Yolo-Chili: An Efficient Lightweight Network Model for Localization of Pepper Picking in Complex Environments

HaiLin Chen , Ruofan Zhang , Jialiang Peng , Hao Peng , Wenwu Hu , Yi Wang * , Ping Jiang *

Posted Date: 29 April 2024

doi: 10.20944/preprints202404.1916.v1

Keywords: Chili Detection,Automatic Picking,Neural Network,Model Compression

Article

Yolo-Chili: An Efficient Lightweight Network Model for Localization of Pepper Picking in Complex Environments

HaiLin Chen [1], RuoFan Zhang [1], JiaLiang Peng [1], Hao Peng [1], WenWu Hu [2], Yi Wang [1,*] and Ping Jiang [2,*]

[1] School of Information and Intelligent Science and Technology, Hunan Agricultural University, Changsha, Hunan;1152400900@stu.hunau.edu.cn (C.H.)
[2] College of Mechanical and Electrical Engineering, Hunan Agricultural University, Changsha, China
* Correspondence: wangyi@hunau.edu.cn;

Abstract: Currently there are fewer depth models applied to pepper picking detection, while the existing generalized neural networks have problems such as large model parameters, long training time, and low model accuracy.In order to solve the above problems, this paper proposes a Yolo-chili target detection algorithm for chili pepper detection. First, the classical target detection algorithm yolov5 is used as a benchmark model, and an adaptive spatial feature pyramid structure combining the attention mechanism and the idea of multi-scale prediction is introduced to improve the model's detection effect on occluded peppers and small target peppers. Secondly, a three-channel attention mechanism module is introduced to improve the algorithm's long-distance recognition ability and reduce the interference of redundant testers. Finally, the quantized pruning method is used to reduce the model parameters and realize the lightweight processing of the model. Applying the method to the homemade chili pepper dataset, the AP value of chili pepper reaches 93.11%; the accuracy rate is 93.51% and the recall rate is 92.55%.The experimental results show that yolo-chili is able to achieve accurate and real-time pepper detection under complex orchards.

Keywords: chili detection; automatic picking; neural network; model compression

1. Introduction

In 2021, China's pepper planting area accounted for 36.72% of the global planting area, and the production accounted for nearly half of the world, but at present, the degree of mechanized picking in China is low, because the current target detection algorithms can't effectively identify the specific location of the pepper.

Deep learning algorithms have been proven to be the most robust target detection methods for automatic fruit picking, and many researchers have used different target detection methods for mAP and detection speed [1–13]. Considering the accuracy and speed, Addie [14] et al. used a variant of yolov4 and Deep SORT to provide a robust real-time pear fruit counter for a mobile application, which provides an effective support for automatic pear fruit picking and yield prediction. For the effects of environmental factors such as stem and leaf shading, uneven illumination, and fruit overlapping, Lawal [15] proposed the YOLOFruit algorithm, which uses a spatial pyramid and a feature pyramid network to extract the detailed features, resulting in a fruit detection network with an average detection accuracy of 86.2% and a detection time of 11.9 ms. Li [16] achieved 94.77% accuracy and 25.86 ms detection speed by segmenting the red region of tomato using HSV in the detection frame of yolov4 and using the tomato target with segmented area exceeding a certain percentage as the output. Similarly for the problem of picking peppers in natural environments, guo [17] et al. introduced a deformable convolution and coordinate attention module in yolov5, which improved the mAP by 4.6% compared to the original model, and achieved a real-time detection speed

 doi:10.20944/preprints202404.1916.v1

of 89.3 frames/sec on a mobile picking platform. However, due to the diversity of pepper picking devices, the complex structure and large parameters of the above model make it difficult to deploy it to mobile hardware devices for real-time detection.

Many researchers realized the problem that huge models are difficult to be deployed to mobile devices, so they started to explore the path of lightweight models. Yang [18] et al. used 76×76 detection head with CBAM attention mechanism network added to yolov4-tiny network, which reduces the number of model parameters while effectively solving the problem of occlusion and the low accuracy of small tomato recognition. While wang et al. [19] added CBAM to FPN to learn the correlation of features between different channels by assigning weights to the features of each channel, to strengthen the transmission of deep information of the network structure, so as to reduce the interference of the complex background on the target recognition, and this kind of detection network has fewer network layers and occupies low memory. However, this is only using a lightweight network, and on the basis of which the accuracy is improved, and there is no substantial change. In contrast, Sun et al. [20] obtained a small baseline model based on YOLOv5s by adding phantom structures and adjusting the overall width of the feature map, and introduced the migration learning technique, which realized a fast and accurate identification of apple chilis while occupying less computational resources. Similarly, rui et al. [21] proposed a classification model for pepper quality detection based on the combination of migration learning and convolutional neural network, which achieved fast convergence and performance improvement in pepper detection. However, this kind does not achieve the lightweight of the model, and it is more to reduce the resources to achieve the model training. On the other hand, zhou et al. [22] started from the equipment requirements, eliminated the feature mapping used for detecting large targets in the YOLOX model, sampled the feature mapping of small targets through the nearest neighbor value, spliced the surface features with the final features, perturbed the gradient of the SiLU activation function, and optimized the loss function at the output, which resulted in a reduction of the number of model parameters by 44.8%, and an increase in the speed of model detection by 63.9% with excellent performance. Zhang et al. [23] implemented a GhostNet feature extraction network with a coordinate attention module in YOLOv4 and introduced deeply differentiable convolution to reconstruct the neck and YOLO head structure, thus realizing a lightweight apple detection model. However, these methods have limited changes to the model parameters, and the model performance is also degraded by the decrease of parameters, which has some defects. Aiming at these problems, Wang et al. [24] used migration learning to establish a YOLO V5s detection model, and at the same time used a channel pruning algorithm to prune the YOLO V5s model, and fine-tuned the pruned model, which achieved an apple detection accuracy of 95.8%, with an average detection time of 8 ms /sheet, and the model size of only 1.4 MB, which effectively reduces the model size and ensures the effectively reduce the model size and ensure the model performance.

The success of the above methods proves the success of target detection in the field of fruit picking, but due to the problems of dense growth of chili fruits, uneven fruit size, severe occlusion of fruits by branches and leaves and similar backgrounds in chili pepper picking, it is difficult to target chili pepper fruits for efficient picking with the above methods [25–32]. At the same time, some of the current general-purpose models have problems such as insufficient model detection performance, large environmental interference factors, large model structure, and slow inference speed. In order to develop a deep learning model to meet the actual picking needs and realize intelligent picking of chili peppers. In this paper, in view of the existing problems of the current pepper picking model, it is proposed to use the three-channel attention mechanism network to help the neural network to extract the long-distance pepper information, to improve the model's ability to recognize small target peppers, and to solve the problem that the current CBAM can not extract the long-distance information effectively. At the same time, the backbone network based on yolov5 is trained and the same detection mechanism is used, so as to ensure that the model can be transplanted to different devices and has the function of real-time detection. Then, a multi-scale prediction algorithm is established to improve the prediction layer structure of yolov5 so that it can detect peppers of different shapes, such as large, medium, small and medium-sized peppers, and improve the detection

 doi:10.20944/preprints202404.1916.v1

ability of small-targeted peppers. Finally, a multi-scale adaptive feature fusion pyramid is established to improve the model performance by introducing an adaptive spatial feature pyramid structure and combining the attention mechanism to suppress the background noise that causes interference, and at the same time adaptively fusing the features of different scales in the final prediction results.

2. Materials and Methods

2.1. Data Acquisition

The chili pepper dataset used in this study was obtained from a chili pepper trellis garden in Changsha, Hunan, China. We collected different light conditions at 8:30 a.m., 1:00 p.m., and 5:00 p.m. on May 7, 2022, November 2, 2022, 2023 August 10, 2023, and September 17, 2023 at 8:30 a.m., 1:00 p.m., and 5:00 p.m. Images of chili peppers under different light conditions were collected. The image resolution was 4000 × 4000 pixels. A total of 1456 raw images were collected, of which 762 were of densely distributed chili peppers and 696 were of sparsely distributed chili peppers. Among the densely distributed chili pepper images there were journals in which the chili fruits were occluded from each other, occluded by leaves, and appeared in multiple targets. The details are shown in Figure 1. The dataset is publicly available at https://www.kaggle.com/datasets/jingxiche/chili-data.

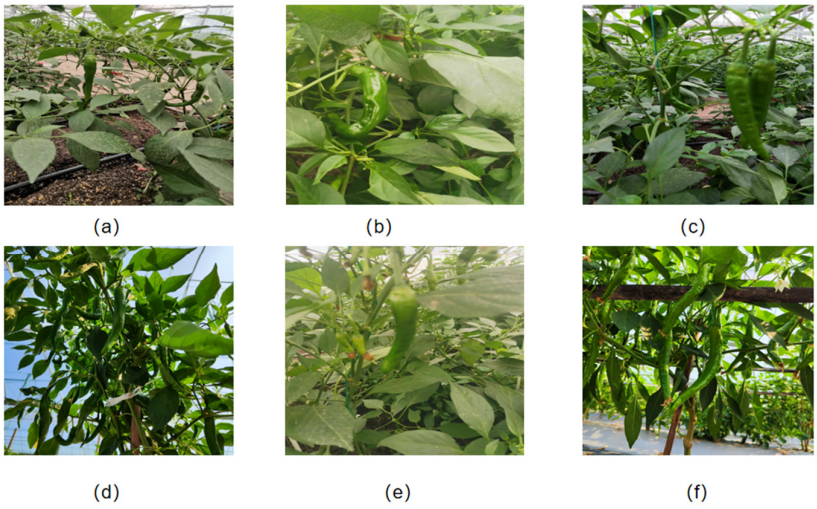


Figure 1. Photographs of chili peppers in different light with different shooting angles. Figures a–c show peppers photographed in cloudy weather, and Figures d–f show peppers photographed in sunny weather. The shooting angles of chili peppers were categorized into top, top and flat view.

2.2. Data Acquisition

We manually labeled 1456 original images using LabelImg to divide the dataset and test set in the ratio of 8:2. Due to the problem of the original data being too small, we expanded the dataset to 13176 by adding Gaussian noise (mean= 0, variance= 0.001), random rotation, random brightness change, and random scaling to the dataset, so as to improve the model's generalization ability and to ensure the model's practical adaptability. Meanwhile, in order to improve the model's recognition ability for small target chili peppers, the pre-trained backbone weights on the coco dataset are also used to improve the model's detection ability.

 doi:10.20944/preprints202404.1916.v1

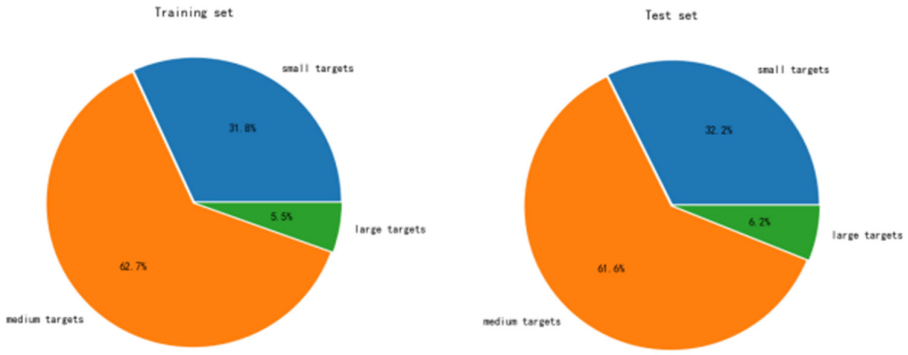


Figure 2. Distribution of chili peppers at different scales in the chili pepper dataset.

2.3. Experimental Environment

In this paper, the performance of yolo-chili is investigated for the experimental environment as shown in Table 1.

Table 1. experimental environment.

Configure	Para
CPU	core i5-11400H
GPU	Nvidia GeForce RTX 3050TI
Accelerated environment	CUDA10.1 CUDNN7.5.0
development environment (computer)	Pycharm2020.1.3
operating system	Windows 10　64-bit system
software environment	Anaconda 4.8.4
storage environment	Memory 16.0GB Mechanical Hard Disk 2T

2.4. HFFN (Hierarchical Feature Fusion Network) Hierarchical Feature Fusion Network Module

In the process of chili pepper detection, it is inevitable that chili pepper targets with different levels of size will appear in the same image, which will seriously interfere with the recognition accuracy of chili peppers. However, the original yolov5 feature pyramid is only applicable to the detection of chili pepper targets with a small degree of hierarchical change in chili pepper size, and performs poorly in the detection process where there are chili peppers with large hierarchical changes in an image. In this paper, we introduce adaptive spatial feature fusion (ASFF) in the model to address the above drawbacks, and set the convolution kernel in ASFF to a size of 3*3 to adapt it to the chili pepper targets in this paper's dataset. Therefore, yolo-chili contains a total of three ASFF prediction layers, which are responsible for processing different levels of chili pepper feature information, among which the first layer is the smallest layer of the feature map, with a channel number of 512, which is responsible for processing the feature information of small-scale chili peppers. The second layer is the layer with moderate feature map size and channel number 256, which deals with the feature information of medium scale chili peppers. The third layer is the layer with the largest feature map, channel number 128, dedicated to processing feature information of large-scale peppers. The yolo-chili containing three ASFF prediction layers is able to handle chili data with large variations in chili levels in the same map, and thus is fully adapted to the task of chili detection in orchards under complex conditions. The HFFN structure is shown in Figure 3.

 doi:10.20944/preprints202404.1916.v1

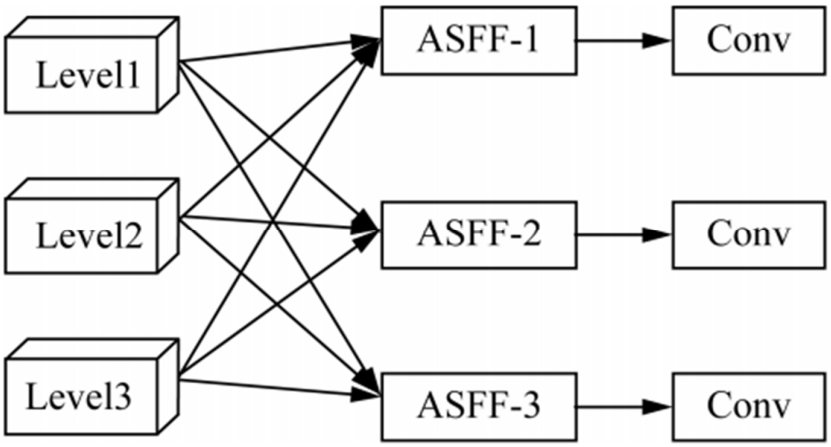


Figure 3. HFFN Structure.

2.5. Three-Channel Attention Mechanism

In HFFN, although yolo-chili can effectively enhance the detection ability of the model for different layers of targets, it also brings a large amount of environmental noise to the model's detection, which makes the final detection results receive interference. Therefore, to address the above problems, this paper proposes a three-channel attention mechanism model. Because the three-channel attention mechanism consists of CBAM attention mechanism and CA attention mechanism, it is abbreviated as CBCA module. In this paper, it is added before the feature processing layer of the model and combined with the model, so that the feature information processed by the model is the feature processed by the attention mechanism. In this way, the features processed by the model are the enhanced and effective features, and the three-channel attention mechanism module also suppresses the information interference from the complex background in the process, which improves the model performance. The three-channel attention mechanism module, shown in Figure 4, includes a spatial attention module, a channel attention module, and a coordinate attention module, and is a product of the linkage between the three.

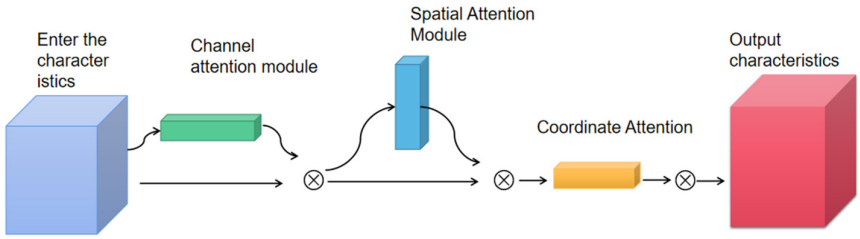


Figure 4. Diagram of the Three-Channel Attention Mechanism Structure.

The channel attention mechanism is an adaptive spatially selective attention module for dynamically learning and adjusting the importance of different channels (feature maps). It helps the model to effectively interact and transfer information between different channels of the feature map to enhance the model's representation of the input data. It mainly realizes the deep information representation of pepper targets in images by weighting the convolutional features of the channels. As shown in Equation (1), the final efficient information representation is obtained through constant weighting, see Figure 5 for details.

 doi:10.20944/preprints202404.1916.v1

$$Mc(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F_{avg}^{c})) + W_1(W_0(F_{max}^{c}))) \quad (1)$$

$$Ms(F) = \sigma(f^{7\times7}([AvgPool(F);MaxPool(F)])) = \sigma(f^{7\times7}([F_{avg}^{s};F_{max}^{s};])) \quad (2)$$

$$Yc(i,j) = Xc(i,j) \times G_c^h(i) \times G_c^h(j) \quad (3)$$

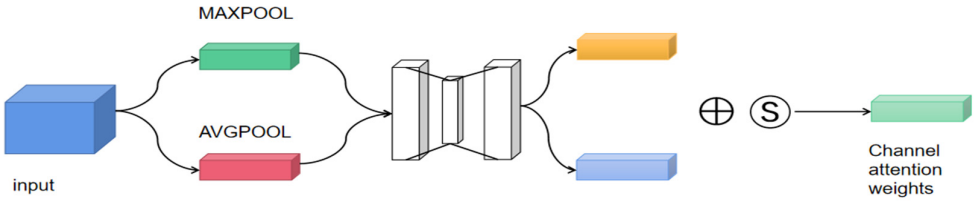


Figure 5. Channel attention module.

The spatial attention mechanism is used to dynamically learn and adjust the importance of different spatial locations. It helps the model to effectively interact and transfer information between different spatial locations of the feature map to enhance the model's representation of the input data. Specifically, spatial attention is a channel compression technique that performs average pooling and maximum pooling in channel dimensions, respectively. As shown in Equation (2) the combination of average pooling and maximum pooling is used to obtain a two-channel feature map to determine the specific location of the chili pepper, and the detailed structure is shown in Figure 6.

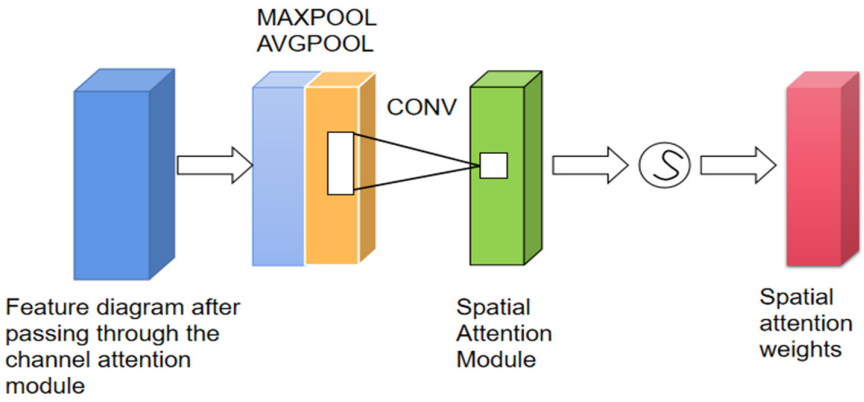


Figure 6. Spatial Attention Module.

Unlike traditional spatial attention, coordinate attention focuses on the absolute coordinate information of each location in the input feature map, not just the features at the spatial location. Therefore, the coordinate attention module can help the model to obtain the absolute coordinate information of the chili peppers, so as to reduce the interference of environmental factors on the model. Coordinate attention with the help of the idea of residual module, in the use of C * H * 1 convolution of the features at the same time using a parallel module to process the feature map, and then through the aggregation to get two independent feature map. As shown in Equation (3) and Figure 7, the final two independent feature maps are multiplied with the input feature map to get the final feature map, thus realizing the absolute expression of coordinate information.

 doi:10.20944/preprints202404.1916.v1

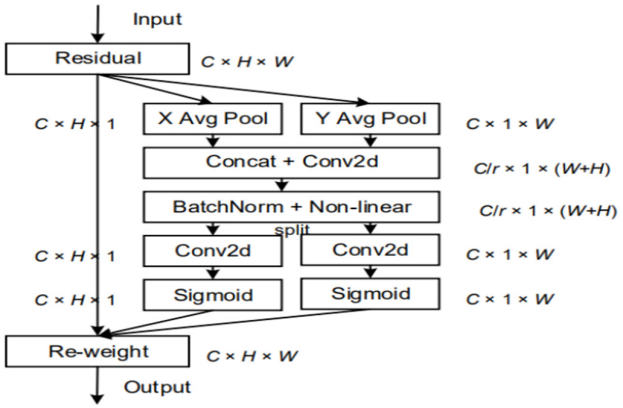


Figure 7. Coordinate Attention Module.

The effective combination of the above modules constitutes a three-channel attention mechanism, which enables the model to effectively capture pepper fruits at different locations when deployed to mobile devices, thus realizing the efficient operation of pepper detection.

2.6. Resolution Adaptive Feature Fusion Network Module

Because data captured using less than the same equipment will be encountered during the pepper detection process, these data will have different resolutions, while image data captured by the same equipment will also consist of different resolutions. In this paper, we find that the images of chili peppers with different resolutions produce different feature maps when input to the model, and therefore do not contribute differently to the model's fusion of different features for prediction. To address the above problems, this paper proposes the resolution adaptive fusion module, which aims to aggregate features of different resolutions. Previous models deal with this kind of problem by adjusting the feature maps of different resolutions to the same resolution and then summing them up. The resolution adaptive fusion module, on the other hand, as shown in Figure 8, adds an additional weight to each input and allows the network to learn the importance of each input feature. And the jump connections from input nodes to output nodes are in the same proportion as they are in the same layer, thus fusing more features without adding much computational cost. In addition, a basic network is constructed using each of the networks composed of top-down and bottom-up and repeated several times to achieve higher level feature fusion.

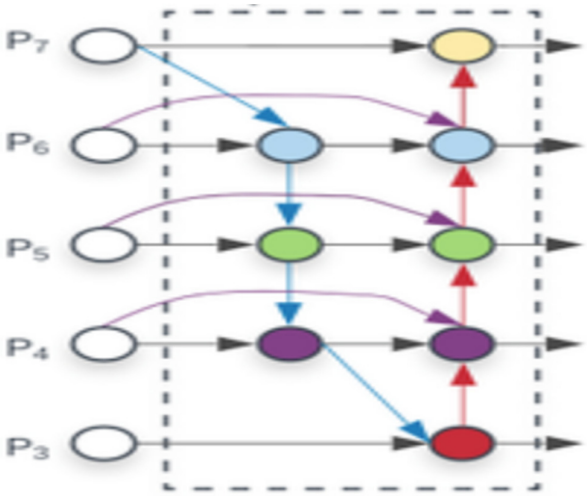


Figure 8. Resolution Adaptive Feature Fusion Network Module.

 doi:10.20944/preprints202404.1916.v1

2.7. Yolo-Chili Network

yolo-chili is shown in Figure 9, which uses yolov5's backbone network to facilitate porting to different devices. yolov5's backbone network is CSPDarknet53, which consists of CBL, BottleneckCSP/C3, and SPP/SPPF. For ease of deployment, yolov5 removes the Focus module. the CBS module consists of a constant combination of Conv+BatchNorm+SiLU used to obtain the depth of the feature map. C3, on the other hand, draws on the residual idea for cross-stage connectivity, which is used to improve the feature transfer efficiency and information utilization. It consists of multiple convolutional layers and residual connections for extracting features from the input image. Compared with CSPDarknet53-tiny in YOLOv4, yolov5 has deeper network structure and stronger feature extraction capability. Meanwhile, for the problem that yolov4 cannot handle multi-scale feature maps better, yolov5 uses FPN for fusion, which improves the model's ability to detect targets of different sizes. In terms of prediction, yolo-chili uses the prediction module of yolov5, but uses ASFF-Detect instead of the original detection layer. It also employs the K-Means algorithm to cluster the anchor frames generated from the dataset, the non-great suppression and confidence threshold filtering to select the prediction frames, and Alpha-IoU instead of CIOU.IoU is computed as the ratio of the area of intersection of the target detection frames (which are usually the frames predicted by the model) with the true labeling frames (Ground Truth) and their concatenation area.The value of IoU ranges from 0 to 1. value ranges from 0 to 1, with larger values indicating a higher degree of overlap between the detected frames and the true labeled frames, and more accurate detection results. α-IoU introduces a parameter α to regulate the calculation of IoU. Specifically, α-IoU is calculated as follows: If the IoU between the detected frame and the real labeled frame is greater than or equal to α, the IoU is directly used as the final evaluation index. If the IoU is less than α, the IoU is multiplied by a factor less than 1 to reduce its influence on the final evaluation index.

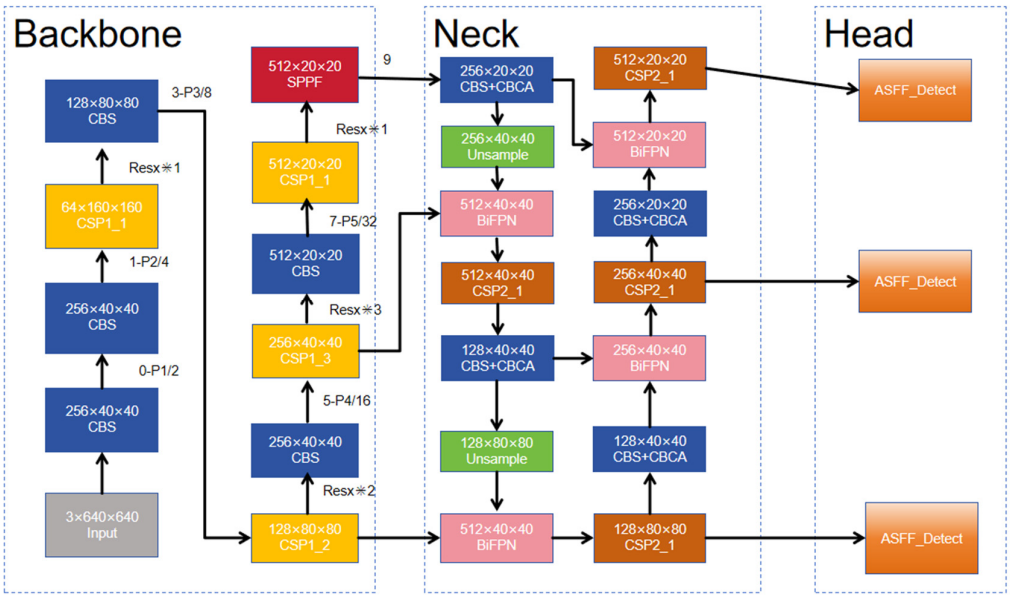


Figure 9. yolo-chili Structure.

3. Results and Discussion

3.1. Parameter Setting

The parameters for the comparison experiments were set as follows: the original size of the image was 640 × 640 pixels, so the input to the model was also adjusted to 640 × 640 × 3. The ratio of the training set to the test set was set to 8:2. The batch size was set to 4, the epoch was set to 100, the initial learning rate was 0.01, the cyclic learning rate was 0.2, and the optimizer used was

 doi:10.20944/preprints202404.1916.v1

SGD(stochastic gradient descent), with a weight decay coefficient of 0.0005, and the iou loss coefficient was set to 0.05.

3.2. Evaluation Indicators

In this study, we use precision, F1 score, accuracy and recall as evaluation metrics to assess the effectiveness of different network models in detection tasks targeting chili pepper images, here Equations (4)–(7) are the formulas for F1, accuracy, precision and recall, respectively.

$$F1=2\times\frac{Precision\times Recall}{(Precision+Recall)} \tag{4}$$

$$Accuracy=\frac{\text{Identify the correct total number of disease and pest images}}{Total\ number\ of\ disease\ and\ pest\ images} \tag{5}$$

$$Precision=\frac{TP}{TP+FP} \tag{6}$$

$$Recall=\frac{TP}{TP+FN} \tag{7}$$

Where TP is the number of positive samples predicted by the classifier with positive true results, i.e., the number of correctly identified positive samples; FP is the number of negative samples predicted by the classifier with negative true results, i.e., the number of incorrectly predicted negative samples; and FN is the number of positive samples predicted by the classifier with negative results but positive true results, i.e., the number of underreported positive samples. Accuracy is the ratio of the total number of correctly recognized pepper images to the total number of pepper images.

3.3. Yolo-Chili Ablation Test Performance Comparison

In this paper, ablation experiments were conducted on the test set using yolo-chili to verify the feasibility of the model optimization strategy based on YOLO-chili. As shown in Table 2, it can be seen that after adding HFFN, three-channel attention mechanism and Resolution Adaptive Feature Fusion Network Module, all the indexes are improved. However, by adding only HFFN, all the performances are decreased, which is due to the problem of positive and negative sample confusion that occurs when yolo-chili performs different levels of feature fusion, and at the same time, the features of the level fusion have a great deal of background noise, which seriously interferes with the model's prediction; therefore, by adding the three-channel attentional mechanism prior to the HFFN, the performance of the model has been significantly improved due to the three channel attention mechanism suppresses the interference of background noise and highlights the fruit features. However, due to the addition of different modules, the computational complexity and memory consumption of the network increased accordingly.

Table 2. ablation test performance comparison.

Yolo-chili	HFFN	Three-channel attention mechanism	Resolution Adaptive Feature Fusion Network Module	AP(Average Precision) (%)	precision(%)	recall (%)
✓				83.24	91.33	81.77
✓		✓		91.39	92.74	91.65
✓	✓			82.32	87.93	81.62

 doi:10.20944/preprints202404.1916.v1

✓			✓	85.54	93.54	82.15
✓	✓		✓	92.24	93.42	91.19
✓	✓	✓		91.27	93.20	91.62
✓	✓	✓	✓	94.11	94.42	92.25

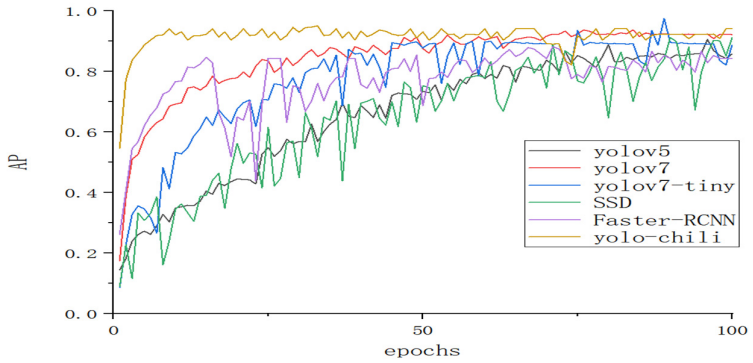
3.4. Comparison of the Performance of Different Object Detection Models.

The comparison between the YOLO-chili model and the currently mainstream object detection models, including Faster-RCNN, SSD, YOLOv7, yolov7-tiny, and YOLOv5, is presented in Table 3. The average precision mean of the YOLO-chili model is respectively 10.48, 2.87, 0.18, 0.49, and 3.09 percentage points higher than the other five models. Among them, the single-stage detection network model SSD has the lowest recognition accuracy, and the two-stage detection model Faster-RCNN has the largest number of parameters, thus leading to slower inference speed. The average precision mean and inference speed of YOLOv7 are improved compared to Faster-RCNN and SSD, but it still cannot meet the requirement of real-time detection for pepper fruits. Although the precision of the YOLO-chili model is only slightly higher than that of yolov7-tiny, the parameter count is much higher than that of yolov7-tiny. While it can meet the requirement for real-time detection of pepper fruits, further optimization is still necessary.

Table 3. Detection results of different target detection algorithms.

Models	Parameters/×106M	FLOPs/G	Model size/MB	AP (%)	precision(%)	recall (%)
Yolov5	7.24	16.6	14.1	85.53	91.33	81.77
Yolov7	37.49	123.5	74.5	92.39	94.24	91.65
Yolov7-tiny	6.51	14.2	12.1	89.32	93.93	87.62
SSD	26.29	62.8	93.3	91.24	93.42	73.19
Faster-RCNN	137.10	370.2	111.5	83.63	67.84	81.62
Yolo-chili	11.4	21.2	18.7	94.11	94.42	92.25

As can be seen in Figure 10, the YOLO-chili model has the fastest fitting speed, while the curve change of Faster-RCNN is obviously unstable, which is due to the efficiency of the yolo series model as a one-stage model itself. The yolo-chili, on the other hand, is due to the possession of efficient computational power and the use of transfer learning to acquire sufficient prior knowledge. At the same time, traditional SSDs may not be able to adapt to the complex and changing environment of the chili dataset and the complexity of the model, thus making it the slowest to train. For the complex environment of the chili pepper dataset yolov7-tiny does not seem to outperform yolo7, but both are using a migration learning approach and therefore both are slower to fit. It can be inferred from the results that the model performance of both YOLOv7 and YOLO-chili is suitable for real-time detection of chili peppers.



 doi:10.20944/preprints202404.1916.v1

Figure 10. Contrast experimental AP curve.

3.5. Reducing Model Size Using Quantitative Pruning

The ultimate goal of this paper is to deploy the real-time detection model to different hardware devices, so lightweighting is a necessary optimization step, so we use the quantitative pruning algorithm to prune the model to reduce the number of model parameters and improve the speed of the model by pruning the channels that account for a lower percentage of importance in the model. First, we use yolo-chili to train the model in the fitted state, and then perform quantitative pruning on the trained model. At the same time, the sparsity of the lower weight layer is trimmed from 0.5 to 0.9 and then the model is quantized and compressed. After that, this paper retrains the yolo-chili model until it converges. This method can effectively reduce the model parameters, model computational complexity, and the size of the weight file while preserving the accuracy of the model. The results are shown in Table 4. The original model parameters are 18.7M, and after quantized training of the model weight file, the pruned model parameters are 9.64M, and the model accuracy reaches 93.66%, which is only 0.45% decrease in accuracy, while the model volume is reduced by half, and the FPS is only 65, which makes yolo-chili fully adaptable to a variety of different mobile devices to accomplish the real-time detection tasks. Although the effect of FPS is reduced, this is acceptable compared to the improvement in detection performance.

Table 4. Quantitative pruning results.

Models	Size/MB	AP	Recall	Precision	FPS
Yolo-chili	18.7	94.11	92.25	94.42	94
pruned_quantized_model	9.64	93.66	0.97	0.97	87

The detection results of yolo-chili are shown in Figure 11, which shows that yolo-chili can effectively identify the location information of chili peppers in complex situations such as multilayered targets, cloudy skies, and occlusion, so this can prove the effectiveness of yolo-chili.

Figure 11. yolo-chili test results.

 doi:10.20944/preprints202404.1916.v1

4. Conclusions

In this paper, we propose yolov5 based pepper target detection algorithm yolo-chili. the initial yolov5 model performs poorly in recognizing small target peppers, dimly lit peppers and clusters of peppers. Therefore, we propose an HFNN to improve the detection for different layers of target peppers. Meanwhile, we added a long-range information extraction module to CBAM and constructed a three-channel attention mechanism network to reduce the effect of complex background on chili pepper detection, so as to improve the detection effect. Finally, the Alpha-IOU loss function is used to replace the original IOU function, and the resolution adaptive feature fusion network module is used to fuse the resolutions of different features, and quantized pruning is used to control the size of the model to ensure the lightweight of the model. The experimental results show that yolo-chili can be fully adapted to the task of picking peppers in real situations, and also has a lightweight detection speed to accomplish real-time detection in real production. In the next step, the research will further use yolo-chili for real-time detection of different peppers to realize the intelligence and modernization of the pepper picking task.

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

Funding: This study was funded by the National Key Research and Development Program (2022YFD2002001) under the title of “Research on key common technologies and system development for intelligent harvesting of special cash crops”.

Data Availability Statement: https://www.kaggle.com/datasets/jingxiche/chili-data

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

 doi:10.20944/preprints202404.1916.v1

References

1. Fu L, Duan J, Zou X, et al. Fast and accurate detection of banana fruits in complex background orchards[J]. IEEE Access, 2020, 8: 196835-196846.
2. Mathew M P, Mahesh T Y. Leaf-based disease detection in bell pepper plant using YOLO v5[J]. Signal, Image and Video Processing, 2022: 1-7.
3. Tian Y, Yang G, Wang Z, et al. Apple detection during different growth stages in orchards using the improved YOLO-V3 model[J]. Computers and electronics in agriculture, 2019, 157: 417-426.
4. Liu T H, Nie X N, Wu J M, et al. Pineapple (Ananas comosus) fruit detection and localization in natural environment based on binocular stereo vision and improved YOLOv3 model[J]. Precision Agriculture, 2023, 24(1): 139-160.
5. Gai R, Chen N, Yuan H. A detection algorithm for cherry fruits based on the improved YOLO-v4 model[J]. Neural Computing and Applications, 2023, 35(19): 13895-13906.
6. Jiang M, Song L, Wang Y, et al. Fusion of the YOLOv4 network model and visual attention mechanism to detect low-quality young apples in a complex environment[J]. Precision Agriculture, 2022: 1-19.
7. Yang G, Wang J, Nie Z, et al. A lightweight YOLOv8 tomato detection algorithm combining feature enhancement and attention[J]. Agronomy, 2023, 13(7): 1824.
8. Tian Y, Wang S, Li E, et al. MD-YOLO: Multi-scale Dense YOLO for small target pest detection[J]. Computers and Electronics in Agriculture, 2023, 213: 108233.
9. Lin Y, Huang Z, Liang Y, et al. AG-YOLO: A Rapid Citrus Fruit Detection Algorithm with Global Context Fusion[J]. Agriculture, 2024, 14(1): 114.
10. Yang S, xing Z, Wang H, et al. Maize-YOLO: a new high-precision and real-time method for maize pest detection[J]. Insects, 2023, 14(3): 278.
11. Zhao Y, Yang Y, Xu X, et al. Precision detection of crop diseases based on improved YOLOv5 model[J]. Frontiers in Plant Science, 2023, 13: 1066835.
12. Karthikeyan M, Subashini T S, Srinivasan R, et al. YOLOAPPLE: Augment Yolov3 deep learning algorithm for apple fruit quality detection[J]. Signal, Image and Video Processing, 2024, 18(1): 119-128.
13. Tang R, Lei Y, Luo B, et al. YOLOv7-Plum: advancing plum fruit detection in natural environments with deep learning[J]. Plants, 2023, 12(15): 2883.
14. Parico A I B, Ahamed T. Real time pear fruit detection and counting using YOLOv4 models and deep SORT[J]. Sensors, 2021, 21(14): 4803.
15. Lawal O M, Huamin Z, Fan Z. Ablation studies on YOLOFruit detection algorithm for fruit harvesting robot using deep learning[C]//IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2021, 922(1): 012001.
16. Li T, Sun M,Ding X, et al. Identification of ripening tomatoes based on YOLO v4+ HSV[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(21).
17. Guo J, Xiao X, Miao J, et al. Design and Experiment of a Visual Detection System for Zanthoxylum-Harvesting Robot Based on Improved YOLOv5 Model[J]. Agriculture, 2023, 13(4): 821.
18. Yang J, Qian Z,Zhang Y, et al. Real-Time Tomato Recognition in Complex Environments with Improved YOLOv4-tiny[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(9).
19. Wang L, Qin M,Lei J, et al. Blueberry Ripeness Recognition Based on Improved YOLOv4-Tiny[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(18).
20. Sun F, Wang Y,Lan P, et al. Apple Fruit Disease Recognition Based on Improved YOLOv5s and Migration Learning[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(11).
21. Ren R, Zhang S, Sun H, et al. Research on pepper external quality detection based on transfer learning integrated with convolutional neural network[J]. Sensors, 2021, 21(16): 5305.
22. Zhou J, Hu W, Zou A, et al. Lightweight detection algorithm of kiwifruit based on improved YOLOX-s[J]. Agriculture, 2022, 12(7): 993.
23. Zhang C, Kang F, Wang Y. An improved apple object detection method based on lightweight YOLOv4 in complex backgrounds[J]. Remote Sensing, 2022, 14(17): 4150.
24. Wang D, He D. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning[J]. Biosystems Engineering, 2021, 210: 271-281.
25. Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129: 1789-1819.
26. Wang F, Jiang J, Chen Y, et al. Rapid detection of Yunnan Xiaomila based on lightweight YOLOv7 algorithm[J]. Frontiers in Plant Science, 2023, 14: 1200144.
27. Fu L, Yang Z, Wu F, et al. YOLO-Banana: a lightweight neural network for rapid detection of banana bunches and stalks in the natural environment[J]. Agronomy, 2022, 12(2): 391.
28. Fang W, Guan F, Yu H, et al. Identification of wormholes in soybean leaves based on multi-feature structure and attention mechanism[J]. Journal of Plant Diseases and Protection, 2023, 130(2): 401-412.
29. Abade A, Ferreira P A, de Barros Vidal F. Plant diseases recognition on images using convolutional neural networks: A systematic review[J]. Computers and Electronics in Agriculture, 2021, 185: 106125.

 doi:10.20944/preprints202404.1916.v1

30. Zeng W, Li M. Crop leaf disease recognition based on Self-Attention convolutional neural network[J]. Computers and Electronics in Agriculture, 2020, 172: 105341.
31. Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern recognition, 2018, 77: 354-377.
32. Yu L, Xiong J, Fang X, et al. A litchi fruit recognition method in a natural environment using RGB-D images[J]. Biosystems Engineering, 2021, 204: 50-63.