, but the detection performance was poor for small targets. Wei et al. [6] performed feature extraction of images in Ohta space and fed the extraction results into the Otsu thresholding algorithm for automatic detection to achieve the automatic recognition of fruits by a picking robot in a natural environment. Lu and Sang [7] used the segmentation results of citrus color difference maps and normalized RGB maps under different illuminations to build contour segments and derive partial order relationships for citrus detection, but the detection speed was slow. Xiong et al. [8] combined Fuzzy C-Means (FCM) with a one-dimensional random signal histogram to reject the nighttime image background and used the Otsu algorithm to segment the litchi, thereby determining the picking point and providing technical support for the vision system of the picking robot. The above-mentioned traditional identification methods have solved the detection problems of some fruits and vegetables to a certain extent. Still, there is limited research on detecting fresh shiitake mushrooms with diverse morphology, dense growth, shading, and variable field of view. In addition, the traditional methods rely excessively on manual extraction of features and scene information and have poor real-time performance, robustness, and generalization capabilities; thus, they cannot satisfy the actual working needs of picking robots.

In recent years, with the rapid development of deep learning, its application in image recognition has been increasing. Compared with manual feature extraction methods, neural networks can extract multilevel features of images though unsupervised or weakly supervised learning, which has stronger generalization ability and enables significant improvement in target detection performance [9]. Currently, deep learning-based image detection algorithms fall into two main categories. The methods in one class are based on region suggestion and include faster regions with convolutional neural network features (Faster R-CNN) [10] and region-based fully convolutional networks (R-FCNs) [11]. The core idea is to first obtain a suggested region and then perform accurate classification and location prediction within that region. Lamb and Chuah [12] proposed a low-cost strawberry detection system based on convolutional neural networks, but the detection speed was only 1.63 frames. Yu et al. [13] combined ResNet-50 [14] with a feature pyramid network (FPN) [15] as the backbone network of mask regions with a convolutional neural network (Mask R-CNN) [16] for the real-time detection of strawberries in unstructured environments. Lin et al. [17] combined Red-Green-Blue-Depth (RGB-D) sensors with a modified FCN for guava detection and localization. Mu et al. [18] combined a Faster R-CNN with ResNet-101 and used migration learning to detect unripe tomato fruits. In their novel identification approach for kiwis, Liu et al. [19] fused aligned RGB and near-infrared (NIR) pictures with a Faster R-CNN. These methods can accurately classify and predict the positions of fruits and vegetables, but the detection speed is slow and does not satisfy the real-time requirements of picking robot operations. The methods in the other class are region-free suggested methods such as You Only Look Once (YOLO) [20], Single-Shot MultiBox Detector (SSD) [21], and CenterNet [22]. The methods in this class transform the localization and classification of the detection process into a regression problem, which in turn improves the detection speed. Koirala et al. [23] proposed the MangoYOLO model for the real-time detection of mangoes and estimation of their yield, but its recognition scenario was rather homogeneous. Li et al. [24] improved YOLOv4-tiny [25] and combined it with migration learning to perform training in stages to detect ripe grapes. Lu and Sang [26] added a convolutional attention module called the convolutional block attention module (CBAM) [27] to the YOLOv4 feature fusion network and included adaptive layers and large-scale feature maps to detect the ripeness of apples. Wang et al. [28] proposed the DSE-YOLO model for the detection of strawberries at different growth stages, but this method increased the number of model parameters. Although the methods in the region-free proposal class have faster detection speeds than those in the region suggestion class, their network structures are complex, contain numerous parameters, require higher equipment performance, and consume more computational resources.

In summary, despite the emergence of numerous deep learning-based fruit and vegetable detection methods, most of them are only applicable to a single scene or target, cannot guarantee detection speed and accuracy simultaneously, and require high-performance equipment. Due to the complex environment in mushroom sheds, fresh shiitake mushrooms are of various forms, exhibit dense growth, are easy to shade, and have variable fields of view, all of which can seriously affect the detection accuracy and effectiveness of the picking robot. Based on the above-mentioned analysis, improving the detection speed while ensuring adequate detection accuracy was the focus of this study and is a research hotspot in the field of fruit- and vegetable-picking robots [29–31].

In this study, we took the image detection problem of fresh shiitake mushrooms in a mushroom shed as the research object, transformed the localization and classification loss of fresh shiitake mushrooms in the detection process into a regression problem, and proposed a lightweight fresh shiitake mushroom detection model called MYOLO to promote the development and application of fresh shiitake mushroom-picking robots. The main contributions of this study are as follows.

(1) To improve the detection speed with guaranteed detection accuracy, the YOLOv3 backbone network Darknet53 was replaced with the lightweight GhostNet16 to compress the model. Spatial pyramid pooling (SPP) was introduced at the end of the backbone network to enrich the expression capability of the final feature map and improve the detection and classification accuracy of small fresh shiitake mushrooms.

(2) A feature fusion network called ASA-FPN was designed that consisted of a FPN, shuffle attention network (SANet), and adaptive spatial feature fusion (ASFF) to improve the detection and localization accuracy of the model for fresh shiitake mushrooms and to enhance its ability to detect densely occluded fresh shiitake mushrooms.

(3) CIoU was used as the regression loss function of the bounding box to improve the problem of slow regression during model training. In addition, migration learning was utilized in the training process to improve the accuracy and generalization ability of the network.

The remainder of this paper is organized as follows. Section 2 describes the image acquisition, annotation, and dataset partitioning methods as well as the improvement of the lightweight YOLOv3-based structural model. Section 3 introduces the experimental design. Section 4 presents the experimental results and analysis. Finally, Section 5 summarizes the conclusions and topics for future work.

2. Materials and Methods

2.1. Image Acquisition

A fresh shiitake mushroom dataset was collected in a mushroom plantation from April to June 2022, and 1803 fresh shiitake images were acquired using Shengyue industrial cameras (Camera model AHD10802P-USB, manufacturer is Weixin Vision, country of origin is China) and mobile phones with a camera resolution of 720 × 480. The mobile phones were not fixed according to the resolution of the shooting angle. The fresh shiitake mushroom images were collected mainly during the daytime, and the filming simulated the picking process of picking robots by constantly adjusting the filming angle and distance. The dataset was divided into two main parts: images of fresh shiitake mushrooms on mushroom stakes and images of picked mushrooms, which included different types, sizes, and distribution densities of fresh shiitake mushrooms. With reference to national regulations and common market classification methods in China, the mushrooms were classified into three categories according to the cracks and shapes of their heads: cracked-surface mushrooms, plane-surface mushrooms, and malformed-surface mushrooms (including malformed cracks and planes) [32]. According to field surveys, the market value of cracked-surface mushrooms is much higher than that of plane-surface mushrooms, and the value of the malformed mushrooms is the lowest [33]. Thus, fresh mushrooms need to be sorted after harvesting. To improve production efficiency, simultaneous picking, detection, and sorting of fresh mushrooms was considered in this study. Due to the low production of

malformed mushrooms, some fresh malformed mushroom images were collected via the Internet, considering the versatility of the visual system. Finally, 1416 images of fresh mushrooms were acquired through field photography and Internet collection, including 274 cracked-surface mushrooms, 351 plane-surface mushrooms, 278 malformed fresh mushrooms, and 513 mixed fresh mushrooms (multiple species, dense shading, multiple fields of view, etc.). The species classifications and shapes of fresh shiitake mushrooms are shown in Figure 1.

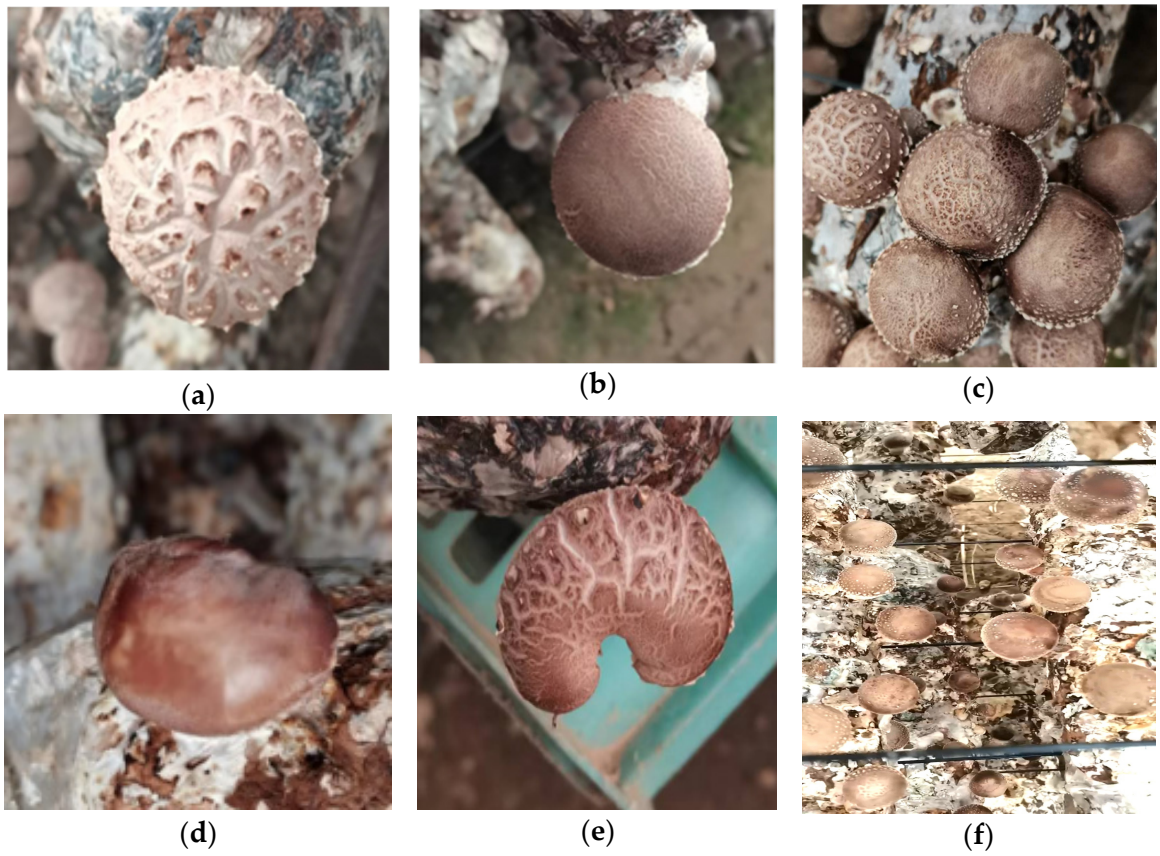


Figure 1. Varietal classification and shapes of various fresh shiitake mushrooms. (**a**) Cracked-surface mushroom; (**b**) Plane-surface mushroom; (**c**) Dense overlap; (**d**) Malformed plane-surface mushroom; (**e**) Malformed cracked-surface mushroom; (**f**) Large field of view.

2.2. Image Datasets

The datasets used in this study were in the PASCAL.VOC2007 format. The regions of fresh shiitake mushrooms in the images were manually labelled with rectangular boxes using the LabelImg software to obtain an Extensible Markup Language (XML) file in VOC format; examples of various types of fresh shiitake mushroom labelling are shown in Figure 2. The labelling was conducted with fresh shiitake mushrooms in a manually observable, full-labelling manner, with all shielded shiitake mushrooms in the image labelled according to their visible size and identified by human eye observation. After labelling, the training, validation, and test sets were allocated according to a 6:1:3 ratio, where 638 images were randomly selected as the training set, 141 as the validation set, and 425 as the test set.

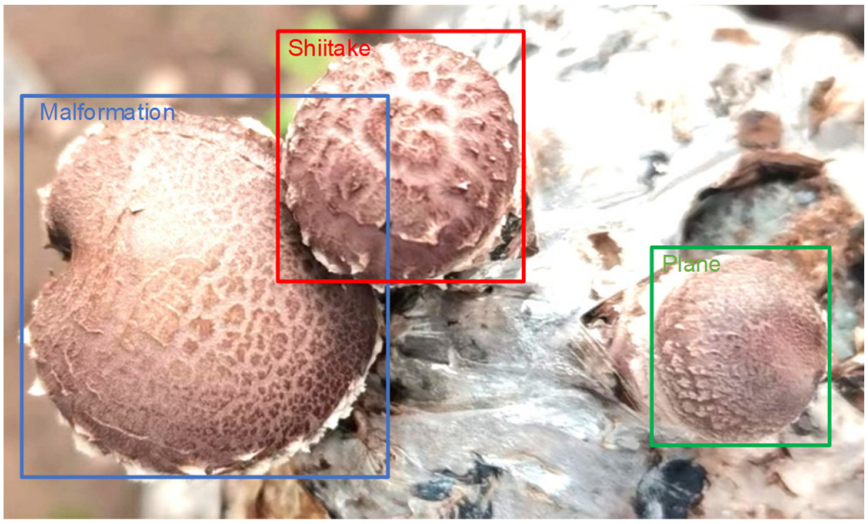


Figure 2. Example labels for various types of fresh shiitake mushrooms.

In this study, the dataset was expanded using data augmentation techniques to increase the variety of the experimental data, boost the model generalizability, and prevent overfitting. The fresh shiitake mushroom images were extended by using the following five techniques: rotation, Gaussian noise, contrast enhancement, brightness variation, and mosaic data enhancement [34,35], as shown in Figure 3. The first four enhancement techniques were implemented through the built-in Python-based OpenCV functions, whereas the mosaic method involves random cropping of four images stitched together into one new image as the training data. After these operations, the training set was expanded to 3184 images, the validation set to 563 images, and the test set to 662 images. The test set included a selection of data randomly adjusted for changes in brightness to simulate light sources.

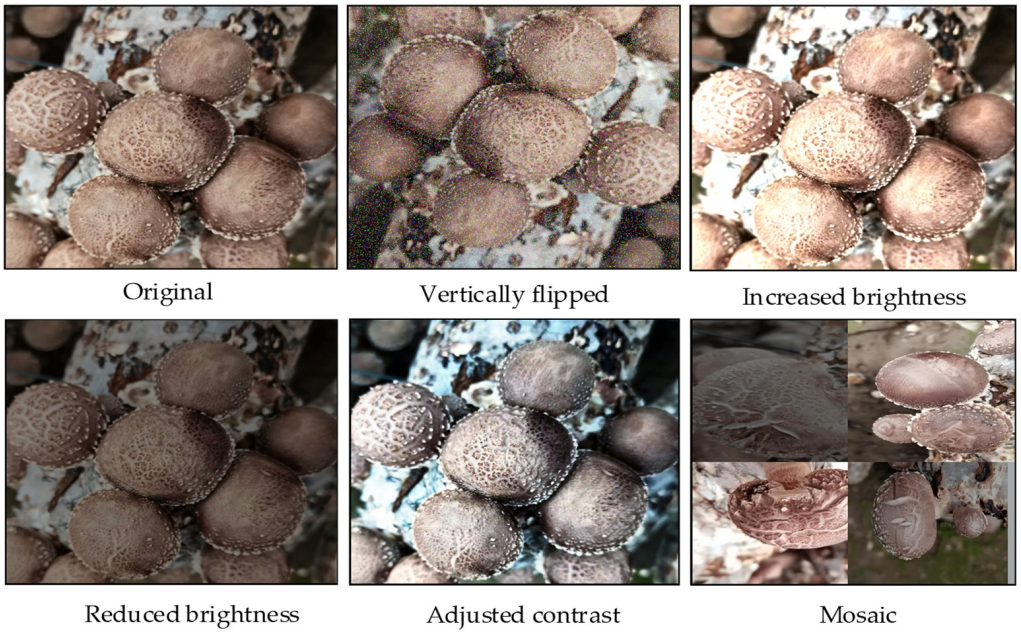


Figure 3. Various types of data enhancement effects.

2.3. Problems with the YOLOv3 Model

YOLO [20] is a single-stage object detection method based on a regression model proposed by Redmon, which uses the Overfeat algorithm proposed by Sermanet in 2013 to

make predictions based on global image information. YOLOv3 [36] is an improved version of the YOLO network structure in which the complete connection layer and last pooling layer are removed from YOLO and a fully convolutional network structure is adopted, enabling the network to extract higher-resolution features. Its network consists of the backbone DarkNet53, neck FPN and three-branch prediction structure. DarkNet53 is a feature extraction network involving ordinary convolution, which improves the learning ability of the network by adding a residual structure between the convolution layer and the lower sampling layers to reduce the loss caused by gradient disappearance. Due to the use of ordinary convolution, the number of network parameters in DarkNet53 increases dramatically with the number of convolutional layers, which can generate many nonessential calculations and directly affect the detection speed of the model. In addition, the model uses an FPN as the neck network, predicts three feature layers at different scales using three branches, and fuses differently sized feature layers obtained by downsampling. This approach enhances the reuse of information from different feature layers, but there is variability between different feature scales, and when there are both small and large targets in the image, the contradiction between features at different layers will lead to poor detection and localization accuracy, and the ability to detect variable fields of view and densely occluded groups is also relatively weak.

2.4. Model Improvements

2.4.1. MYOLO Network Structure

To address the above-mentioned problems of the YOLOv3 model; overcome the effects of complex conditions such as diverse morphology, dense growth, easy occlusion, and variable field of view of fresh shiitake mushrooms; and improve the detection and localization accuracy of fresh shiitake mushrooms further, this paper proposes a new lightweight fresh shiitake mushroom detection model called MYOLO. This model is based on YOLOv3, and the "M" denotes that fresh shiitake mushrooms are the detection targets. MYOLO mainly consists of a backbone network, a neck network and a prediction network, the backbone network is responsible for extracting picture information, the neck network performs further feature extraction (such as location, category) on the obtained picture information, the prediction network is located behind the backbone network and neck network (as shown in Figure 4a), predicts the target and scores. MYOLO utilizes the regression idea of the YOLOv3 model, with few ghost modules (the principle will be described in Section 2.4.2) as the main body to build a lightweight GhostNet16, as well as SPP modules. The lightweight backbone network of MYOLO is formed. The network compresses the model and reduces the operational computation required for general convolution while maintaining accuracy to improve the real-time detection performance. Through the introduction of the SPP module, local and global features can also be effectively fused to enhance the detection and classification accuracy of the model further for small, fresh shiitake mushrooms. Furthermore, to utilise the features extracted by the backbone network further to improve the detection and localisation accuracy of MYOLO for fresh mushrooms and to enhance the detection of densely occluded fresh mushrooms, a new neck network called ASA-FPN, consisting of FPN, SANet, and ASFF, was designed. SANet is located after Concat, the key position of FPN, at the intersection between different scales of information in a feature fusion network, which is conducive to obtaining rich feature information and improving the localization accuracy of the model. ASFF is located at the end of the FPN and can effectively suppress the inter-scale variability and improve the accuracy of detecting densely occluded fresh shiitake mushrooms by MYOLO. The MYOLO model ASFF network outputs three prediction layers afterwards for use in the prediction network, and its network architecture is shown in Figure 4.

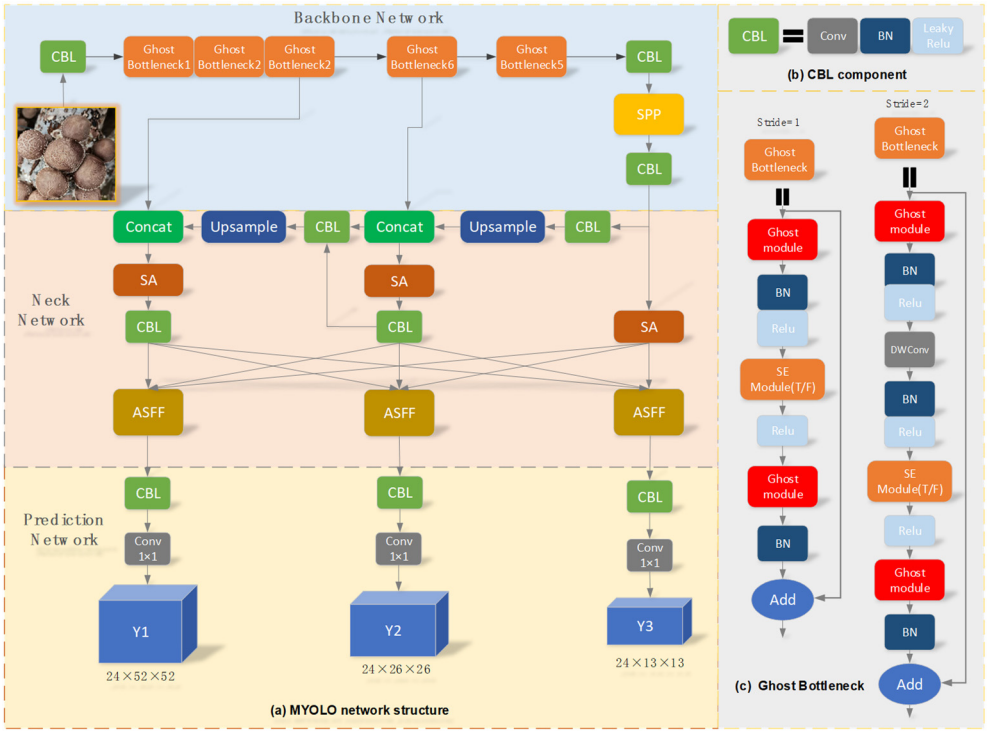


Figure 4. MYOLO network structure.

After prediction layers Y1, Y2, and Y3 are processed in the prediction network, the images are divided into 52 × 52, 26 × 26, and 13 × 13 grids, respectively, and three anchor frames are generated for each grid. During the training process, the anchor frame providing the target information is chosen as the prediction result after the sizes and locations of the anchor frames are continually changed. Prediction layer Y1 has a small grid, which is effective for finding tiny targets in the picture; prediction layer Y2 has a moderate grid, which is good for detecting intermediate targets; and prediction layer Y3 has a large grid, which is good for detecting large targets. The final prediction parameters for the picture are included in each channel of the prediction layer. The specific structure of the prediction layer (with a 13 × 13 grid as an example) is shown in Figure 5. The prediction parameters of each prediction layer include the prediction frame centre coordinates (X, Y), prediction frame length and width (W and H, respectively), prediction frame confidence level (C), score of fresh shiitake mushrooms in the prediction frame (Score), and the number of predicted bounding boxes for which each grid is responsible (B).

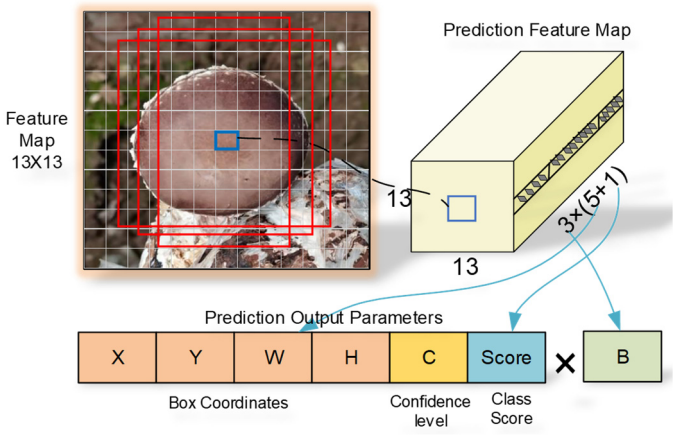


Figure 5. Structure of the 13 × 13 prediction layer.

2.4.2. GhostNet16 Network Structure

GhostNet [37] is a novel end-to-end network architecture proposed by Huawei Noah's Ark (Noah's Ark Lab is a laboratory based in China where Huawei conducts basic research on artificial intelligence. http://dev3.noahlab.com.hk/index.html, accessed on 6 February 2023), which is based on a series of low-computation linear transformations to provide additional low-redundancy features. GhostNet is mainly stacked by Ghost Bottleneck, which consists of a Ghost module, squeeze-and-excitation (SE) module [38], and depthwise (DW) separable convolution [39], as shown in Figure 4c. When the step size is 1, the Ghost module in front of the Ghost Bottleneck acts as an extension layer to increase the number of channels, and the Ghost module behind it acts as a compression layer to reduce the number of input channels to match the Shortcut path. The SE module is also used in the Ghost Bottleneck to adjust the weights of each channel adaptively to strengthen the critical channels and suppress the minor ones, thereby improving the network performance. When the step size is 2, a downsampling layer and DW separable convolution with a step size of 2 are used to build the Shortcut path to achieve a lightweight model. The above-mentioned mechanism reduces the number of parameters in the model and improves timeliness [40].

Figure 6 illustrates the process of extracting and generating feature maps by using the Ghost module. Here, $\varnothing_z$ denotes a low computational linear transformation process, z denotes the zth linear operation to generate the feature map, and i denotes the ith feature extraction performed by the backbone network. Suppose that the input feature map size is $c^i_{input} * w^i_{input} * h^i_{input}$, which is divided into s_i parts, the output feature map size is $c^i_{output} * w^i_{output} * h^i_{output}$, the convolution kernel is $k_i * k_i$, and the size of each linear operation kernel is $d_i * d_i$. The ordinary convolutional computation used by DarkNet53 is T^i_c, as shown in Equation (1); the Ghost module used by GhostNet is T^i_g, as expressed in Equation (2). Assuming that k_i is equal to d_i and that s_i is much smaller than c^i_{output} (which is exactly what happens in practice), the compression ratio r_i calculated using ordinary convolution and the Ghost modules can be derived according to Equations (1) and (2), as shown in Equation (3). Based on Equation (4), when the backbone network needs to perform feature extraction n times, the computation required by the Ghost module will be exponentially reduced compared with ordinary convolution (R_n), and the timeliness will be significantly improved.

$$T^i_c = c^i_{input} * k^2_i * c^i_{output} * h^i_{output} * w^i_{output} \tag{1}$$

$$T^i_g = \frac{c^i_{output}}{s_i} * h^i_{output} * w^i_{output} * (c^i_{input} * k^2_i + (s_i - 1) * d^2_i) \tag{2}$$

$$r_i = \frac{T^i_c}{T^i_g} \approx \frac{s_i * c^i_{input}}{s_i + c^i_{input} - 1} \approx s_i > 1 \tag{3}$$

$$R_n = r_1 * r_2 * r_2 \ldots r_{n-1} * r_n \gg \min(s_i)^n \tag{4}$$

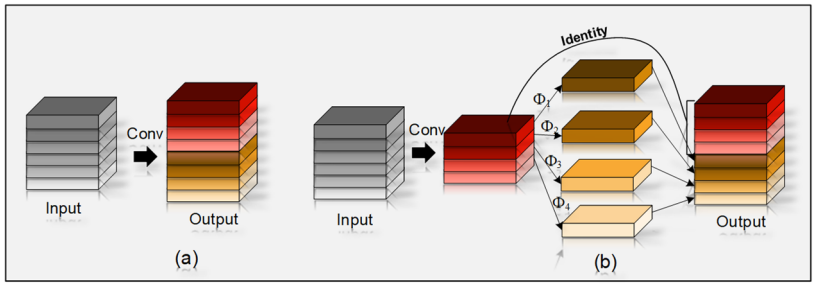


Figure 6. Process of extracting and generating feature maps by using the Ghost module. (**a**) Ordinary convolution; (**b**) Ghost module.

Based on the above-mentioned analysis, to solve the problem of the large number of DarkNet53 parameters in the backbone network of YOLOv3, we built a lightweight GhostNet16 as the backbone network of MYOLO by drawing on the GhostNet network structure. The GhostNet16 network parameters are listed in Table 1, and the network structure is shown in Figure 4a. Because the neck network of MYOLO requires three different scales of feature inputs to provide multiscale information for prediction and classification, feature layers 5, 11, and 16, which were adjusted by the SE module, were used as input features for the MYOLO neck network.

Table 1. Hardware and software configuration.

Feature Layer	Input	Component Units	Stride	Attention
0	416 × 416 × 3	Conv2d 3 × 3	2	Add
1	208 × 208 × 16	G-bneck 3 × 3	1	No
2	208 × 208 × 16	G-bneck 3 × 3	2	No
3	104 × 104 × 24	G-bneck 3 × 3	1	No
4	104 × 104 × 24	G-bneck 5 × 5	2	Add
5	52 × 52 × 40	G-bneck 5 × 5	1	Add
6	52 × 52 × 40	G-bneck 3 × 3	2	No
7	26 × 26 × 80	G-bneck 3 × 3	1	No
8	26 × 26 × 80	G-bneck 3 × 3	1	No
9	26 × 26 × 80	G-bneck 3 × 3	1	No
10	26 × 26 × 80	G-bneck 3 × 3	1	Add
11	26 × 26 × 112	G-bneck 3 × 3	1	Add
12	26 × 26 × 112	G-bneck 5 × 5	2	Add
13	13 × 13 × 160	G-bneck 5 × 5	1	No
14	13 × 13 × 160	G-bneck 5 × 5	1	Add
15	13 × 13 × 160	G-bneck 5 × 5	1	No
16	13 × 13 × 160	G-bneck 5 × 5	1	Add

2.4.3. SPP Network Structure

To fuse local and global features effectively and improve the model detection performance for small fresh shiitake mushrooms, the spatial pyramidal pooling (SPP) structure was incorporated into the backbone network of MYOLO [41–43]. SPP is a structure consisting of three different scales of maximum pooling layers, as shown in Figure 7. These are H: feature map height, W: width, and C: number of channels. First, the input H × W × C feature map is executed three times with different convolution kernel sizes in block pooling to extract feature information from different sizes of perceptual fields; second, the feature map is obtained by pooling operation to normalize its size. Next, the three feature maps are merged with the original feature map on the channels to obtain the H × W × 4C feature map. Finally, the spliced feature map is passed on the subsequent network to increase the perceptual field, enrich the expression capability of the final feature map, and enhance the detection performance of the model for small fresh shiitake mushrooms further.

2.4.4. ASA-FPN Network Structure

The new neck network, ASA-FPN, designed in this study consists of three main components: FPN, SANet, and ASFF. The main structure of FPN was shown in Figure 4 and will not be repeated here, and the SANet and ASFF networks introduced are further analysed, as described below.

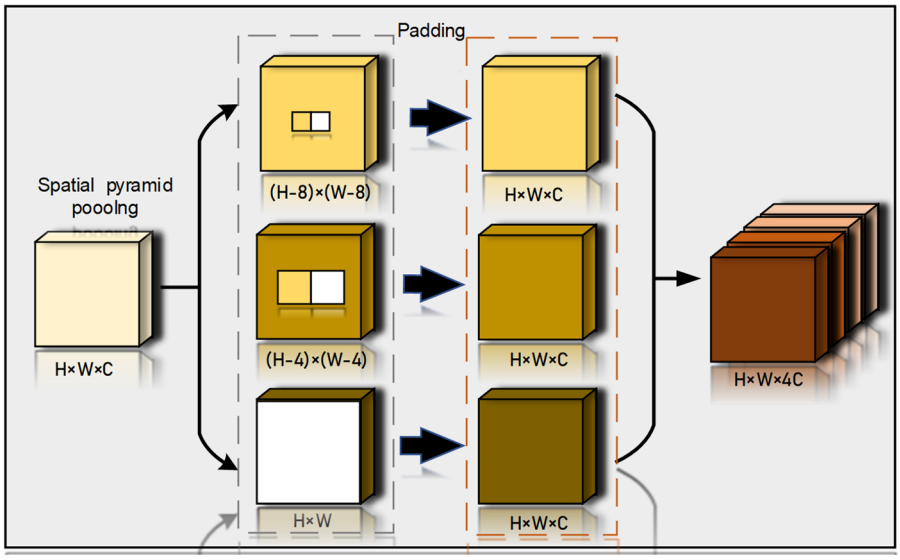


Figure 7. SPP network structure diagram.

To improve the detection and localization accuracy of fresh shiitake mushrooms in complex environments, an attention mechanism based on feature grouping and channel replacement (SANet) [44] was introduced at the intersection of information at different FPN scales. The network structure of SANet is shown in Figure 8. First, input X is split into groups according to the channel dimensions; second, for split features X_k, they are further split into two branches along the channel dimension for learning channel attention features X_{k1} and null domain attention features X_{k2}. X_{k1} is extracted using a combination of GAP, Scale, and Sigmoid. X_{k2} is first extracted using the group norm (GN) for spatial-level feature extraction, followed by enhancement using $F_c(\cdot)$. After the two attention calculations, X_k is obtained by fusing the two types of attention features through Concat. Next, the channel shuffle operation is used for inter-group communication. Finally, the feature map output has the same size as the input. The above-mentioned attention mechanism gives higher weight to the fresh shiitake mushroom feature information to suppress the influence of the background information, improving the accuracy of fresh shiitake mushroom detection and localization in complex scenes.

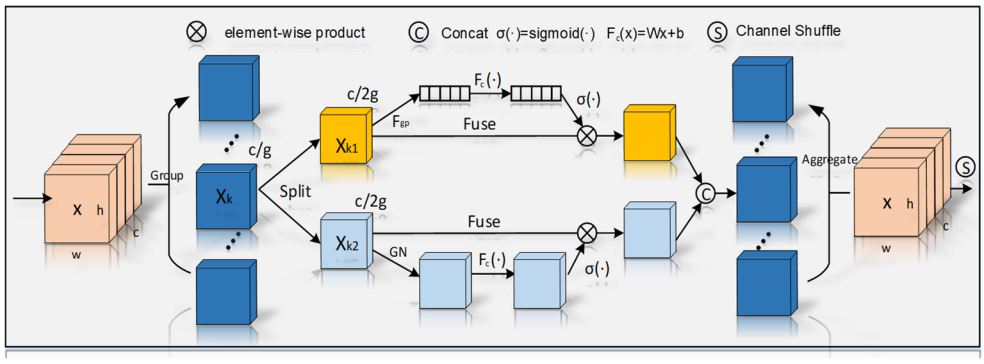


Figure 8. SA network architecture diagram.

In an FPN, multiscale prediction still faces conflicts, although simple feature fusion can improve feature layer information. An ASFF network [45] was incorporated at the end of the ASA-FPN. The new fusion network can adaptively learn the weight information at different scale feature levels to reduce conflicts when there are densely distributed objects. In Figure 9, a 13 × 13 × 256 feature map is used as an example. First, to solve the different-scale problem, 52 × 52 × 64 and 26 × 26 × 128 feature maps are unified in dimension and

downsampled to a 13 × 13 × 256 feature map (if the upsampling of the unified dimension transformation is also performed from the deep feature map to the shallow feature map). Subsequently, the feature maps of each layer are compressed by the 1 × 1 convolution block to generate three 13 × 13 × 16 feature maps and then by the Softmax function to extract the 13 × 13 × 3 multi-scale feature-level weight information. Then, it is multiplied by the downsampled feature layers P1, P2, and P3 (each feature map is multiplied times only 13 × 13 × 1 feature maps) and the outputs are summed. Finally, the effective feature maps responsible for target prediction are obtained. Through the above-mentioned operation, the inter-scale variability can be effectively suppressed to improve the accuracy of the detection network for densely shaded fresh shiitake mushrooms, reducing the missed detection rate.

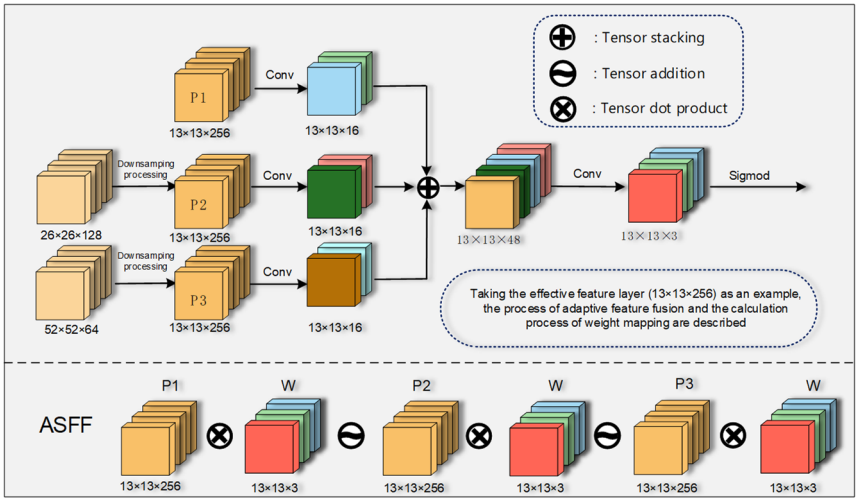


Figure 9. ASFF network structure.

3. Experimental Design

3.1. Network Training

Table 2 displays the hardware and software combinations utilized in this work for model training and testing. We used migration learning for establishing the model parameters to increase the recognition accuracy of the network [13]. Initialization provides the network with fast learning capabilities, eliminates network overfitting, and improves the generalization of the network for fresh shiitake mushroom detection in a mushroom shed environment.

Table 2. Hardware and software configuration.

Hardware or Software	Configuration
CPU	Intel i9-10700H
RAM	24 GB
SSD	256 GB
Operating system	Window 10
GPU	NVIDIA GeForce GTX 2080Ti 11 GB
Development environment	Python 3.8, Pytorch 1.12, CUDA 11.3

The MYOLO network was first used to train the PASCAL.VOC2007 dataset for 200 stages. After the training was completed, the higher mean average precision (*mAP*) weight file was then selected as the pretraining model for the MYOLO network. Before training, the sizes of the nine anchor boxes were calculated using the K-means clustering algorithm [46], where (42, 22), (67, 44), and (71, 89) correspond to prediction layer Y1; (108, 63), (137, 101), and (99, 148) correspond to prediction layer Y2; and (210, 132), (157, 184), and (254, 253) correspond to prediction layer Y3, as shown in Figure 10. The above operation

makes the actual size of the anchored box closer to the size of fresh shiitake mushrooms in the dataset, which is conducive to improving the detection and localization accuracy. In the training process, the epoch was set to 400 and the batch size was set to 16. The stochastic gradient descent [47] optimizer was used for training, with the initial learning rate set to 0.001, momentum set to 0.937, and weight decay set to 0.0001, and the cosine annealing method was employed to update the learning rate. The model was saved once at the end of each epoch, and the performance metrics of the detection model were recorded in real time through the matplotlib tool, with the training taking a total of 4 h 45 min.

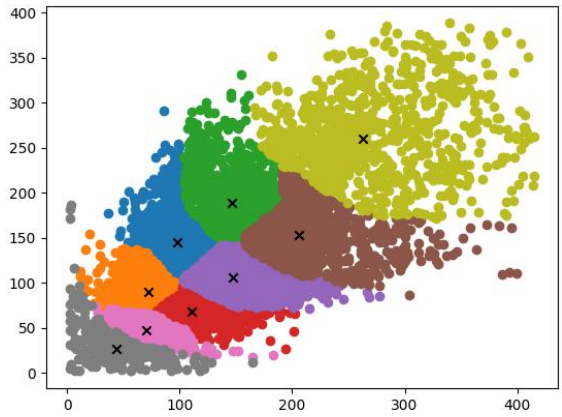


Figure 10. Generation of nine candidate boxes by K-means (K = 9).

3.2. Loss Function

During the model training process, there is an uncertainty error between the predicted and true values. The objective of the loss function is to reduce this error continuously so that the value predicted by the model is as close to the corresponding true value as possible. The loss function of MYOLO consists of three main components: bounding box loss, confidence loss, and category loss. Among them, the bounding box loss function is regressed by the CIoU [48] function to improve the convergence efficiency, which is calculated as follows:

$$Loss = L_{CIoU} + L_{conf} + L_{class} \tag{5}$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(p,g)}{c^2} + \beta v \tag{6}$$

$$IoU = \frac{A \cap B}{A \cup B} \tag{7}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^g}{h^g} - arctan\frac{w^p}{h^p}\right)^2 \tag{8}$$

$$\beta = \frac{v}{(1 - IoU) + v} \tag{9}$$

where A and B are the areas of the two boxes; IoU denotes the degree of overlap between the two boxes; p and g are the centroids of the predicted and actual boxes, respectively; c is the diagonal length of the smallest external rectangular box in the box; β is the weight; v is the parameter measuring the consistency of the length, width, and ratio; and w and h are the width and height of the box, respectively.

$$L_{conf} = \sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{i,j}^{obj}\log(p_i) - \lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{i,j}^{noobj}\log(1 - p_i) \tag{10}$$

$$L_{class} = -\lambda_{class}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{i,j}^{obj}\sum_{n_i \in classes} BCE(n, n^*) \tag{11}$$

$$BCE(n, n^*) = -n\log n^* - (1 - n)\log(1 - n^*) \tag{12}$$

where S is the number of grids; B is the number of prior frames in each network; n and n^* are the values of the actual and predicted categories of the jth a priori box of the ith grid, respectively; $BCE(n, n^*)$ is the cross entropy loss; $I_{i,j}^{obj}$ is 1 for the jth a priori box of the ith grid with a target and 0 for no target; $I_{i,j}^{noobj}$ is the jth a priori box of the ith grid, which is 1 when there is no target and 0 when there is a target; p is the probability that the target exists in the current prior frame; λ_{noobj} is the loss of confidence weights without objects; and λ_{class} denotes category loss weights.

3.3. Model Evaluation

In this study, to verify the accuracy of the MYOLO detection model, fresh shiitake mushroom recall (R), precision (P), $F1$ score, average precision (AP), and mAP were used as evaluation indicators [49]. The specific formulae for the above-mentioned indicators are as follows:

$$R = \frac{TP}{TP + FN} \tag{13}$$

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$AP = \int_0^1 P(R)dR \tag{16}$$

where TP is the number of fresh shiitake mushrooms detected correctly, FP is the number of fresh shiitake mushrooms detected incorrectly, and FN is the number of fresh shiitake mushrooms missed. For each category in the target detection, a $P - R$ curve can be plotted based on the accuracy and recall.

4. Experimental Results and Analysis

4.1. Experimental Results

The loss curves during training and validation set mAP values are shown in Figure 11.

As shown in Figure 11a and Table 3, MYOLO uses CIoU as the bounding box loss function, and the training time is shortened by 29 min compared to MYOLO-R, which improves the regression speed and convergence effect during model training. At the same time, the mAP of MYOLO is increased by 7.3% compared to MYOLO-N in Table 3, indicating that the introduction of transfer learning also greatly improves the accuracy of model detection [50–52]. In addition, the loss curves of MYOLO(Train) and MYOLO(Val) demonstrate that the rate of decline is the fastest in the first 150 rounds of training and then gradually becomes slower. Although the experiment was set to train for 400 rounds, the validation set loss had stabilized after 230 rounds, and after 300 rounds, the validation set loss started to increase slowly while the training set loss was still decreasing, indicating that the model had been overfitted [53]. Figure 11b depicts the change curve of the validation set mAP (IoU threshold set to 0.5) during the training process. As the number of training rounds increases, the mAP curve also increases gradually, reaching a peak near round 220, and the curve has a decreasing trend after round 270. Therefore, in this study, the maximum value of mAP for each of the 20 rounds before and after the 220th round was taken as the final model weight. After testing, the maximum value appeared in the 222nd round, when mAP = 96.66%.

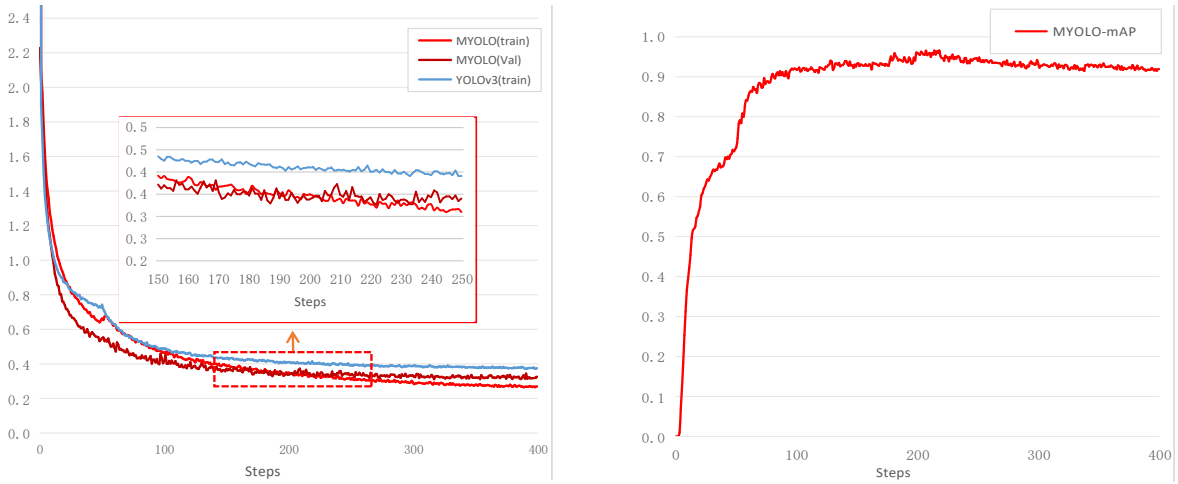


Figure 11. Loss curves and validation set *mAP* values during training. (**a**) Loss curve during training; (**b**) Validation set *mAP* curve during training.

The final model was tested using the test set, which contained images of three types of fresh shiitake mushrooms in a mushroom shed in various complex situations, as shown in Figure 2. The P-R curves of the three types of fresh shiitake mushroom images (Figure 12) indicate that the model achieved AP values of 96.15% for cracked-surface mushrooms, 95.60% for plane-surface mushrooms, and 98.39% for malformed-surface mushrooms, in addition to an *mAP* of 97.03% and an average detection speed of 19.78 ms. The above-mentioned test results prove that the detection accuracy and speed of the proposed model satisfy the practical requirements of a picking robot and that it can be applied to the automated detection of fresh shiitake mushrooms in mushroom sheds.

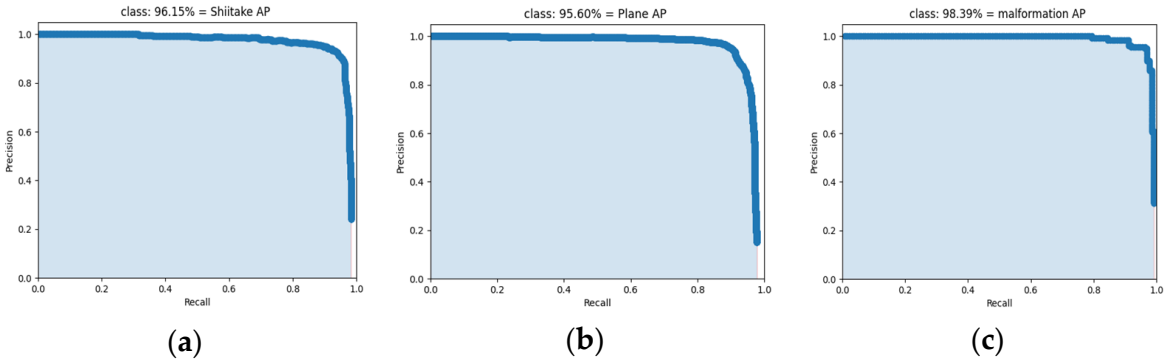


Figure 12. $P-R$ curves of the YOLO-M detection model. (**a**) Cracked-surface mushroom; (**b**) Plane-surface mushroom; (**c**) Malformed-surface mushroom.

Table 3. Training time and migration learning experiment results (IoU threshold set to 0.5).

Model	CIOU	Migration Learning	mAP	Training Time (Epoch = 400)
YOLOv3	×	√	94.85%	5 h 52 min
MYOLO-R	×	√	96.31%	5 h 14 min
MYOLO	√	√	97.03%	4 h 45 min
MYOLO-N	√	×	89.73%	4 h 47 min

4.2. Ablation Experiments

Ablation experiments are usually conducted on complex neural networks to explore the effects of network-specific substructures or training strategies and parameters on model generation and are important guides in the design of neural network structures [54]. To evaluate the effectiveness and feasibility of the lightweight model, MYOLO, proposed in this paper, the performance of GhostNet16, SPP, and ASA-FPN networks was verified individually by performing ablation experiments. As the MYOLO model is derived from YOLOv3, it was used as the benchmark for the ablation experiments.

Table 4 shows the results of the ablation experiments performed on GhostNet16, SPP, and ASA-FPN networks using the test set. YOLO-A is the model with GhostNet16 applied; YOLO-B is the model with GhostNet16 and SPP applied; and YOLO-C is the model with GhostNet16, SPP, and ASA-FPN applied. YOLO-A exhibits no significant change in *mAP* compared to YOLOv3, but the number of model parameters is reduced by 38.65 MB, and the detection speed is reduced by 19.19 ms compared to Darknet53, which shows that utilizing the lightweight GhostNet16 as a backbone network can reduce the number of model parameters without affecting the detection accuracy. Furthermore, *mAP* improved by 0.74% for YOLO-B compared to YOLO-A and 1.39% for MYOLO compared to YOLO-B. The experimental results demonstrate that the models using SPP and ASA-FPN can improve the detection performance without significantly affecting the number of model parameters. The proposed MYOLO model improves the *mAP* by 2.18% and decreases the number of model parameters by 32.16M. Therefore, the ablation experiments show that MYOLO is effective and feasible, considering the balance between model parameters and accuracy.

Table 4. Test results obtained for the four algorithms on the test set (IoU threshold set to 0.5).

Model	FPN	GhostNet16	SPP	ASA-FPN	mAP	Total Parameters	Speed
YOLOv3	√	×	×	×	94.85%	61.53 M	35.94 ms
YOLO-A	√	√	×	×	94.90%	22.88 M	17.45 ms
YOLO-B	√	√	√	×	95.64%	23.93 M	18.01 ms
YOLO-M	×	√	√	√	97.03%	29.37 M	19.78 ms

Figure 13 presents the detection results for the different models in the mushroom shed. The different species of fresh shiitake mushrooms are indicated by the different coloured bounding boxes; the yellow boxes correspond to missed fresh shiitake mushroom detections. Here, A1–A3 are the YOLOv3 detection results, B1–B3 are the YOLO-A detection results, C1–C3 are the YOLO-B detection results, and D1–D3 are the MYOLO detection results. Figure 13 demonstrates that the replacement of the backbone network, GhostNet16, does not influence the detection effect, but like in the original model, missed detection is more serious for dense occlusions and large field-of-view situations, as shown in B1 and B2. YOLO-B can detect small fresh shiitake mushrooms more accurately but still has some problems with mushroom localization, as depicted in C3. Using the ASA-FPN neck network MYOLO can detect mushrooms more accurately in dense occlusions and complex environments to maintain better detection, localization accuracy, and improves the missed detection under a large field of view effectively, as demonstrated by column D.

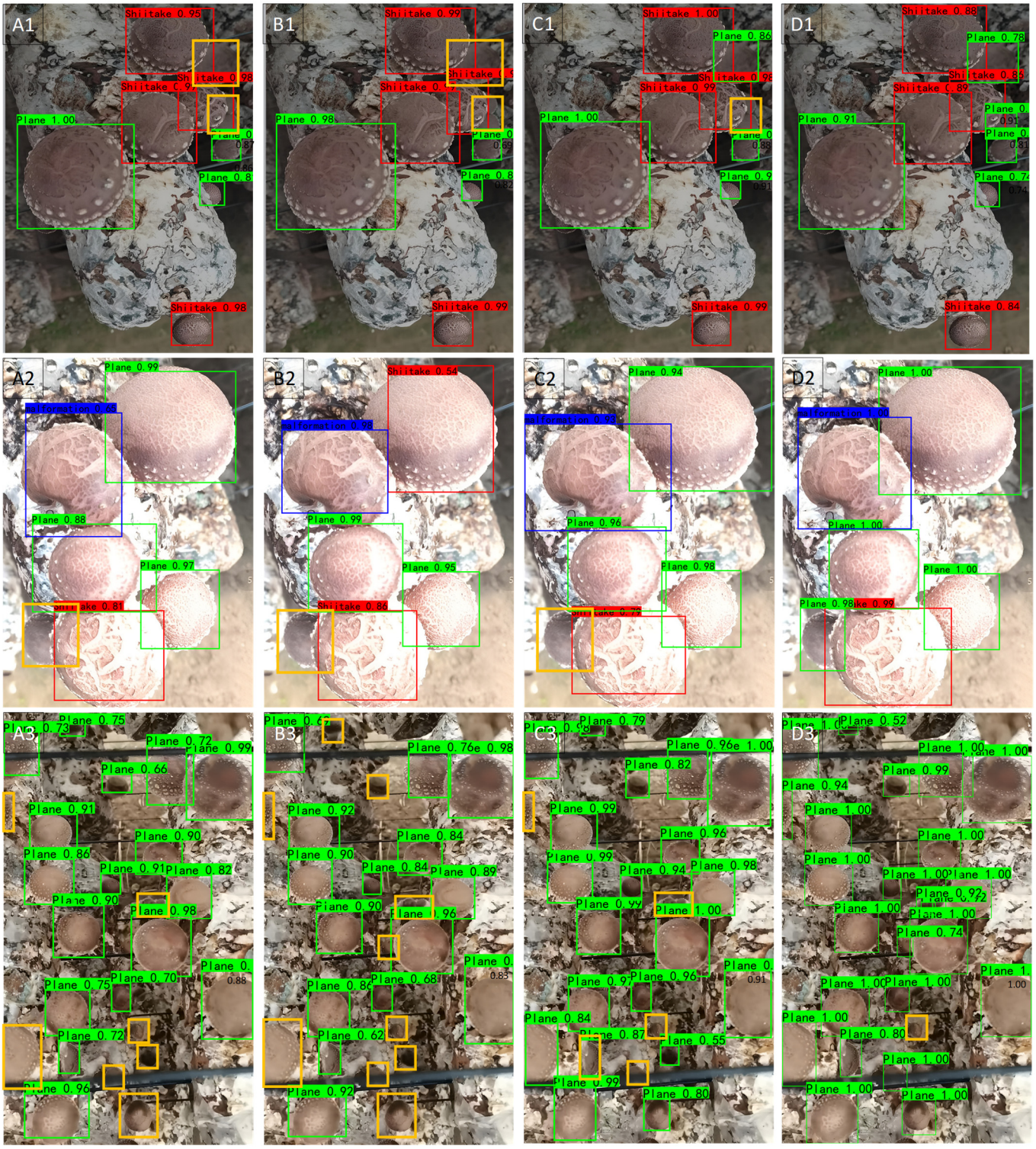


Figure 13. Comparison of improved module performance. (The yellow box is the target of the missed detection).

4.3. Multiscene Detection Performance Analysis

To demonstrate the detection performance of MYOLO under complex conditions such as diverse morphologies of fresh shiitake mushrooms, dense growth, occlusion, and variable fields of view, the test set was further classified to construct the above-mentioned complex scenarios. Accordingly, the performance of MYOLO was tested against

those of four mainstream target detection algorithms: Faster RCNN, SSD, YOLOv3, and YOLOv5-m [55].

The test set consisted of 663 images of fresh shiitake mushrooms, which were divided to obtain 239 images of dense occlusion, 241 images of different lighting, and 135 images of large fields of view. The detection results are shown in Figure 14. The average accuracy of YOLO-M for the test set was 97.03%, better than those of all other target detection algorithms. The lightened YOLO-M target detection model outperformed the other algorithms in different scenarios for the detection of cracked-surface mushrooms, plane-surface mushrooms, and malformed-surface mushrooms, indicating that YOLO-M can better detect and classify fresh mushrooms. Compared with the *mAP* of YOLOv3, that of YOLO-M is improved by 3.56% for the detection of dense occlusion, which indicates that the addition of ASFF solves the problem of dense occlusion of fresh shiitake mushrooms to a certain extent. The largest difference was in the large field-of-view case, where the *mAP* for YOLO-M is 9.64%, 10.88%, 5.00%, and 3.70% higher than the *mAP* values for Faster RCNN, SSD, YOLOv3, and YOLOv5-m, respectively, proving the powerful ability of YOLO-M to extract fresh shiitake mushroom surface features in the large field-of-view case and the effectiveness of SPP and SANet incorporation.

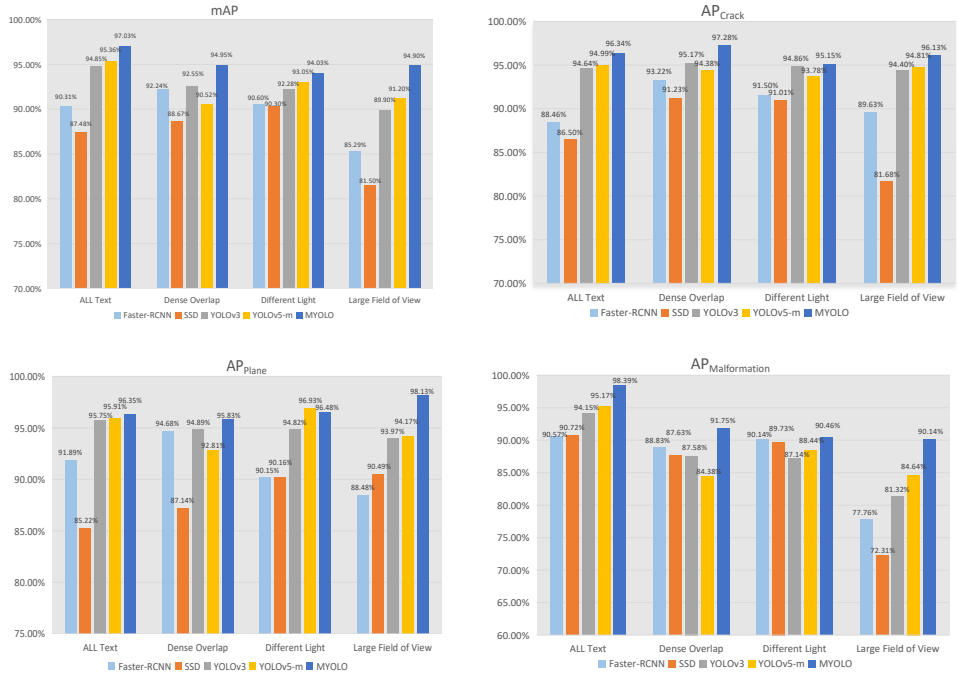


Figure 14. Detection results for each algorithm in different scenarios (IoU threshold set to 0.5).

Figure 15 reveals that all five detection algorithms mentioned above are more accurate in identifying the species of fresh shiitake mushrooms when the actual scene contains more species of fresh shiitake mushrooms. Lowercase A is a mixed scene, lowercase B is a dense occlusion scene, lowercase C and D are different lighting scenes, and lowercase E is a large field of view scene. For cases in which the target is small owing to the shooting distance, MYOLO achieves higher accuracy than the other algorithms, as depicted in Figure 15a,e. Both Faster-RCNN and MYOLO can accurately identify fresh shiitake mushrooms even when they are densely distributed and in the presence of shading, which reflects the high adaptability of both, as revealed by Figure 15b. Regarding the effects of light intensity, MYOLO also ensures good stability in detection; in particular, when there is insufficient light, fresh shiitake mushrooms that are obscured can be detected, as shown in Figure 15d. For densely obscured fresh shiitake mushrooms, as shown in Figure 15b,c, MYOLO also showed its stability, and the rest of the models produced missed detections when occluded, while MYOLO did not. Under large field-of-view conditions, all fresh shiitake mushrooms

are detected as plane-surface mushrooms as the back pattern of fresh shiitake mushrooms is not obvious at a distance; when the camera comes closer, fresh shiitake mushrooms are classified correctly, as shown in Figure 15a,d,e. Although some small fresh shiitake mushrooms are still missed in the large field-of-view situation, Figure 15e shows that MYOLO minimizes the missed detection. Thus, MYOLO can accurately detect fresh shiitake mushrooms in various complex environments, and its prediction of the position and category of frames is accurate and has strong robustness.

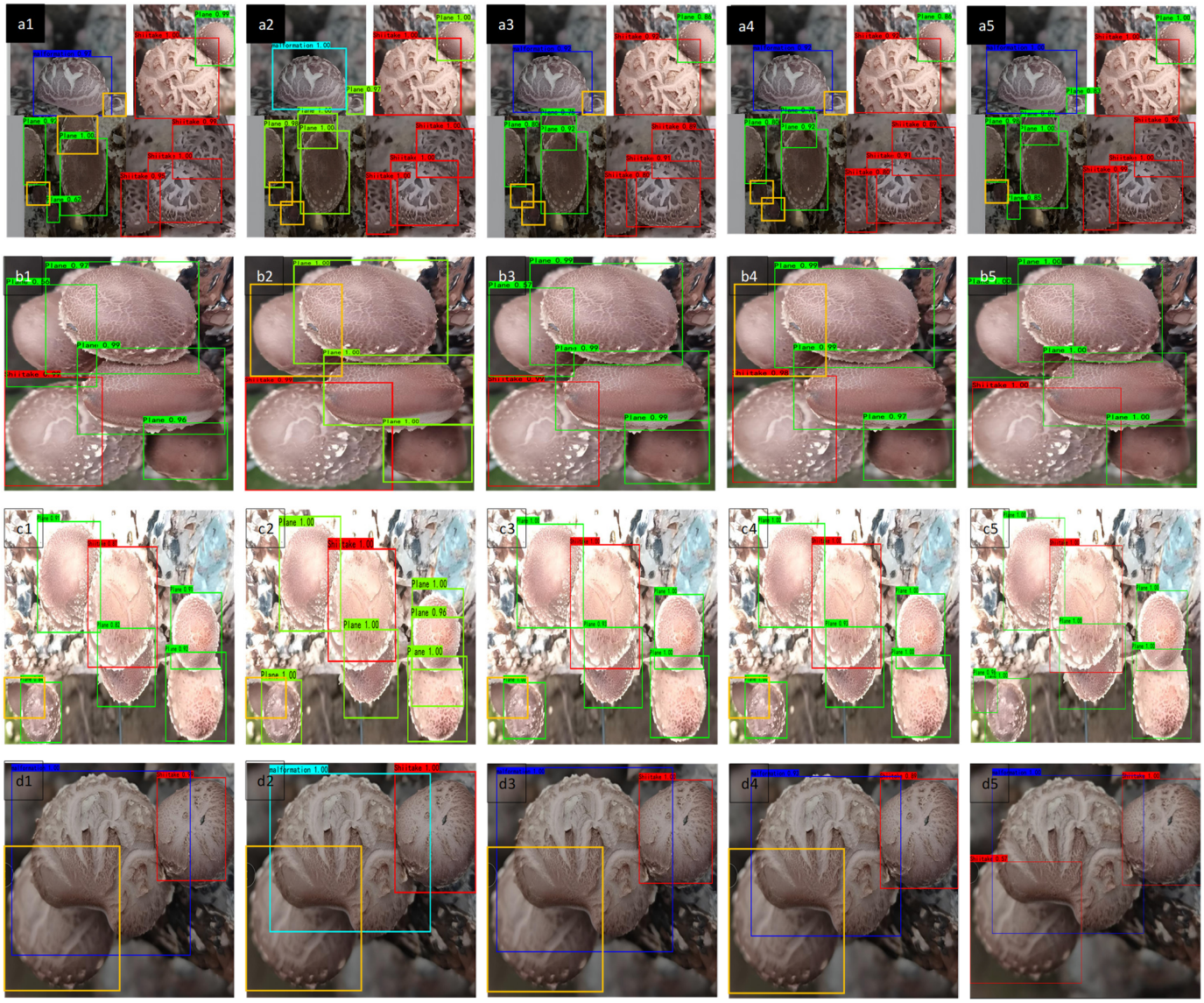
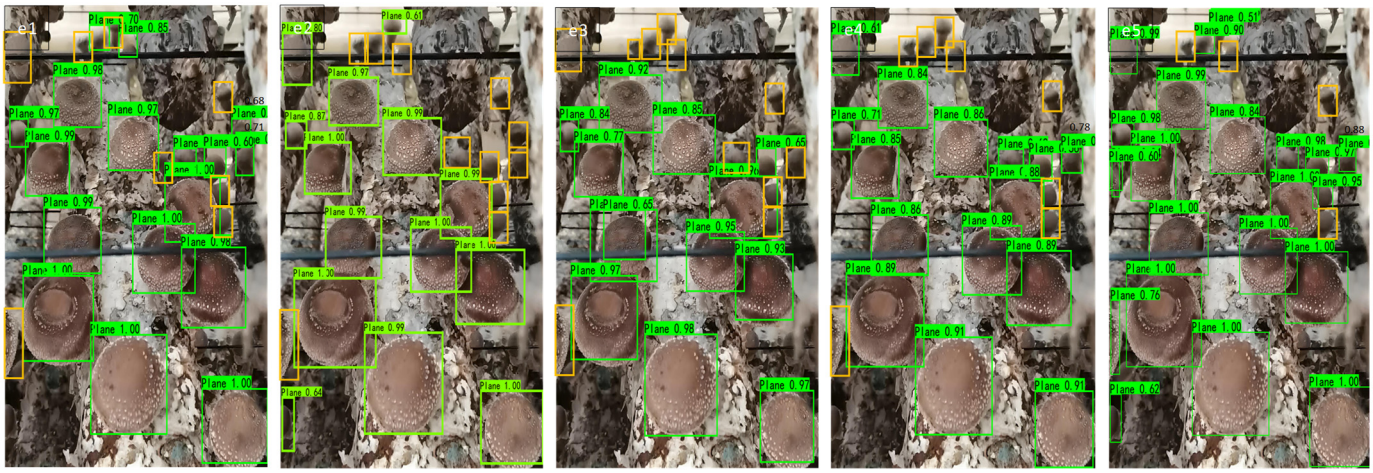


Figure 15. *Cont.*



Faster RCNN **SSD** **YOLOv3** **YOLOv5-m** **MYOLO**

Figure 15. Detection results for each algorithm in different scenarios. (The yellow box is the target of the missed detection).

4.4. Feasibility Analysis of Picking Robot Applications

To demonstrate the feasibility of using MYOLO in picking robot applications, we compared its time performance and model complexity with those of four mainstream target detection algorithms—Faster RCNN, SSD, YOLOv3, and YOLOv5-m—on a test set, and the results are shown in Table 5. After the lightweighting process, MYOLO had an *mAP* of 97.03%, the number of model parameters was 29.8 MB, detection speed was 19.78 ms, and the number of floating point operations (FLOPs) was 21.36 G. The above-mentioned metrics indicate that MYOLO has higher detection performance than models such as YOLOv3 and can better detect and classify fresh shiitake mushrooms. Although YOLOv5-m is superior in terms of detection time and total number of model parameters, its detection accuracy is not sufficiently high. The detection speed of MYOLO is 19.78 ms, which is ideal for porting and embedded development, and the model inference speed can guarantee that the real-time requirements for the detection and classification process are satisfied, which can meet the practical needs of picking robots for fresh shiitake mushroom detection tasks.

Table 5. Algorithm detection performance comparison (IoU threshold set to 0.5).

Algorithm	FLOP (G)	Total Parameters (M)	Speed (ms)	*F*1 (%)	*mAP* (%)
Faster-RCNN	370.21	137.1	129.65	89.30	90.31
SSD	62.75	26.3	23.02	85.67	87.48
YOLOv3	66.17	62.0	35.94	92.01	94.85
YOLOv5-m	21.38	21.3	17.95	92.33	95.36
MYOLO	21.36	29.8	19.78	94.02	97.03

4.5. Discussion

Unlike fruits such as strawberries [12], pomegranates [17] and kiwi fruit [19], fresh shiitake mushrooms differ not only in size, shape and compactness of different categories, but also in different growing periods within the same category. In addition, fresh shiitake mushrooms grow in the environment of mushroom sheds and are susceptible to complex environmental influences such as light changes, overlaps, shadows, and occlusion. Because of these factors, accurate shiitake mushroom detection is very challenging. Most of the experimental images come from indoor shooting, and a small part comes from the network, so the difference between the images is more obvious, which increases the difficulty of detection greatly. Liu et al. [32] used the YOLOvX algorithm to detect the quality of

shiitake mushrooms, and the mAP was as high as 99.96%, but they did not classify shiitake mushroom varieties, and their experimental scene was too rationalized (moving shiitake mushrooms to a specific shooting site after picking), which did not conform to the actual picking scene. Zhang et al. [56] combined the YOLOv3 algorithm with the Rao-1 algorithm for automatic detection of damaged apples; although the mAP is 5.03% higher than the original method, it does not consider that the complexity of the network model is too high, and the detection speed is not reliable in actual deployment. Bazame et al. [57] used lightweight YOLOv3 to classify and detect coffee cherries, and although the detection speed was improved, the mAP was only 84%, resulting in frequent false or missed detections in the detection task. The above author's improvement of YOLO cannot balance the relationship between detection speed and detection accuracy, and is not suitable for the algorithm deployment of sorting robots.

From the results of Table 4 and Figure 5, it can be seen that the detection of various fresh shiitake mushrooms by the MYOLO detection model has achieved good results, and its mAP reaches 97.03% under the total test set, 94.95% under the secret shadow test set, 94.03% under different light datasets, and 94.90% under the large field of view dataset. The above results not only indicate that the MYOLO network model can adapt to changes in image quality and complex environment, but also prove its robustness. At the same time, the time performance and model complexity of MYOLO are also compared with the other four detection models in Table 5, the detection speed of MYOLO is 19.78 ms and the model complexity is 21.36 G, which met the speed requirements of real-time picking.

From what has been discussed above, it can be seen that MYOLO can detect fresh shiitake mushrooms in complex scenarios, balancing the relationship between detection speed and detection accuracy, which ensures detection accuracy while meeting the requirements of detection speed, it overcomes the disadvantages of the current YOLO algorithm in fruit and vegetable detection effectively [32,56,57]. So MYOLO is convenient for network deployment to mobile devices, and it is more applicable to the detection of fresh shiitake mushrooms. For crop detection in other applications (such as: crop classification and localization, disease degree estimation, etc.), Zhang et al. [58] combined YOLOv5x and SE for weed crop classification and lettuce localization, for which the *mAP* was as high as 97.3%, detection speed was 19.3 ms, detection speed and accuracy reached a good balance. Gao et al. [59] used the automatic tandem dual BlendMask deep learning framework and ResNet-50 and FPN as the backbone network of the blank mask to evaluate the severity of Fusarium head blight in wheat, and the average accuracy of Fusarium head wilt severity classification reached 91.80%. In future research, methods applied in other aspects of crop detection can also provide ideas for the algorithm improvement of picking robots.

5. Conclusions and Future Work

For the visual perception element of a picking robot, this study proposed a lightweight MYOLO detection model to address the problems of diverse morphology, dense growth, easy occlusion, and variable field of view in the detection of fresh shiitake mushrooms, providing a theoretical basis for a vision detection system for picking robots. A lightweight GhostNet16 was constructed as the backbone network in the MYOLO model to improve the network detection speed for fresh shiitake mushrooms. SPP was introduced to improve the detection accuracy of the model for small fresh shiitake mushrooms. In addition, a new feature fusion network, ASA-FPN, was designed to improve the detection and localization accuracy of fresh shiitake mushrooms and to increase the detection of fresh shiitake mushrooms under large field-of-view conditions to a certain extent. The results showed that the MYOLO model has high accuracy and speed for detecting different categories of fresh shiitake mushrooms.

In addition, the recognition results of four models were compared on 663 images. The comparison metrics revealed that the *F*1 value of MYOLO was 4.72%, 8.35%, 2.01%, and 1.69% higher than those of Faster RCNN, SSD, YOLOv3, and YOLOv5-m, and the *mAP* was 6.72%, 9.55%, 2.18%, and 1.67% higher, respectively; the detection speed was 19.78 ms,

which is 31.15, 3.38, and 6.45 times higher than those of the first three types of models. With high detection accuracy and real-time performance, MYOLO can meet the needs of picking robots for real-time detection of multiple categories of fresh shiitake mushrooms in mushroom sheds.

In the future, we will continue to improve the MYOLO model in terms of fresh shiitake mushroom grade classification, light intensity, and the effects of various complex environments.

Author Contributions: Conceptualization: P.C. and H.F.; data curation: H.F.; formal analysis: H.F.; funding acquisition: P.C.; investigation: P.C. and H.F.; methodology: H.F.; project administration: P.C. and H.F.; resources: P.C., H.F. and K.L.; software: H.F.; supervision: P.C. and H.F.; validation: H.F. and J.Z.; visualization: H.F. and S.L.; writing—original draft: H.F.; writing—review and editing: P.C., H.F. and K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Central Government Guides Local Science and Technology Development Foundation Projects (grant no.ZY19183003), Guangxi Key Research and Development Project (grant no.AB20058001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Vaishnavi, M.; Sharma, A.; Tiwari, J.; Singh, S.; Sharma, S. Production of edible mushrooms to meet the food security: A review. *J. Posit. Psychol.* **2022**, *6*, 4316–4325. [CrossRef]
2. Wang, M.; Zhao, R. A review on nutritional advantages of edible mushrooms and its industrialization development situation in protein meat analogues. *J. Funct. Foods.* **2023**, *3*, 1–7. [CrossRef]
3. Cheute, V.M.S.; Backes, E.; Corrêa, R.C.G. The Global Market for Mushrooms, Their Uses as Dietary Supplements and Associated Safety Issues. In *Edible Fungi: Chemical Composition, Nutrition and Health Effects*; The Royal Society of Chemistry: London, UK, 23 November 2022; Volume 383, ISBN 978-1-83916-401-9.
4. Tang, Y.; Chen, M.; Wang, C.; Luo, L.; Li, J.; Lian, G.; Zou, X. Recognition and localization methods for vision-based fruit picking robots: A review. *Front. Plant Sci.* **2020**, *11*, 510. [CrossRef] [PubMed]
5. Arefi, A.; Motlagh, A.M.; Mollazade, K.; Teimourlou, R.F. Recognition and localization of ripen tomato based on machine vision. *Australian J. Crop Sci.* **2011**, *5*, 1144–1149. [CrossRef]
6. Wei, X.; Jia, K.; Lan, J.; Li, Y.; Zeng, Y.; Wang, C. Automatic method of fruit object extraction under complex agricultural background for vision system of fruit picking robot. *Optik* **2014**, *125*, 5684–5689. [CrossRef]
7. Lu, J.; Sang, N. Detecting citrus fruits and occlusion recovery under natural illumination conditions. *Comput. Electron. Agri.* **2015**, *110*, 121–130. [CrossRef]
8. Xiong, J.; Lin, R.; Liu, Z.; He, Z.; Tang, L.; Yang, Z.; Zou, X. The recognition of litchi clusters and the calculation of picking point in a nocturnal natural environment. *Biosyst. Eng.* **2018**, *166*, 44–57. [CrossRef]
9. 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]
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2016**, *39*, 1137–1149. [CrossRef]
11. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387. [CrossRef]
12. Lamb, N.; Chuah, M.C. A strawberry detection system using convolutional neural networks. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2515–2520.
13. Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agri.* **2019**, *163*, 104846. [CrossRef]
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
16. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision(CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2961–2969.
17. Lin, G.; Tang, Y.; Zou, X.; Xiong, J.; Li, J. Guava detection and pose estimation using a low-cost RGB-D sensor in the field. *Sensors* **2019**, *19*, 428. [CrossRef] [PubMed]

18. Mu, Y.; Chen, T.-S.; Ninomiya, S.; Guo, W. Intact detection of highly occluded immature tomatoes on plants using deep learning techniques. *Sensors* **2020**, *20*, 2984. [CrossRef] [PubMed]
19. Liu, Z.; Wu, J.; Fu, L.; Majeed, Y.; Feng, Y.; Li, R.; Cui, Y. Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion. *IEEE Access* **2019**, *8*, 2327–2336. [CrossRef]
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Computer Vision—ECCV 2016. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland; pp. 21–37.
22. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
23. Koirala, A.; Walsh, K.; Wang, Z.; McCarthy, C. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. *Precis. Agri.* **2019**, *20*, 1107–1135. [CrossRef]
24. Li, H.; Li, C.; Li, G.; Chen, L. A real-time table grape detection method based on improved YOLOv4-tiny network in complex background. *Biosyst. Eng.* **2021**, *212*, 347–359. [CrossRef]
25. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
26. Lu, S.; Chen, W.; Zhang, X.; Karkee, M. Canopy-attention-YOLOv4-based immature/mature apple fruit detection on dense-foliage tree architectures for early crop load estimation. *Comput. Electron. Agri.* **2022**, *193*, 106696. [CrossRef]
27. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
28. Wang, Y.; Yan, G.; Meng, Q.; Yao, T.; Han, J.; Zhang, B. DSE-YOLO: Detail semantics enhancement YOLO for multi-stage strawberry detection. *Comput. Electron. Agri.* **2022**, *198*, 107057. [CrossRef]
29. Saleem, M.H.; Potgieter, J.; Arif, K.M. Automation in agriculture by machine and deep learning techniques: A review of recent developments. *Precis. Agri.* **2021**, *22*, 2053–2091. [CrossRef]
30. Fang, L.; Wu, Y.; Li, Y.; Guo, H.; Zhang, H.; Wang, X.; Xi, R.; Hou, J. Using channel and network layer pruning based on deep learning for real-time detection of ginger images. *Agriculture* **2021**, *11*, 1190. [CrossRef]
31. Zulkiflfley, M.A.; Moubark, A.M.; Saputro, A.H.; Abdani, S.R. Automated apple recognition system using semantic segmentation networks with group and shuffle operators. *Agriculture* **2022**, *12*, 756. [CrossRef]
32. Liu, Q.; Fang, M.; Li, Y.; Gao, M. Deep learning based research on quality classification of shiitake mushrooms. *LWT* **2022**, *168*, 113902. [CrossRef]
33. Yu, L.; Pu, Y.; Cen, H.; Li, J.; Liu, S.; Jing, N.; Ge, J.; Lv, L.; Li, Y.; Xu, Y.; et al. A lightweight neural network-based method for detecting estrus behavior in ewes. *Agriculture* **2022**, *12*, 1207. [CrossRef]
34. Xiang, R.; Zhang, M.; Zhang, J. Recognition for stems of tomato plants at night based on a hybrid joint neural network. *Agriculture* **2022**, *12*, 743. [CrossRef]
35. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
36. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
37. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 19–21 June 2018; pp. 7132–7141.
39. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
40. Fu, L.; Li, S.; Kong, S.; Ni, R.; Pang, H.; Sun, Y.; Hu, T.; Mu, Y.; Guo, Y.; Gong, H. Lightweight individual cow identification based on Ghost combined with attention mechanism. *PLoS ONE* **2022**, *17*, e0275435. [CrossRef]
41. 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]
42. Wang, N.; Qian, T.; Yang, J.; Li, L.; Zhang, Y.; Zheng, X.; Xu, Y.; Zhao, H.; Zhao, J. An enhanced YOLOv5 model for greenhouse cucumber fruit recognition based on color space features. *Agriculture* **2022**, *12*, 1556. [CrossRef]
43. Yu, Z.; Liu, Y.; Yu, S.; Wang, R.; Song, Z.; Yan, Y.; Li, F.; Wang, Z.; Tian, F. Automatic detection method of dairy cow feeding behaviour based on YOLO improved model and edge computing. *Sensors* **2022**, *22*, 3271. [CrossRef] [PubMed]
44. Zhang, Q.L.; Yang, Y.B. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, 6–11 June 2021; pp. 2235–2239.
45. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516.
46. MacQueen, J. Classification and Analysis of Multivariate Observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965, 27 December 1965–7 January 1966; Volume 5, pp. 281–297.
47. Robbins, H.; Monro, S. A stochastic approximation method. *Annals Math. Stat.* **1951**, *22*, 400–407. [CrossRef]

48. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
49. Liu, C.Y.; Wu, Y.Q.; Liu, J.J.; Sun, Z. Improved YOLOv3 network for insulator detection in aerial images with diverse background interference. *Electronics* **2021**, *10*, 771. [CrossRef]
50. Khasawneh, N.; Mohammad, F.; Luay, F. Detection of K-complexes in EEG signals using deep transfer learning and YOLOv3. *Cluster Comput* **2022**, 1–11. [CrossRef]
51. Cong, P.; Lv, K.; Feng, H.; Zhou, J. Improved YOLOv3 Model for Workpiece Stud Leakage Detection. *Electronics* **2022**, *11*, 3430. [CrossRef]
52. Khasawneh, N.; Faouri, E.; Fraiwan, M. Automatic Detection of Tomato Diseases Using Deep Transfer Learning. *Appl. Sci.* **2022**, *12*, 8467. [CrossRef]
53. Huang, J.; Qu, L.; Jia, R.; Zhao, B. O2u-net: A Simple Noisy Label Detection Approach for Deep Neural Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 3326–3334.
54. He, X.; Cheng, R.; Zheng, Z.; Wang, Z. Small object detection in traffic scenes based on YOLO-MXANet. *Sensors* **2021**, *21*, 7422. [CrossRef]
55. Ultralytics. YOLOv5. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 18 May 2020).
56. Zhang, M.; Liang, H.; Wang, Z.; Wang, L.; Huang, C.; Luo, X. Damaged Apple Detection with a Hybrid YOLOv3 Algorithm. *Inf. Process* **2022**, *in press*. [CrossRef]
57. Bazame, H.C.; Molin, J.P.; Althoff, D.; Martello, M. Detection, classification, and mapping of coffee fruits during harvest with computer vision. *Comput. Electron. Agri.* **2021**, *183*, 106066. [CrossRef]
58. Zhang, J.L.; Su, W.H.; Zhang, H.Y.; Peng, Y. SE-YOLOv5x: An Optimized Model Based on Transfer Learning and Visual Attention Mechanism for Identifying and Localizing Weeds and Vegetables. *Agronomy* **2022**, *12*, 2061. [CrossRef]
59. Gao, Y.; Wang, H.; Li, M.; Su, W.H. Automatic Tandem Dual BlendMask Networks for Severity Assessment of Wheat Fusarium Head Blight. *Agriculture* **2022**, *12*, 1493. [CrossRef]