applied sciences

MDPI

Review

A Review of Target Recognition Technology for Fruit Picking Robots: From Digital Image Processing to Deep Learning

Xuehui Hua [1], Haoxin Li [2], Jinbin Zeng [2], Chongyang Han [2], Tianci Chen [2], Luxin Tang [3] and Yuanqiang Luo [2,*]

[1] College of Automotive Engineering, Foshan Polytechnic, Foshan 528100, China
[2] College of Engineering, South China Agricultural University, Guangzhou 510642, China
[3] Guangdong Industrial Robot Integration and Application Engineering Technology Research Center, Guangzhou Institute of Technology, Guangzhou 510540, China
* Correspondence: luoyq@scau.edu.cn

Abstract: Machine vision technology has dramatically improved the efficiency, speed, and quality of fruit-picking robots in complex environments. Target recognition technology for fruit is an integral part of the recognition systems of picking robots. The traditional digital image processing technology is a recognition method based on hand-designed features, which makes it difficult to achieve better recognition as it results in dealing with the complex and changing orchard environment. Numerous pieces of literature have shown that extracting special features by training data with deep learning has significant advantages for fruit recognition in complex environments. In addition, to realize fully automated picking, reconstructing fruits in three dimensions is a necessary measure. In this paper, we systematically summarize the research work on target recognition techniques for picking robots in recent years, analyze the technical characteristics of different approaches, and conclude their development history. Finally, the challenges and future development trends of target recognition technology for picking robots are pointed out.

Keywords: fruit; picking robots; deep learning; target detection; image processing

check for updates

Citation: Hua, X.; Li, H.; Zeng, J.; Han, C.; Chen, T.; Tang, L.; Luo, Y. A Review of Target Recognition Technology for Fruit Picking Robots: From Digital Image Processing to Deep Learning. *Appl. Sci.* **2023**, *13*, 4160. https://doi.org/10.3390/app13074160

Academic Editors: Rui Yao and Hancheng Zhu

Received: 22 February 2023
Revised: 17 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, the booming fruit-growing industry has made fruit picking an important production segment that employs more than 60% of the labor force. With the aging population and urbanization further highlighting the labor shortage, a significant increase is seen in the cost of picking [1]. Therefore, intelligent agricultural fruit-picking equipment and picking robots that can improve picking efficiency and reduce picking costs have now become an important research direction. Figure 1 sets out several representative fruit-picking robots.

For picking robots, the recognition ability of the vision system is particularly important in the picking process. Identifying the fruit efficiently and accurately is a prerequisite for completing the picking task [2]. In 1968, Schertz and Brown [3] introduced machine vision to fruit recognition, which enabled the rapid development of fruit-harvesting robots, including the first vision recognition system for apples established by Parrish et al. [4] in 1977. Traditional digital image processing techniques and target recognition techniques based on deep learning make up the majority of the approaches used to identify fruits. These techniques extract features such as color, geometric shape, and texture in images for matching. Such methods have matured after researchers invested years in them. However, the actual working environment of picking robots is much more complex. For example, the change in illumination conditions easily affects the color features of the fruits in images, and the occlusion of the fruits by background objects such as branches and leaves easily affects the geometric features of the fruits in images. This brings problems such as low recognition accuracy, poor model real-time, and low robustness under complex environments. Therefore, automated picking experiments on picking robots pose challenges to

conventional digital image processing techniques. In 2012, the AlexNet network was proposed and won the ImageNet image recognition competition with excellent performance. Since then, deep learning has received a lot of attention from academics and has been extensively used to recognize fruit targets. Altaheri et al. [5] achieved the recognition of dates by combining AlexNet and VGG16 networks, and the recognition accuracy exceeded 90% in the unobstructed cases. The deep-based target recognition method uses a multilayer perceptron structure, where both low-level features and high-level features are analyzed [6] and used to recognize targets. Therefore, compared to traditional digital image processing techniques, this method offers higher recognition accuracy and greater robustness, versatility, and generalization [7].

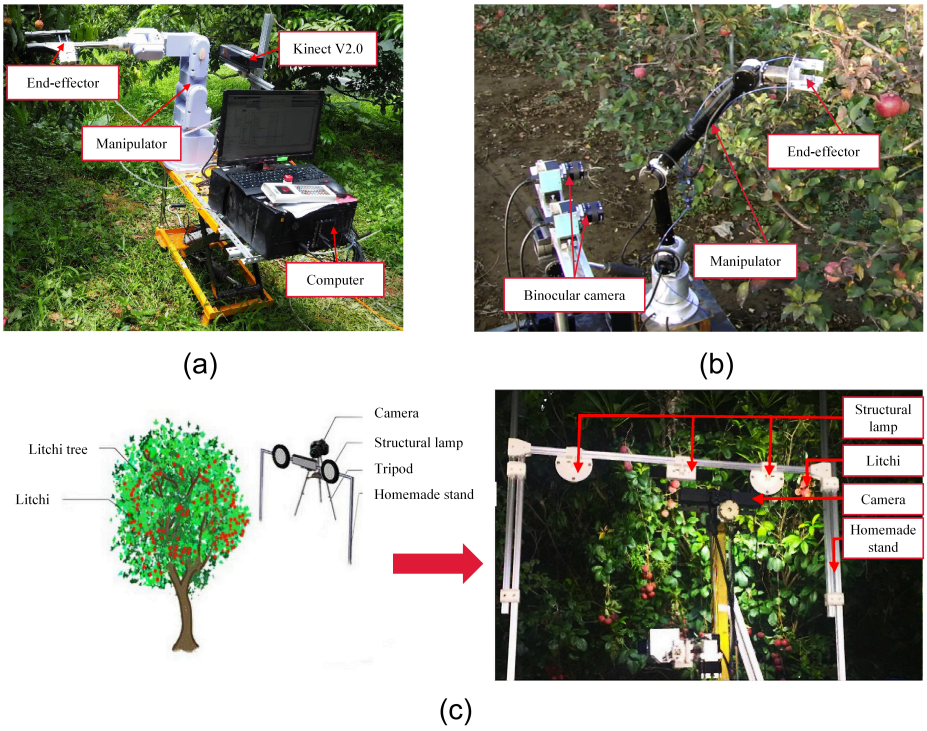

Figure 1. Representative forms of picking robots. (**a**) Litchi-harvesting robot (Adapted with permission from Ref. [8]. 2020, Li et al.); (**b**) Apple-harvesting robot (Adapted with permission from Ref. [9]. 2015, Si et al.); (**c**) Another litchi-harvesting robot (Adapted with permission from Ref. [10]. 2020, Liang et al.).

Although deep learning technology has been widely used in fruit recognition, it still cannot meet the needs of picking robots in the actual working environment, as they cannot fully address the collision between the end-effector and the obstacle objects, such as branches, during the fruit-picking process. Therefore, it is necessary to reconstruct the fruits and branches in 3D to obtain 3D information. In this paper, we summarize the target recognition techniques of picking robots over the past few years by focusing on analyzing and combing through the advantages and disadvantages of different fruit recognition methods. Figure 2 shows the species and quantity of fruit involved in the citation. In addition, this paper summarizes the problems and challenges of target recognition technology for picking machines and proposes the future development trend of this technology for feasibility reference by other researchers.

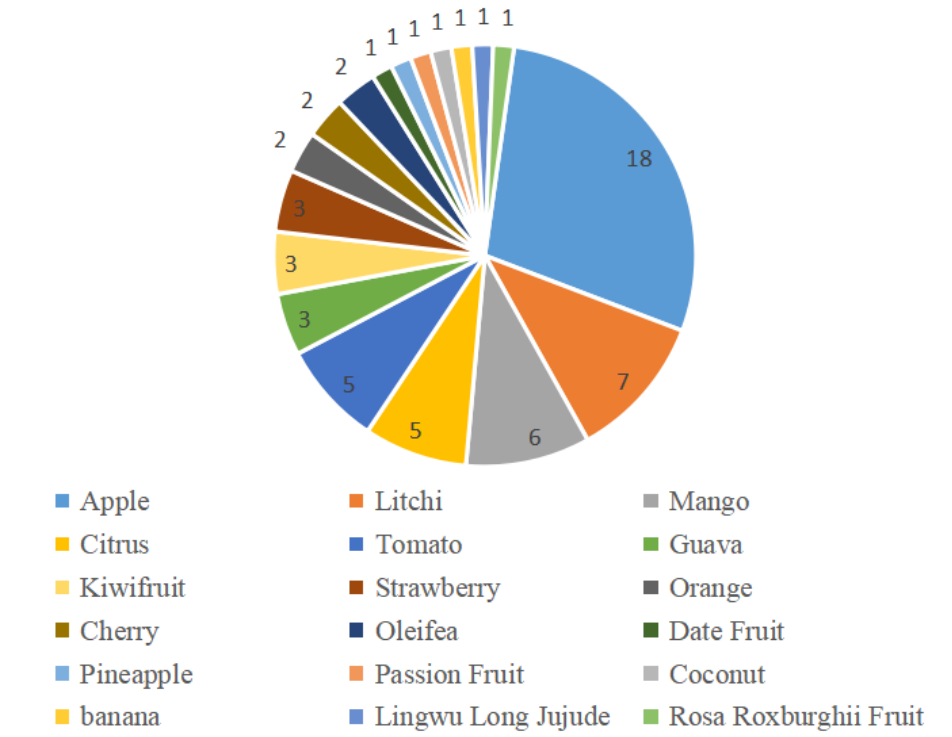

Figure 2. Species and quantity of fruit involved in the citation.

2. Traditional Digital Image Processing Techniques

2.1. Color Feature-Based

The target recognition method is based on color feature segments and recognizes the target fruit in images with background objects such as leaves by predicting the information of pixel points and combining the color space, such as RGB HSI and HSV. The target recognition method based on color features can be used when the target fruit has distinct colors and color features that differ significantly from those of the background objects. For example, Li et al. [11] used color features as recognition features to distinguish pineapples from the background, and the correct recognition rate of the target fruit was 90% during sunny weather and 60% when cloudy. To improve the accuracy of fruit recognition segmentation, Bulanon et al. [12] performed segmentation recognition of Fuji apples by using color difference and optimal thresholding. The chromatic aberration measure in this approach is to enhance the image so that the optimal threshold depends on the maximum grayscale variance of the chromatic aberration. The experiment showed that the overall recognition success rate was higher than 80%, but the recognition based on color features failed to eliminate the influence of changing lighting conditions, and the recognition segmentation algorithm in this paper had an error rate of 18% under backlight conditions. Therefore, when a picking robot uses color feature-based recognition segmentation of target fruits, its recognition accuracy and efficiency will be affected by the type and ripeness of fruits, the complexity of the background, and different lighting conditions.

To further recognize and segment the target fruits with different maturity levels, Zhou et al. [13] used RGB and HSI color spaces on Fuji apples (Figure 3). The difference between the different channels of RGB in the captured image is an important basis for the initial segmentation of the apples. In addition, in order to reduce the influence of light conditions, the threshold of the saturation channel in the HSI image was used to segment the red apples. The system was evaluated using regression coefficients of 0.8 for automatic recognition counts of unripe and ripe apples and 0.85 for manual counts of

apples, but the accuracy of the system was impacted by the occlusion of the target fruit by background objects.

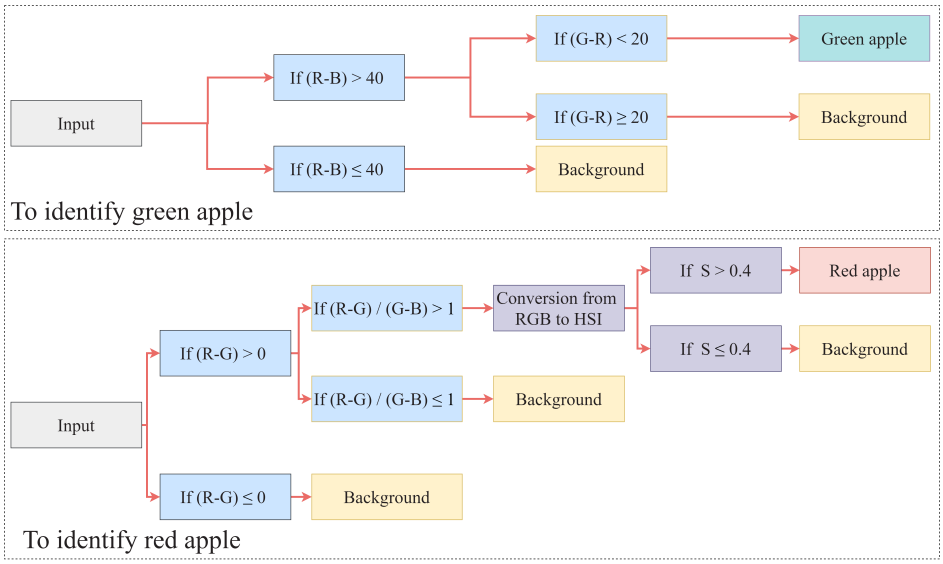

Figure 3. Flowchart of different steps in fruit segmentation. (Adapted with permission from Ref. [13]. 2012, Zhou et al.).

2.2. Geometric Feature-Based

Geometric feature-based recognition provides a new way for picking robots to recognize and segment target fruits, especially when the target fruit is similar to its background in color or when changes in lighting conditions significantly affect its color. Using geometric feature-based recognition can generally achieve better results than color feature-based recognition. Whittaker et al. [14] proposed a system for locating tomatoes in natural environments. The system combines geometric features and global pixels for analysis so that recognization can be successful regardless of the ripeness of the target fruit (in other words, color variation).

Although the recognition segmentation algorithm based on geometric features is not generally affected by lighting conditions, the change in the geometric parameters of the target fruit caused by occlusion becomes the main problem. When leaves, stems, etc., occlude the fruit, the geometry of the fruit in the image is compromised, and the fruit cluster formed by the occlusion between the fruits will affect the recognition effect. To solve this problem, Hannan et al. [15] segmented and identified target oranges in fruit clusters through edge extraction and perimeter-based detection and addressed the problem of changing lighting conditions and fruit clustering. In This study, the overall recognition accuracy of the system is shown to be 90% under changing light conditions and shading through testing, and the overall performance of the picking robot is effectively improved.

2.3. Texture Feature-Based

In general, the surface of the fruit is smoother than the environmental objects such as leaves and stems, so the texture feature becomes an important fruit feature to be separated from the background. Because this analysis process is not sensitive to the color of the target fruit, many researchers will use texture features in picking robots to identify and segment the target fruit. The Gabor texture analysis proposed by Zhang et al. [16] in 2002 is an efficient processing approach. Kurtulmus et al. [17] achieved the identification and segmentation of unripe green citrus by using it and the "Eigenfruit" method, and the recognition accuracy rate was 75.3%. Their paper shows that the background complexity and the size of the target fruit under different conditions affect the recognition accuracy of the system. To improve the accuracy, Chaivivatraku et al. [18] proposed a texture analysis-

based technique for plant green fruit detection and experimentally extracted 24 texture features on pineapples and bitter melons for segmentation and recognition, and the results showed that the detection rate of a single-width of pineapple reached 85%, and the detection rate of a single width of bitter melon reached 100%.

2.4. Multi-Feature Fusion Based

Although the target fruit may be distinguished from the backdrop by a single characteristic, it is not a wise choice. This is because changes in lighting conditions affect the brightness of the target fruit, which in turn, leads to changes in its color. In addition, the occlusion of the target fruit by background objects such as branches and leaves can also compromise the geometric features of the target fruit in the image. In order to further improve the performance of the recognition system of the picking robot, a variety of features can be fused to form part of the recognition algorithm. A recognition method based on color features and texture features was proposed in the literature [19], but accuracy is easily affected by the color characteristics of the target fruit. Therefore, Payne et al. [20] proposed an edge detection filter for this problem to overcome the disadvantage of the main influence position of color features in the algorithm and demonstrated through experiments that the effectiveness and detection effect were significantly improved.

2.5. Comparison and Summary of Traditional Digital Image Processing Techniques

Table 1 compares the research results of traditional digital image processing techniques in some references, from which the following four conclusions can be drawn. (a) The recognition method based on color features, although easily affected by lighting conditions and susceptible to disturbance in natural environments, is more suitable for working in artificially structured environments. (b) Although the recognition method based on geometric features avoids the influence of lighting conditions, the recognition performance is poor in the case of occlusion and overlapping, so it is more vulnerable to interference in the natural environment. (c) Since the fruit surface is relatively smooth, the use of texture features can effectively achieve the recognition function, but the method is easily affected by environmental noise and offers poor performance in complex orchard environments. (d) The feature fusion-based method can effectively avoid the limitations of individual features and improve the performance of the visual recognition system of picking robots, but its performance still cannot meet the requirements of operation in a natural environment. Therefore, it is difficult to extend fruit recognition technology based on traditional digital image processing techniques to practical applications, and a more effective recognition method should be found.

Table 1. Comparison of research results of traditional digital image processing techniques.

Traditional Digital Image Processing Techniques	Application to Crops	Technical Characteristics and Performance Indicators	Limitations	References
Color feature-based	Pineapple, apple	Can significantly segment the fruit from the background, with a combined recognition success rate of 80%	Easily affected by lighting conditions	[11–13]
Geometric feature-based	Tomatoes, oranges	Capable of acquiring fruit outline information, with a combined recognition success rate of 90%	Vulnerable to fruit occlusion, fruit overlap and fruit volume	[14,15]

Table 1. *Cont.*

Traditional Digital Image Processing Techniques	Application to Crops	Technical Characteristics and Performance Indicators	Limitations	References
Texture feature-based	Citrus, pineapple	Bitter melon can significantly segment the fruit from the background, with a combined recognition success rate of 85%	Vulnerable to the growth environment of fruit trees, fruit shading, etc.	[16–18]
Multi-feature fusion based	Oranges, apples, mangoes	Makes up for the shortcomings of a single feature and can more accurately segment the fruit from the background, with a combined recognition success rate of 90%	Although the recognition success is improved by feature fusion, it cannot offset the influence of natural environmental factors, such as lighting conditions	[19,20]

3. Deep Learning-Based Target Recognition Techniques

3.1. Deep Learning-Based Target Detection Techniques

In recent years, many scholars have devoted themselves to the research and application of deep learning. Figure 4 shows the development of target detection algorithms based on deep learning. R-CNN, Fast R-CNN, and Faster R-CNN are typical representatives of two-stage target detection models based on classification. YOLO, SSD and RetinaNet are typical single-stage target detection models based on regression.

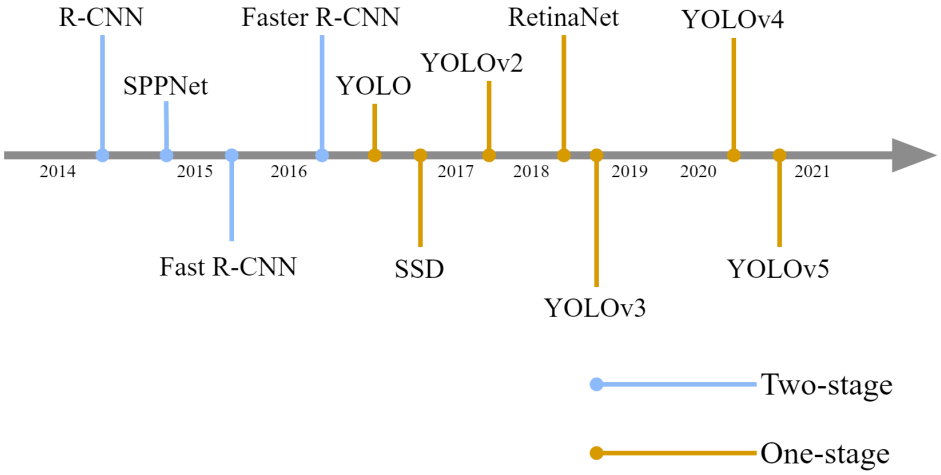

Figure 4. The evolution of deep learning-based target detection algorithms.

3.1.1. Classification-Based Two-Stage Target Detection Techniques

Girshick et al. [21] introduced the R-CNN model inspired by Alexnet [22], which was a great breakthrough in target detection and laid the foundation for the later R-CNN series networks. The R-CNN is based on four steps: first, input an original image; second, generate about 2000 candidate regions based on the input image using the selective search [23] algorithm, and preprocess each candidate region into a uniformly fixed size; third, extract region features using the Alexnet network; and finally, classify the extracted region features using the SVM classification algorithm. Girshick et al. [21] showed through experiments that the mAP on the VOC 2012 dataset was 53.3%. The proposed R-CNN has greatly facilitated the development of target detection techniques, but it requires excessive time to generate candidate regions.

To address the problem that R-CNN only has forward transmission and no shared computing, He et al. [24] proposed a Spatial Pyramid Pooling Network (SPP-Net). SPP-Net realized the shared computation of convolutional features by introducing adaptive size pooling, which greatly reduced the amount of computation and improved the detection speed compared with R-CNN. In 2015, Girshick [25] proposed the Fast R-CNN target detection algorithm. This algorithm enhanced and optimized the detection speed and accuracy by combining R-CNN and SPP-Net. Fast R-CNN optimized SPP-Net in that it enabled adjustable parameters. The feature extraction network and classification algorithm in the R-CNN algorithm were optimized, and VGG16 was used in lieu of the AlexNet network, and the SoftMax classifier is used in lieu of the SVM classifier. Fusing R-CNN and SPP-Net, the Fast R-CNN excellent performance has been widely recognized in a large number of practices. Zhou et al. [26] improved Fast R-CNN for recognition detection of key organs in tomatoes by fusing automatically extracted RGB and grayscale image features. A double convolutional chain Fast R-CNN was proposed, and the highest average precision (AP) of this recognition algorithm was concluded to be 70.33%, 63.99%, and 44.95%, respectively, for the recognition of three key organs of tomatoes.

The main reason why SPP-Net and Fast R-CNN take more time is that more specialized region proposals are generated by the selective search algorithm during operation. Therefore, a new algorithm for generating candidate regions, the Region Proposal Networks (RPN), is proposed. The sliding window adopted by the RPn algorithm effectively solves the problem of producing too many candidate regions and saves computing resources while each candidate region is given a score. To improve the performance of the target detection algorithm, Ren et al. [27] proposed the Faster R-CNN target detection model, which included the feature extraction network, RPN, and Fast R-CNN. The Faster R-CNN consists of multiple steps of candidate region generation, feature extraction, and classification through RPN to achieve end-to-end training, and its detection performance is substantially improved. Many researchers have already applied it to picking robots. For the problem of different ripening times of mangoes, Wang et al. [28] detected mango flower spikes by using the Faster R-CNN model to facilitate the picking robot to select targets for early harvest. However, the experiments showed that the algorithm offered a low mAP value and unsatisfactory detection accuracy. For circumstances where the target fruit is occluded, Gao et al. [29] classified the detection of the fruit on the tree by using the Faster R-CNN model, which helped the picking robot plan the picking path plan as well as avoid damage to its end-effector from the obstacles during the picking process. Taking apples as the research object, the system classified them into four categories and proved the average classification accuracy of 87.9% for the four categories through experiments. To address the problem of changing lighting conditions caused by the changing of time, Song et al. [30] developed a vision recognition system for picking robots that work all day long. By constructing a Faster R-CNN model and using images of kiwifruit collected under different lighting conditions for training, the accuracy of the test was 87.16%, and the model was able to recognize different kinds of fruits with good robustness.

Since the PRN generation candidate region becomes a larger region after mapping from the anchor of the feature map against the original image, the Faster R-CNN algorithm does not work well for small targets. Many researchers have improved and optimized the Faster R-CNN algorithm by addressing this problem. For example, Tu et al. [31] proposed a Multi-scale Faster R-CNN (MS-FRCNN) based on fusing local and global information. By using color feature information and depth information in images captured by RGB-D cameras, it improves the Faster R-CNN model for small target passion fruit. This approach proposes a new measure for the Faster R-CNN model for small target detection difficulty. In addition, it is challenging for Faster R-CNN to successfully detect fruits in the presence of target occlusion and complex backgrounds with different target morphology and sizes. Therefore, some researchers have proposed different solutions to this problem. To address the problems of slow recognition and poor robustness of the recognition system, Fu et al. [32], based on the Faster R-CNN, combined ZFNet and backpropagation to extract

features from the model, together with the gradient random descent technique to speed up the convergence of the model. It was demonstrated through experiments that these improvements adapt better to the natural environment with changing lighting conditions and have better robustness against the subjectivity and limitations of artificial feature selection. To improve the recognition accuracy and reduce the leakage rate in the case of high similarity between target fruit and background, Parvathi et al. [33] proposed a Faster R-CNN improvement method for detecting coconuts. The method first used image enhancement technology to enhance the collected images and then used ResNet-50-optimized Faster R-CNN model for training and finally achieved better detection results. On this basis, Sun et al. [34] employed the K-means clustering algorithm for optimization and used it to detect tomato organs, and the mAP of experimental results was significantly improved.

3.1.2. Classification-Based One-Stage Target Detection Techniques

(1) YOLO series

In 2015, Redmon et al. [35] proposed the first regression-based one-stage target detection algorithm, You Only Look Once (YOLO), based on the GoogLeNet [36] model. Although YOLO has the biggest advantage over the classification-based two-stage target detection algorithm in terms of fast detection speed, it still has shortcomings in other aspects. First, because YOLO generates a limited number of grid cells, and each grid cell can only correspond to one category, YOLO is less effective in detecting dense and small objects. Second, YOLO is less effective in detecting objects when the size of the predicted objects differs greatly from the size of the training set. Subsequently, many scholars have improved the YOLO algorithm and further proposed YOLOv2, YOLOv3, YOLOv4, YOLOv5, and other algorithms.

YOLOv2 [37] addresses three aspects of YOLO: detection accuracy, detection speed, and the number of detected objects, and improves the detection performance of the model by introducing structures such as Batch Normalization (BN) layers [38], anchor boxes, and improving the backbone network. Xiong et al. [39] improved the detection performance of the model by using the YOLOv2 target detection model for target detection of green mangoes by firstly taking images of a single mango species by UAV and, secondly, feeding the images into the YOLOv2 model for training, eventually identifying that the mAP of the training model was 86.4%. However, several experiments confirmed that its detection effect was reduced when there were too many mangoes causing fruit shading and poor lighting conditions. To address this problem, reference [40] proposed an improved YOLOv2-based detection approach for green mangoes by using a densely connected Tiny-YOLO-dense network structure to fuse features. By training this model, the accuracy for green mango detection exceeded 90%, and the improved target detection model offered better detection performance than Faster RCNN, especially in the case of background occlusion or fruit overlapping. However, this approach requires manual annotation of the foreground regions of the samples that are obscured or overlapped. Although YOLOv2 offers various improvements over YOLO, the model cannot perform multi-scale prediction.

YOLOv3 [41] achieved multi-scale prediction by introducing Feature Pyramid Networks (FPN) [42]. Inspired by ResNet [43], the DarkNet53 [41] network is designed as the backbone network based on DarkNet19. Because of the excellent target detection performance of YOLOv3, many researchers have applied it to picking robots and employed it for fruit detection. To improve the efficiency of picking robots, Liang et al. [10] used the YOLOv3 model to detect litchi fruit during the night and tested the YOLOv3 model by using YOLOv3, SSD, and Faster R-CNN for images with high luminance, normal luminance, and low luminance, respectively. The YOLOv3 model yielded the best detection performance. Its lowest AP value was 89.3% in the low luminance with an average detection time of 0.026 s. Due to the complex growth environment of fruits, there are often cases where the background objects occluded fruits, or fruits overlap, or the fruits and background objects are highly similar. This further complicates the task of picking robots. Therefore, it is vital to improve the detection of target fruits by picking robots in complex environ-

ments. To address the problems of fruit occlusion by branches and leaves, fruit overlapping, and features that change with the growth cycle and thus affect the detection performance of the recognition system, Tian et al. [44] proposed an improved YOLOv3 model for apple detection. The improved YOLOv3 model is used to replace the low-resolution layers in the original YOLOv3 model by using the DenseNet model, which effectively reduces the loss of feature information during the propagation of feature maps, mitigates gradient disappearance, enhances feature propagation, and promotes feature multiplexing and fusion. The tests showed that YOLOv3-DenseNet offered good detection performance for occluded and overlapped apples. Based on this, Liu et al. [45] proposed a YOLO-Tomato model for detecting tomatoes in complex environments. The YOLO-Tomato model not only adopted the DenseNet model to promote feature reuse but also applied the circular bounding box (C-Bbox). The C-Bbox offers two advantages: first, the shape of tomatoes matches the C-Bbox more accurately. Secondly, the C-Bbox has fewer parameters, which reduces the complexity of the calculation. The experiments showed that YOLO-Tomato was able to reduce the effect of changing lighting conditions and offered better robustness.

The YOLOv3-Tiny model is the lightweight version of the YOLOv3 model. Based on the YOLOv3 model, the YOLOv3-Tiny model streamlines the backbone network and other features for better real-time performance. To enable recognition systems of the picking robot to be mounted on small hardware devices, Fu et al. [46] proposed a high-accuracy and high-speed deep YOLOv3-tiny network (DY3TNet) target detection model that can be used for detecting kiwifruit in orchards under all-weather conditions. The DY3TNet model is based on the YOLOv3-tiny model and improved by using multiple 1×1 convolutional layers to reduce the operation load and increase the detection speed. The DY3TNet target detection model was used for the detection experiments of kiwifruit images under different lighting conditions, and the AP was 90.05% under changing light conditions, offering good detection performance. In addition, Xu et al. [47] proposed a Light-YOLOv3 model based on the improved YOLOv3 for green mango detection in complex environments. The model not only optimizes the residual network and NMS to different degrees but also incorporates the MSCA unit. These improvements make the computational speed of the Light-YOLOv3 model on the embedded platform five times faster than the original YOLOv3 model and also demonstrate good performance under different shaded lighting conditions, which provides technical support for improving the efficiency of mango-picking robots.

YOLOv4 [48] offered four improvements over YOLOv3: first, it introduced Mosaic [49] data to enhance operations; second, it fused Darknet53 with CSPNet [50] and used it as a backbone network; third, it introduced the SPP network to expand the perceptual field; fourth, it used FPN+Path Aggregation Network (PAN) [51] to enhance feature fusion; and fifth, CIoU_loss was used as the loss function at the prediction end. These improvements significantly enhanced the detection performance compared to the YOLOv3 model. In reference [52], by using YOLOv3 and YOLOv4 to detect bananas under different lighting conditions, YOLOv4 was proven better than YOLOv3 in terms of detection accuracy and speed. Some researchers improved the YOLOv4 model by combining residual neural networks, densely connected networks, and attention mechanisms. Zheng et al. [53] proposed an RC-YOLOv4 model, which fused residual neural networks for tomato detection. The RC-YOLOv4 model fused ResNet networks in the backbone network of the original YOLOv4 model to construct a new backbone network, the R-CSPDarknet53. The C-SPP network replaced the original SPP network to reduce the loss of feature information. The experimental results showed that this improved approach greatly enhanced the success rate of the RC-YOLOv4 model for detecting tomatoes. Gai et al. [54] proposed a YOLOV4-DenseNet model incorporating a densely connected network for cherry detection to address small target fruits. The YOLOV4-DenseNet model uses DenseNet in lieu of the CSPDarknet53 utilized in the original YOLOV4 and adopts Leaky ReLU as the loss function to promote feature reuse and fusion to improve the network's detection performance for cherries. Through experiments using images of cherries, the detection performance of the

YOLOV4-DenseNet model is proven to be better, which can realize intelligent picking and improve the efficiency of picking robots.

To make the model easier to deploy in small embedded devices, the YOLOv4 model can be simplified using the channel pruning technology. This approach is to simplify the model by identifying and distinguishing the channels of the network, removing less important channels, and retaining important channels, thereby reducing the parameters to be stored by the model. Wu et al. [55] proposed an approach to improve the YOLOv4 model in combination with the channel pruning algorithm for the detection of apple blossoms. The improved model achieved excellent results in terms of being lightweight, especially in that the number of parameters was reduced by 96.74% while the accuracy was maintained. In addition, as a lightweight version of the YOLOv4 model, the YOLOv4-tiny model has a simpler structure and fewer parameters, making it more suitable for deployment in small mobile terminals. Tang et al. [56] proposed a YOLO-Oleifera model based on the improved YOLOv4-tin model for the detection of oil-seed camellia fruit. The YOLO-Oleifera model adopts two measures to optimize the clustering algorithm and adds a small-scale convolution kernel to optimize the YOLOv4-tin model, and the optimized model eliminates the shortcomings that fall into local optimization and enhances the operation speed. In order to meet the real-time requirements of the picking robot, the detection speed of the model is further improved. Zhang et al. [57] proposed an RTSD-Net model for the detection of strawberries by reducing the convolutional layer of the YOLOv4-tin model backbone network. Experiments showed that the RTSD-Net model has the advantage of lightweight technical specifications without significantly compromising accuracy.

YOLOv5 implements the same Mosaic data enhancement on the input side and introduces adaptive anchor frame calculation to facilitate the identification of targets in different sizes. In addition, YOLOv5 introduces the Focus module for image slicing operation and designs two sets of CSPNet, which are applied in Backbone and Neck, respectively. On the output side, YOLOv5 adopts GIoU_loss as the loss function of the Bounding box. Comparing the detection performance of YOLOv5 and YOLOv4, there is not much difference between them in terms of accuracy, except that YOLOv5 performs better in terms of detection speed and is preferred by many researchers and also widely used in fruit recognition detection. In the literature [58], the YOLOv5 model can be used to detect litchi fruits and detect the main stem of picking, effectively improving the picking efficiency of litchi-picking robots. To improve the intelligence of picking robots, Cheng et al. [59] judged whether the fruit was pickable by ripeness, achieved the recognition detection of citrus by using the YOLOv5 model and then combined the RGB image information and the 4-pass ResNet34 network to distinguish the ripeness of the target fruit, and discovered through experiments that the accuracy of the modified approach was 95.07%. In addition, migration learning is an efficient way to improve detection speed. Wang et al. [60] proposed a YOLOv5s-based detection model for apple calyx by using migration learning and channel pruning. The improved YOLOv5s model reduced the model parameters and weight volume by about 71% and substantially improved the computational speed to obtain a detection accuracy of 93.89%. In addition, target fruit detection in complex environments is a significant challenge for picking robots. How to further improve the YOLOv5 model to adapt it to the actual working environment of picking robots is an important research direction. Lyu et al. [61] proposed a YOLOv5-CS model for the detection of green citrus in natural environments. The YOLOv5-CS model is improved in that it optimizes data augmentation and loss function while incorporating the cbam attention mechanism. Yan et al. [62] proposed an improved YOLOv5s-based lightweight detection model for apple detection to address the problem that occluded fruits might prevent the picking robot from performing its task. The improved YOLOv5s model improves the detection accuracy of the model by enhancing the CSP module and the initial anchor frame size of the backbone network and proposes to incorporate the SE module into the backbone network.

It is shown experimentally that the improved YOLOv5s model compresses the volume by 9.29%, while the mAP value improves it by 5.05%.

(2) SSD model

Liu et al. [63] proposed a Single Shot Multi-box Detector (SSD) target detection model, whose core idea is to implement a feature mapping by designing a small convolutional filter and generating and predicting the category and offset of the bounding box. The SSD model uses multi-scale prediction, using a large-scale feature map to detect small objects and a small-scale feature map. The SSD model increases the detection speed without compromising the detection accuracy compared to the classification-based two-stage target detection model, which basically achieves a balance between the two, and many researchers have applied it to fruit target detection. Peng et al. [64] used the ResNet-101 model to optimize the backbone network, increased the depth of the feature extraction network, and optimized the SSD model by migration learning and stochastic gradient descent techniques. The system also achieves target detection of litchi, apple, navel orange, and emperor citrus by yielding an mAP of 89.53%, which is higher than that of the original SSD network. The detection effect of different kinds of fruits in the experiment is not substantially different, which reflects that the system has high robustness. Wang et al. [65] proposed a more lightweight and improved SSD model for the detection of Lingwu jujube in the complex operating environment of the Lingwu jujube picking robot. Through experiments, it was shown that the mAP of the improved SSD model was 96.6% for detecting the target of Lingwu jujube, which further promoted the development of intelligent picking technology for Lingwu jujube.

The SSD detection model uses feature pyramids to detect targets at different scales, but the method makes it difficult to fuse features at different scales, and the feature map characterization ability of small targets overlaps, so it offers poor results in detecting small targets. The Feature Fusion Single Shot Multibox Detector (FSSD) [66] is a feature fusion approach used to solve the problem of difficult detection of small targets. On this basis, in reference [67], multi-output optimization was used for the convolutional layer to obtain a recognition model for a small target, and the model was applied to the detection of small litchi in images taken by UAVs, proving its effectiveness.

(3) RetinaNet

In 2007, Lin et al. [68] proposed the RetinaNet model, which was an effective one-stage detection algorithm. The RetinaNet model utilized the FPN neck network for feature fusion, the RestNet network as the backbone network for feature extraction, and the FCN for the completion of regression and prediction. The initial stage of the regression classification procedure had a sample imbalance issue since the model did not create candidate regions. The RetinaNet model's prediction process used the Focal Loss Loss Function, which, by adjusting the weights of the classification of various samples, successfully worked out the issue of an imbalance in the amount of positive and negative samples during the model's training process. Facing the complex environment in the actual orchard, it identified apples [69–71], Rosa roxburghii fruit [72], and camellia oleifera fruit [73] in a real orchard, RestNet performed better. To improve RetinaNet's ability to detect apples in complex environments, Sun et al. [70] enhanced its backbone network, neck network, and loss function by optimizing each of them. These optimizations increased the feature extraction ability and rate of convergence, and the test indicated a 5.02% improvement in mAP. In addition, Yan et al. [72] suggested various improvement plans for the RetinaNet model and used the enhanced model to recognize Rosa roxburghii tratt. By adopting the K-means + + clustering algorithm, the technique was provided to optimize the Anchor scale as well as the bias in the focal loss function. The optimized model recognized six types of prickly pear images, yielding an average recognition rate of 94.86%, which was 1.8% higher than the original model.

3.1.3. Comparison and Summary of Deep Learning-Based Target Detection Techniques

Figures 5 and 6 show the architecture of representative algorithms in the classification-based two-stage target detection algorithm and the regression-based single-stage target detection algorithm, respectively. In Figure 5, the main trends of the two-stage target detection algorithms are: (a) using parameter sharing to improve the operation speed; (b) proposing new training strategies for end-to-end training. In Figure 6, the main trends of the one-stage target detection algorithms are: (a) proposing networks with stronger feature extraction capabilities to improve the detection performance of the model; (b) compressing the size of the model by a series of simplifications and reducing the number of parameters to improve the model's computational speed.

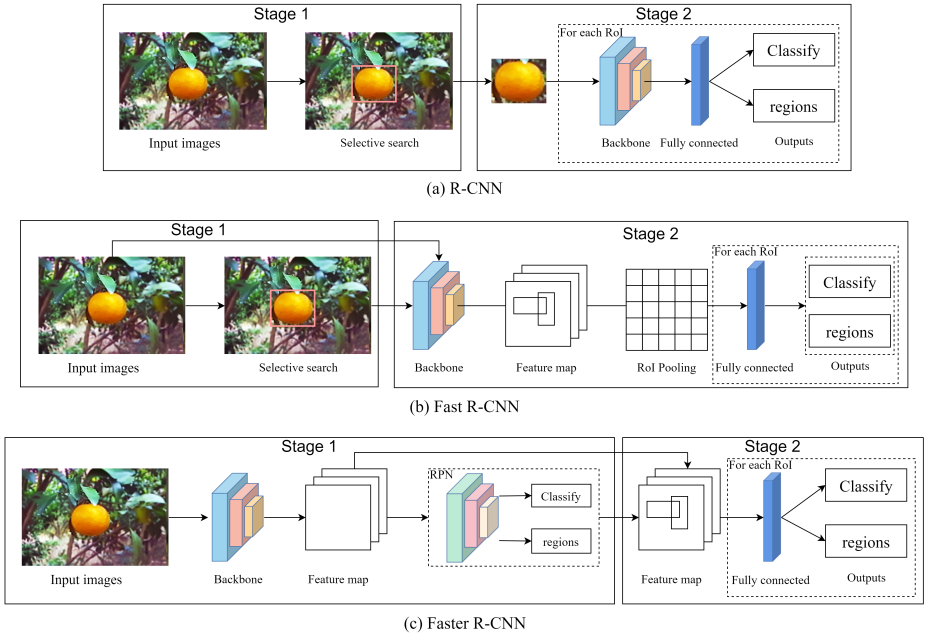

Figure 5. Classification-based two-stage target detection algorithms.

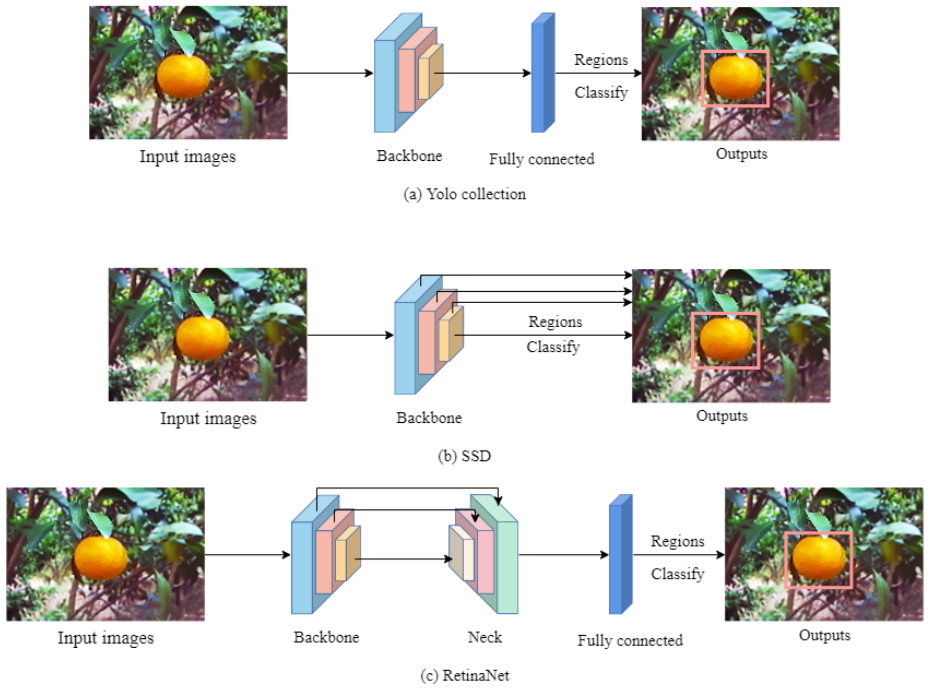

Figure 6. Regression-based one-stage target detection algorithms.

A comparison of the research results of deep learning-based target detection methods is shown in Table 2. Through comparative analysis, two conclusions can be drawn: (a) deep-learning-based target detection methods improve the detection performance in varying degrees compared with traditional digital image processing techniques and are more suitable for complex natural environments; (b) although deep learning-based target detection methods offer strong feature extraction capability, they are subject to environmental interference and thus require a large number of data sets for training, and the training time is long.

Table 2. Comparison of research results of deep-learning-based target detection methods.

Model	Research Object	Technical Characteristics and Performance Indicators	References
Faster R-CNN	Mango flower spikes	This study used Faster R-CNN to identify mango spikes but with a low accuracy of 66% and a recall rate of 48%.	[28]
	Apple	This study used VGG16 to optimize Faster R-CNN, which had good robustness to blocked apples. Its mAP was 87.9%.	[29]
	Kiwifruit	This study used VGG16 to optimize Faster R-CNN, which showed good robustness for fruits under different lighting conditions, and its mAP was 87.61%.	[30]
	Passion fruit	This study enhanced Faster R-CNN's ability to identify small targets by integrating global and local features, with an accuracy of 93.1% and a recall rate of 96.2%.	[31]
	Kiwifruit	This study used ZFNet, backpropagation and stochastic gradient descent techniques to improve Faster R-CNN to realize kiwifruit image detection, with an AP of 89.3%.	[32]
	Coconut	This study utilized ResNet-50 to optimize Faster R-CNN, which showed good robustness for complex environments. Its mAP was 89.4%.	[33]
	Tomato	This study used ResNet-50 and K-means clustering algorithms to optimize Faster R-CNN and achieved the detection of tomato organs. Its mAP was 90.7%.	[34]
YOLO	Green mango	This study used YOLOv2 to identify green mangoes, and its AP was 86.43%. However, the recognition effect would decrease in complex environments if the fruit was blocked.	[39]
	Green mango	This study used Tiny-YOLO to optimize YOLOv2, which improved the recognition performance in fruit overlapping scenarios, and the accuracy rate was 97.02%.	[40]
	Litchi	This study adopted litchi identification at different brightness of YOLOv3, and its mAP was 96.43%.	[10]
	Apple	This study used DensNet to optimize YOLOv3 to improve the recognition accuracy of fruit occlusion scenarios.	[44]
	Tomato	This study used DenseNet and C-Bbox to optimize YOLOv3, showing good robustness under light conditions and divergence under different obscuration conditions.	[45]
	Kiwifruit	This study used different convolution kernels to optimize YOLOv3-Tiny, reducing the volume of the model and yielding an AP of 90.05%.	[46]
	Green mango	This study was a lightweight green mango recognition model designed based on YOLOv3, with an F1 of 97.7% and a volume of 44 MB.	[47]
	Banana	This study used YOLOv4 to identify banana skewers. Its AP was 99.55%, and the average detection time for an image was 44.96 ms.	[52]
	Tomato	This study used ResNet-CSPDarknet53 to optimize YOLOv4, achieving tomato detection. Its accuracy and recall rate were 88% and 89%.	[53]
	Cherry	This study used DenseNet to optimize YOLOv4, which improved the recognition performance of the vision system for small targets, and its mAP was increased by 15% compared with YOLO	[54]
	Apple flowers	This study optimized YOLOv4 using a channel-trimming technique. Its parameter volume was reduced by 96.74%, the volume was 12.46 MB, and the mAP was 97.31%.	[55]

Table 2. *Cont.*

Model	Research Object	Technical Characteristics and Performance Indicators	References
	Oil-seed camellia fruit	This study used the convolutional nucleus of different scales to optimize YOLOv4-Tiny. It showed good stability under lighting conditions, and the recognition performance decreased under severe occlusion conditions.	[56]
	Strawberry	This study used CSPNet to optimize YOLOv4-Tiny. Its detection speed was increased by 25.93%, the accuracy was reduced by only 0.62%, and it was deployed in embedded systems with good performance.	[57]
	Litchi	The detection of litchi was achieved by using YOLOv5, with an mAP of 79.6% and a recall rate of 75.25%	[58]
	Citrus	This study identified and judged the maturity of citrus by combining ResNet34 and YOLOv5, with an accuracy rate of 95.07%	[59]
	Apple calyx	This study optimized YOLOv5s through channel pruning and other techniques. The optimized model physical examination had been reduced by 71%, while mAP had only been reduced by 1.57%.	[60]
	Green citrus	CBAM-optimized YOLOv5 was used in this study to realize the identification and detection of green citrus. Its mAP was 98.23%, and the recall rate was 97.66%.	[61]
	Apple	This study adopted SE optimization YOLOv5s to realize apple identification under different occlusion conditions. Its mAP was 86.75%, and the recall rate was 91.48%.	[62]
SSD	Apple, litchi, navel orange, citrus	This study used ResNet-101, migratory learning and random gradient descent garlic vendor to optimize SSD. Its mAP for the four fruits was 89.53% .	[64]
	Lingwu long jujube	In this study, DenseNet was used to optimize SSD and realize the identification of fruits, with an mAP of 96.6% and a detection speed of 28.05 fps/s	[65]
	Litchi	This study was based on FSSD, optimizes the feature extraction network and realized the detection of litchi. Its AP was 55.79%, and there were omissions and missed detection.	[67]
RetinaNet	Apple, camellia oleifera fruit	They used the RetinaNet model for apple and camellia oleifera fruit recognition, and their APs were 83.1% and 87.9%, respectively.	[69,73]
	Apple	This study was based on RetinaNet, using the Res2Net module to optimize the backbone, BiFPN to optimize the neck, and EIoU Loss as the loss function to achieve the recognition of apples, and its AP was improved by 5.02% compared with that before optimization.	[70]
	Rosa roxburghii tratt	This study was based on RetinaNe, optimizing the loss function and using the K-means + + clustering algorithm to optimize Anchor. It has an AP of 94.86%, improving the original model by 1.8%.	[72]

3.2. Deep Learning-Based Target Segmentation Techniques

3.2.1. Semantic Segmentation Techniques Based on Deep Learning

The deep-learning-based semantic segmentation technique classifies each pixel of the image by a deep learning technique, which can identify the spatial information and basic shape of the target. The traditional CNN network used for segmentation tasks is problematic for its large computation, low computational efficiency, and unsatisfactory segmentation effect. To address such problems, Shelhamer et al. [74] proposed a Full Convolutional Network (FCN) compatible with images of arbitrary size. To input images with arbitrary sizes, the FCN uses the skip layer method to combine the feature maps of each convolutional layer and then employs bilinear interpolation to achieve upsampling to obtain a more delicate segmentation effect. Lin et al. [75] used the FCN algorithm to segment the images of guava and estimated fruit pose using the center position of fruit and

tree branches. However, the stems of the guava were poorly segmented, which meant it was difficult for the picking robot to achieve collision-free picking based on the segmentation results.

Although FCN has greatly promoted the development of image segmentation technology, two problems remain prominent: first of all, some information will be lost after the image is downsampled; second, the utilization of local and global features in the segmentation process is unbalanced. To address these problems, Chen et al. [76] proposed a DeepLab semantic segmentation model to optimize the rough segmentation results with boundaries by constructing a CRF model. Subsequently, Chen et al. successively proposed improved DeepLab models such as DeepLabv2X [77], DeepLabv3 [78], and DeepLabv3+ [79]. The DeepLabv2 model introduced Atrous Spatial Pyramid Pooling (ASPP) to achieve multi-scale segmentation. Li et al. [8] used the DeepLabV3 model to segment the RGB-D camera-acquired litchi images into three categories: background, fruit, and branches, and segmented and localized multiple litchi clusters in a complex environment, which provided a real-time harvesting solution for picking robots to technically support real-time picking. In addition, Peng et al. [80] took a different approach to segmenting lychee images for recognition. The approach uses DeepLabV+ as the base model, avoids information loss by optimizing the feature extraction network, and enhances the computing speed by using migration learning, data enhancement, and coding and decoding structures. It is shown experimentally that the improved DeepLabV3+ semantic segmentation model has a better segmentation effect on litchi-growing branches with 76.5% MIoU.

Badrinarayanan [81] proposed a SegNet semantic segmentation model, which offers more applications for picking robots. The operation is depicted in Figure 7. The SegNet model is improved for the upsampling process of the FCN model by storing Max-pooling indices (Max-Pooling indices) for recovering image information during the upsampling decoding before downsampling. Majeed et al. [82] developed a backbone and branch segmentation method using the Kinect V2 sensor and the SegNet semantic segmentation model. The accurate segmentation of picking points is crucial in the process of fruit picking by picking robots, but the SegNet semantic segmentation model is problematic in that the segmentation of picking points is not clear enough or might be wrong in the process of image segmentation.

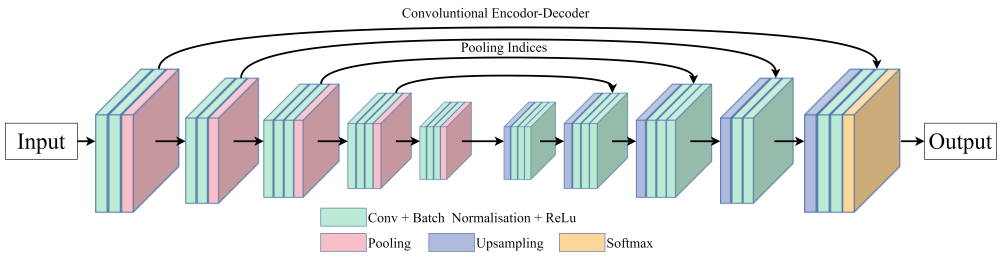

Figure 7. Diagram of SegNet architecture.

3.2.2. Deep Learning-Based Instance Segmentation Techniques

Instance segmentation provides different labels for separate instances of the same class of objects, so the image information provided by instance segmentation is more detailed than that offered by semantic segmentation. The most representative model for segmentation, for instance, is the Mask R-CNN [83] instance segmentation model, which consists of a backbone network ResNet-FPN [42,43], a region proposal network RPN and three branches. The operation process is depicted in Figure 8. The ResNet-FPN extracts the features of the input image. The RPN generates the RoI (Region of Interest), which may contain the detected target based on the features of the image. Finally, the FC and FCN in the three branches perform the target classification and instance segmentation.

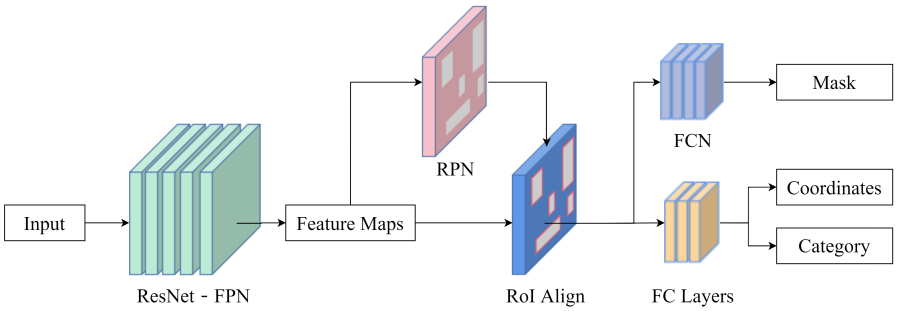

Figure 8. Diagram of Mask R-CNN architecture.

By comparing the detection accuracy of different residual networks as backbone networks, Yu et al. [84] selected ResNet-50 to replace the ResNet-101 network of the original Mask R-CNN instance segmentation model to obtain more features of the image and improve the performance of the visual recognition system of the picking robot. The experiments showed that the improved Mask R-CNN model had a large improvement in detection speed compared with the previous one, and the MIoU of instance segmentation was 89.85%, which offered better generality and robustness in complex environments such as different illumination conditions, occlusion, and fruit overlapping. To further improve the target fruit segmentation accuracy and recognition speed, Jia et al. [85] proposed an improved Mask R-CNN instance segmentation model by combining ResNet and DenseNet and used it for image instance segmentation of apples. The improved Mask R-CNN model not only deepens the depth of the model but also preserves more features of the target fruit in the image and reduces the number of parameters, and the recognition accuracy and recall of the model are 97.31% and 95.70%, respectively, as shown by experiments. However, Wang et al. [86] took a different improvement approach to recognize apples in complex environments by using the ResNet-50 network and FPN instead of the backbone network of the original Mask R-CNN model and also introduced an attention mechanism by using transformer attention. The improved Mask R-CNN model is more accurate. The mAP and recall rates of instance segmentation were 91.7% and 97.1% under different lighting conditions, occlusion, and fruit overlapping. In addition, in the process of recognition, the spatial constraint relationship between the target fruit and the picked stem is fully utilized to obtain more accurate information about the fruit location. To this end, Xu et al. [87] used RGB image and depth image data fusion at the input side while optimizing the RPN of the Mask R-CNN model. The experiments showed that the optimized model had an accuracy of 93.76% and 89.34% for the segmentation of fruits and stems, respectively, and the recognition time of a single image was 0.04 s.

3.2.3. Comparison and Summary of Deep-Learning-Based Target Segmentation Techniques

Table 3 compares the research results of target segmentation methods based on deep learning. Through the summary comparison of the literature, the following two conclusions are drawn: (a) target segmentation methods based on deep learning can better segment the target fruit as well as the fruit branches, which provides more detailed spatial information and facilitates the picking robots to automatically pick fruits without causing any damage to them; (b) both types of image segmentation algorithms require a large number of datasets for training, and the datasets are time-consuming and labor-intensive in the labeling process; (c) semantic segmentation has lower accuracy for fruit recognition in complex environments, while instance segmentation has a longer computing time and has difficulty meeting real-time requirements.

Table 3. Comparison of research results of deep-learning-based target segmentation methods.

Model	Research Object	Technical Characteristics and Performance Indicators	References
FCN	Guava	In this study, FCN was used to achieve image segmentation of guava, and the average accuracy was 89.3%.	[75]
DeepLabV3	Litchi	In this study, DeepLabV3 was used to segment images into backgrounds, fruits and branches, and the detection accuracy of lychee branches was 83.33%.	[8]
DeepLabV3+	Litchi	This study used Xception-65 to optimize DeepLabV3+ and realized the segmentation of litchi growing branches, with an MIoU of 76.5%.	[80]
SegNet	Apple	This study used SegNet to segment the trunks and branches of apple trees in the image, with segmentation accuracy of 92% and 93% and MIoU of 59% and 44%, respectively.	[82]
Mask R-CNN	Strawberry	In this study, Mask R-CNN was used to achieve the segmentation of strawberries in the natural environment, with an AP of 95.78% and a recall rate of 95.41%	[84]
	Apple	This study used ResNet and DenseNet to optimize Mask R-CNN, which realized apple segmentation. Its accuracy rate was 97.31%, and the recall rate was 95.70%.	[85]
	Apple	This study optimized Mask R-CNN using the ResNet of the fusion attention mechanism and characteristic pyramid network. For apple segmentation under different lighting conditions and occlusion conditions, its mAP and recall rates were 91.7% and 97.1%, respectively.	[86]
	Cherry tomato	In this study, the segmentation of cherry tomatoes was realized by optimizing Mask R-CNN, with a segmentation accuracy of 93.76% and 89.34%, respectively.	[87]

3.3. Vision-Based 3D Reconstruction Technology of Fruit

The fruit target recognition method based on 3D reconstruction constructs the spatial coordinates and spatial posture of the fruit using the 3D information obtained by a set of sensors and then guides the picking robot to move to the specified position and adjusts the picking posture. For picking robots, a camera is usually used to acquire images, while machine vision technology is used to obtain fruit 3D information for 3D reconstruction. The process of implementing the vision-based 3D reconstruction technology is shown in Figure 9. The first step is to calibrate the vision system and establish a geometric model of the fruit's geometric position in space corresponding to the information in its image. Secondly, the vision system is used to acquire the image and extract features from the image, and again, stereo matching is performed based on the extracted features. Finally, 3D scene reconstruction is performed based on the geometric model and the results of stereo matching. Vision-based 3D reconstruction technology often uses stereo-vision-based 3D reconstruction technology and RGB-D vision sensor-based 3D reconstruction technology.

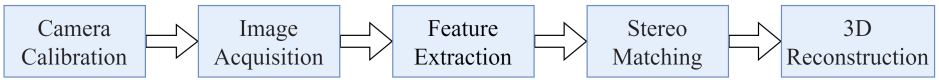

Figure 9. Flow chart of vision-based 3D reconstruction technology.

3.3.1. Three-Dimensional Reconstruction Technology Based on Stereo Vision

Stereo vision-based three-dimensional reconstruction technology uses two or more cameras to capture images from different angles based on the principle of stereo matching to reproduce three-dimensional scenes. The principle of binocular stereo vision 3D matching is shown in Figure 10. The application of 3D reconstruction technology based on stereo vision can effectively enhance the environment perception ability of the fruit-picking robot and increase its working efficiency. Li et al. [88] used a binocular stereo-vision system to obtain the 3D position of apples. The approach is to first detect the apples using Faster R-CNN on the images captured by the vision system, then segment the apples from the images based on color features, and finally, perform stereo matching to identify the three-dimensional spatial information of the apples. However, this approach is less effective when the fruit

is occluded. For the problem of occluded fruits, Si et al. [9] proposed a different solution by using two cameras to build a stereo vision system that identifies apples in orchards. The approach uses the RRM algorithm to identify the spatial 3D position of apples by taking images from different angles with different cameras in the stereo vision system. The system was tested in an actual orchard and showed good performance, but limited by lighting conditions, there were differences in the quality of images taken by the stereo vision system from different angles, which affected the recognition performance of the system. To address this problem, Wang et al. [89] proposed a binocular-vision-based recognition system for lychee in unstructured environments. The approach first optimizes the brightness of the image by using wavelet transform before using the clustering algorithm to segment the lychee from the image segmentation, and finally, the images taken by the left and right cameras are stereo matched to determine the spatial three-dimensional position of the lychee. The system uses wavelet variation to optimize the quality of the image, which improves the robustness of the system under changing lighting conditions.

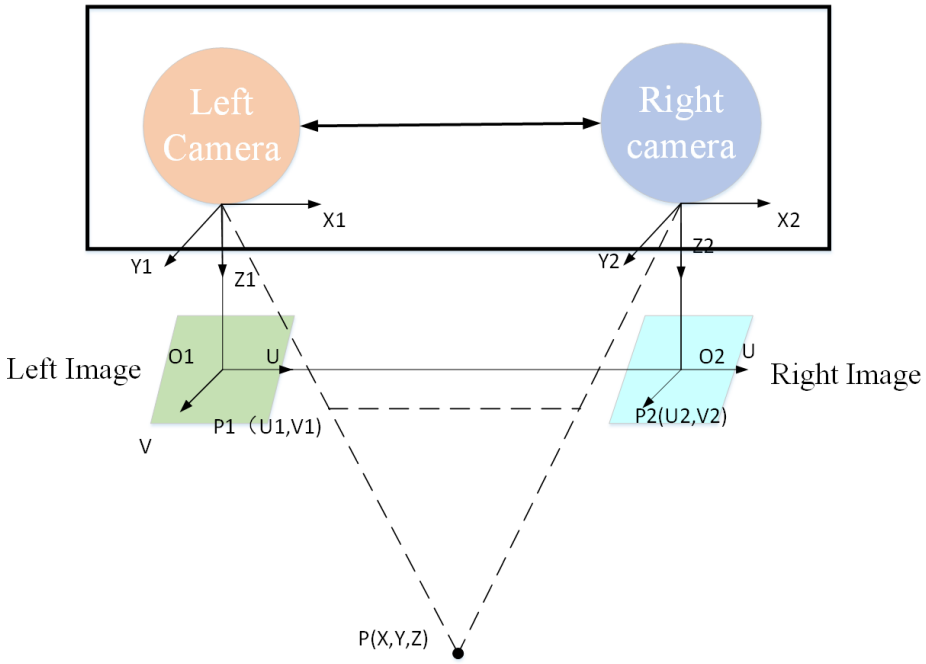

Figure 10. Stereo matching principle based on binocular vision.

3.3.2. Three-Dimensional Reconstruction Technology Based on RGB-D Vision Sensor

RGB-D vision sensors can acquire the sensor-to-object distance information directly through structured light or time-of-flight, offering high measurement accuracy without being easily affected by lighting conditions. RGB-D vision sensors are widely used in the vision recognition system of picking robots. To improve the system's ability to perceive the environment, Tao et al. [90] proposed a point cloud processing-based apple recognition method. Firstly, the point cloud data of the orchard was acquired by using an RGB-D vision sensor, and the color features were fused with the 3D features in the point cloud data to further classify the fruits, branches, and leaves in the image, and then a genetic algorithm and support vector machine were used to classify the data into three categories. According to the test, the accuracy rates of apple, branch and leaf recognition were 92.3%, 88.03% and 80.34%, respectively. As fruit growth in a natural environment is irregular, the branches of fruit trees not only block the fruit but also obstruct the end-effector of the picking robot. For this reason, Lin et al. [91] proposed a method to identify and locate guava fruit using RGB-D vision sensors and successfully reconstructed the fruit and branches in three dimensions. First, Mask R-CNN was used to segment the fruit from the image. Second, a sphere was used to describe the fruit and a cylinder to describe the

branch. Finally, a fitting algorithm was used to process the point cloud to reconstruct the 3D position. The system was tested in an orchard with an F1 score of 83.3% for the 3D reconstruction of fruits and 41.5% for the 3D reconstruction of branches. Because the point cloud collected by the RGB-D vision sensor is sparse for slender branches, which affects the 3D reconstruction effect of branches. For the low recognition accuracy in occluded fruits, Li et al. [92] proposed an apple recognition system based on an RGB-D vision sensor. The system segments the apple fruit in the image by using the YOLACT++ model and then estimates the depth information of the apple by the 3D vision cone before constructing the 3D information of the apple. Through practical experiments in orchards, the apple-picking robot was able to adjust the picking posture of the robotic arm according to the fruit position and size output from this system.

3.3.3. Comparison and Summary of Vision-Based 3D Reconstruction Technology for Fruit

Table 4 compares the research results of vision-based 3D reconstruction technology for fruit. By summarizing the literature, the following three conclusions can be drawn: (a) The 3D reconstruction technology based on stereo vision has high accuracy when measuring close targets, but its difficulty is in carrying out feature points for matching, especially when the fruit is occluded, and the effect of 3D reconstruction is easily affected by lighting conditions. (b) The 3D reconstruction technology based on RGB-D vision sensors measures depth information through structured light or time-of-flight, which can eliminate the influence of changing lighting conditions but has the disadvantages of low accuracy for detecting edge position and low image resolution. (c) The reconstruction process of vision-based fruit reconstruction technology does not involve mechanical motion and depends on the performance of hardware devices, which offers greater potential for future development.

Table 4. Comparing the research results of vision-based 3D reconstruction technology of fruit.

Vision-Based 3D Reconstruction Technology of Fruit	Application to Crops	Technical Characteristics and Performance Indicators	Limitations	References
3D reconstruction technology based on stereo vision	Apple, Litchi	Proximity recognition positioning is high, can be used in the natural environment, comprehensive recognition accuracy of 88%.	Affected by light conditions and when the fruit is shaded, and complicated to calculate when stereo matching.	[9,88,89]
3D Reconstruction Technology Based on RGB-D Vision Sensor	Apple, Guava	Overcomes the influence of changing light conditions, with a wide measurement range and a combined recognition accuracy of 85%.	Sparse point cloud generation for fine branches, low accuracy of object edge positioning.	[90–92]

4. Challenges and Future Directions of Target Recognition Research for Picking Robots

4.1. Existing Challenges

1. An unstructured orchard environment raises the difficulty of fruit recognition. In the actual working environment, the change in lighting conditions easily affects the color characteristics of the target fruit, and the shading of branches, leaves, and other background objects, as well as overlapped fruit, easily lead to missed fruit and other problems in the target recognition system. The traditional target recognition technology is substantially limited to the complex environment of actual work, and the fruit recognition efficiency is low. Although deep learning technology can improve the target recognition performance of picking robots in complex environments, there are still many uncontrollable influencing factors, and the stability of the visual recognition

system needs to be considered. In order to obtain more accurate 3D information on fruits, vision-based 3D reconstruction technology can be used as an effective approach, but the existing 3D reconstruction technology still faces various challenges. For example, the 3D reconstruction technology based on a stereo vision system has difficulty accurately matching the overlapped fruit in three dimensions, and the 3D reconstruction technology based on an RGB-D vision sensor has the problem of insufficient filling rate when collecting slender branches. Therefore, an unstructured orchard environment raises the difficulty of fruit recognition.

2. The real-time nature of the vision system in picking robots makes it difficult to meet the actual production needs. To improve the recognition performance of its vision system, the recognition algorithms are often examined. Although the method can effectively improve the recognition accuracy of the algorithm, this also makes the structure of the algorithm even more complex and the operation time longer. Consequently, actual production needs are often not met.
3. The scale of the data set affects the robustness of the target recognition system of the picking robot. To improve the robustness of the target detection model, the images in the dataset need to contain different lighting conditions, shading degrees, fruit overlapping, and growth cycles of their own characteristics. Therefore, the data images need to be acquired at different times, and at the same time, large-scale acquisition tasks need to be completed within a specific period. Therefore, large-scale data set acquisition is one of the important tasks of fruit target recognition.
4. The generality of target recognition models needs to be improved. In most of the existing studies, the improvement of target detection algorithms for fruits only targets a specific situation, while in a working environment, many different situations may occur simultaneously, such as fruits at different growth stages. Therefore, the development of highly generalized target detection models is beneficial to improve the accuracy of the picking robot's decision-making.

4.2. The Future Directions of Development

1. Research on accurate fruit recognition technology for complex environments. The complex working environment is still one of the constraints for the development of picking robots. Although many researchers have proposed various improvement measures for fruit detection models, different lighting conditions, overlapping and occlusion will still affect the recognition performance of picking robots. At the same time, the complex environment also poses certain challenges to the 3D reconstruction of fruits, and the changes in lighting conditions and fruit overlapping pose challenges to the 3D construction of the visual recognition system. Therefore, it is important to improve the robustness, generalization, and versatility of the visual recognition system for picking robots in complex environments.
2. Research on target recognition technology for fruit in a dynamic environment. During the operation of the picking robot, the fruit is generally not completely stationary, and the wind and picking action will cause the fruit to oscillate dynamically. Although the design of the end-effector of the picking robot can accommodate certain position deviations, the randomness and complexity of fruit oscillation may still lead to the failure of the end-effector picking, thus damaging the picked fruit or the end-effector.
3. Lightweight model research for fruit target recognition. Since picking robots pose high requirements for target detection models, they not only require a fast detection speed to achieve real-time detection but also small model sizes so that the model can be embedded in the devices. Therefore, subsequently, the focus should be on developing lightweight target detection models that can be used in edge devices for real-time fruit detection and improving the performance of visual recognition systems in embedded devices.

4. Target recognition technology research for small-scale data sets. Target recognition technology often requires a large number of data sets covering various aspects. The process of data acquisition and labeling thus takes a longer time and greater effort. However, the research using small-scale data models in existing target recognition techniques is underrepresented. Therefore, how to use small-scale data sets for model training is an important research direction.

Author Contributions: The presented work was under the supervision of Y.L. and H.L.: conceptualization, methodology, software, and writing—original draft; X.H.: validation, writing—review and editing; C.H. and T.C.: methodology and writing—review; J.Z. and L.T.: validation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Guangdong Province Modern Agricultural Industry Technology System Innovation Team Construction Project (Tea)—Tea Industry Innovation Team Facility and Mechanization Post Expert (grant No. 2023KJ120), The "14th Five-Year Plan" Guangdong Province Agricultural Science and Technology Innovation in the Ten main Directions "Unveiling the List of Hanging" Project—Lingnan Characteristic Fruit Intelligent Harvesting Technology (grant No. 2022SDZG03) and Guangdong Province Special fund for scientific and technological innovation strategy ("Climbing Plan" Special fund) (grant No. pdjh2022a0072).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study did not report any data.

Acknowledgments: The authors acknowledge the editors and reviewers for their constructive comments and all the support for this work.

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

References

1. 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]
2. Zhou, Y.; Tang, Y.; Zou, X.; Wu, M.; Tang, W.; Meng, F.; Zhang, Y.; Kang, H. Adaptive Active Positioning of Camellia oleifera Fruit Picking Points: Classical Image Processing and YOLOv7 Fusion Algorithm. *Appl. Sci.* **2022**, *12*, 12959. [CrossRef]
3. Schertz, C.E.; Brown, G. Basic considerations in mechanizing citrus harvest. *Trans. ASAE* **1968**, *11*, 343–0346.
4. Parrish, E.A.; Goksel, A.K. Pictorial pattern recognition applied to fruit harvesting. *Trans. ASAE* **1977**, *20*, 822–0827. [CrossRef]
5. Altaheri, H.; Alsulaiman, M.; Muhammad, G. Date fruit classification for robotic harvesting in a natural environment using deep learning. *IEEE Access* **2019**, *7*, 117115–117133. [CrossRef]
6. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
7. Zhou, H.; Wang, X.; Au, W.; Kang, H.; Chen, C. Intelligent robots for fruit harvesting: Recent developments and future challenges. *Precis. Agric.* **2022**, *23*, 1856–1907. [CrossRef]
8. Li, J.; Tang, Y.; Zou, X.; Lin, G.; Wang, H. Detection of fruit-bearing branches and localization of litchi clusters for vision-based harvesting robots. *IEEE Access* **2020**, *8*, 117746–117758. [CrossRef]
9. Si, Y.; Liu, G.; Feng, J. Location of apples in trees using stereoscopic vision. *Comput. Electron. Agric.* **2015**, *112*, 68–74. [CrossRef]
10. Liang, C.; Xiong, J.; Zheng, Z.; Zhong, Z.; Li, Z.; Chen, S.; Yang, Z. A visual detection method for nighttime litchi fruits and fruiting stems. *Comput. Electron. Agric.* **2020**, *169*, 105192. [CrossRef]
11. Li, B.; Wang, M.; Wang, N. Development of a real-time fruit recognition system for pineapple harvesting robots. In Proceedings of the 2010 Pittsburgh, Pittsburgh, PA, USA, 20–23 June 2010; American Society of Agricultural and Biological Engineers: St. Joseph, Michigan, USA, 2010; p. 1.
12. Bulanon, D.; Kataoka, T.; Ota, Y.; Hiroma, T. AE—Automation and emerging technologies: A segmentation algorithm for the automatic recognition of Fuji apples at harvest. *Biosyst. Eng.* **2002**, *83*, 405–412. [CrossRef]
13. Zhou, R.; Damerow, L.; Sun, Y.; Blanke, M.M. Using colour features of cv. 'Gala' apple fruits in an orchard in image processing to predict yield. *Precis. Agric.* **2012**, *13*, 568–580. [CrossRef]
14. Whittaker, D.; Miles, G.; Mitchell, O.; Gaultney, L. Fruit location in a partially occluded image. *Trans. ASAE* **1987**, *30*, 591–0596. [CrossRef]
15. Hannan, M.; Burks, T.; Bulanon, D.M. A machine vision algorithm combining adaptive segmentation and shape analysis for orange fruit detection. *Agric. Eng. Int. CIGR J.* **2009**, *11*, 1281.
16. Zhang, J.; Tan, T.; Ma, L. Invariant texture segmentation via circular Gabor filters. In Proceedings of the 2002 International Conference on Pattern Recognition, Quebec City, QC, Canada, 11–15 August 2002; Volume 2, pp. 901–904.

17. Kurtulmus, F.; Lee, W.S.; Vardar, A. Green citrus detection using 'eigenfruit', color and circular Gabor texture features under natural outdoor conditions. *Comput. Electron. Agric.* **2011**, *78*, 140–149. [CrossRef]
18. Chaivivatrakul, S.; Dailey, M.N. Texture-based fruit detection. *Precis. Agric.* **2014**, *15*, 662–683. [CrossRef]
19. Payne, A.B.; Walsh, K.B.; Subedi, P.; Jarvis, D. Estimation of mango crop yield using image analysis–segmentation method. *Comput. Electron. Agric.* **2013**, *91*, 57–64. [CrossRef]
20. Payne, A.; Walsh, K.; Subedi, P.; Jarvis, D. Estimating mango crop yield using image analysis using fruit at 'stone hardening'stage and night time imaging. *Comput. Electron. Agric.* **2014**, *100*, 160–167. [CrossRef]
21. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
23. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
25. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
26. Zhou, Y.; Xu, T.; Deng, H.; Miao, T. Recognition method of tomato key organs based on dual convolution Fast R-CNN. *J. Shenyang Agric. Univ.* **2018**, *49*, 65–74.
27. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
28. Wang, Z.; Underwood, J.; Walsh, K.B. Machine vision assessment of mango orchard flowering. *Comput. Electron. Agric.* **2018**, *151*, 501–511. [CrossRef]
29. Gao, F.; Fu, L.; Zhang, X.; Majeed, Y.; Li, R.; Karkee, M.; Zhang, Q. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. *Comput. Electron. Agric.* **2020**, *176*, 105634. [CrossRef]
30. Song, Z.; Fu, L.; Wu, J.; Liu, Z.; Li, R.; Cui, Y. Kiwifruit detection in field images using Faster R-CNN with VGG16. *IFAC-PapersOnLine* **2019**, *52*, 76–81. [CrossRef]
31. Tu, S.; Pang, J.; Liu, H.; Zhuang, N.; Chen, Y.; Zheng, C.; Wan, H.; Xue, Y. Passion fruit detection and counting based on multiple scale faster R-CNN using RGB-D images. *Precis. Agric.* **2020**, *21*, 1072–1091. [CrossRef]
32. Fu, L.; Feng, Y.; Majeed, Y.; Zhang, X.; Zhang, J.; Karkee, M.; Zhang, Q. Kiwifruit detection in field images using Faster R-CNN with ZFNet. *IFAC-PapersOnLine* **2018**, *51*, 45–50. [CrossRef]
33. Parvathi, S.; Selvi, S.T. Detection of maturity stages of coconuts in complex background using Faster R-CNN model. *Biosyst. Eng.* **2021**, *202*, 119–132. [CrossRef]
34. Sun, J.; He, X.; Ge, X.; Wu, X.; Shen, J.; Song, Y. Detection of key organs in tomato based on deep migration learning in a complex background. *Agriculture* **2018**, *8*, 196. [CrossRef]
35. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
36. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
37. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
38. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
39. Xiong, J.; Liu, Z.; Chen, S.; Liu, B.; Zheng, Z.; Zhong, Z.; Yang, Z.; Peng, H. Visual detection of green mangoes by an unmanned aerial vehicle in orchards based on a deep learning method. *Biosyst. Eng.* **2020**, *194*, 261–272. [CrossRef]
40. Xue, Y.; Huang, N.; Tu, S.; Mao, L.; Yang, A.; Zhu, X.; Yang, X.; Chen, P. Immature mango detection based on improved YOLOv2. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 173–179.
41. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
42. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Tian, Y.; Yang, G.; Wang, Z.; Wang, H.; Li, E.; Liang, Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput. Electron. Agric.* **2019**, *157*, 417–426. [CrossRef]
45. Liu, G.; Nouaze, J.C.; Touko Mbouembe, P.L.; Kim, J.H. YOLO-tomato: A robust algorithm for tomato detection based on YOLOv3. *Sensors* **2020**, *20*, 2145. [CrossRef]

46. Fu, L.; Feng, Y.; Wu, J.; Liu, Z.; Gao, F.; Majeed, Y.; Al-Mallahi, A.; Zhang, Q.; Li, R.; Cui, Y. Fast and accurate detection of kiwifruit in orchard using improved YOLOv3-tiny model. *Precis. Agric.* **2021**, *22*, 754–776. [CrossRef]
47. Xu, Z.F.; Jia, R.S.; Sun, H.M.; Liu, Q.M.; Cui, Z. Light-YOLOv3: Fast method for detecting green mangoes in complex scenes using picking robots. *Appl. Intell.* **2020**, *50*, 4670–4687. [CrossRef]
48. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
49. Hao, W.; Zhili, S. Improved mosaic: Algorithms for more complex images. *J. Phys. Conf. Ser.* **2020**, *1684*, 012094. [CrossRef]
50. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
51. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
52. Fu, L.; Wu, F.; Zou, X.; Jiang, Y.; Lin, J.; Yang, Z.; Duan, J. Fast detection of banana bunches and stalks in the natural environment based on deep learning. *Comput. Electron. Agric.* **2022**, *194*, 106800. [CrossRef]
53. Zheng, T.; Jiang, M.; Li, Y.; Feng, M. Research on tomato detection in natural environment based on RC-YOLOv4. *Comput. Electron. Agric.* **2022**, *198*, 107029. [CrossRef]
54. Gai, R.; Chen, N.; Yuan, H. A detection algorithm for cherry fruits based on the improved YOLO-v4 model. *Neural Comput. Appl.* **2021**, 1–12. [CrossRef]
55. Wu, D.; Lv, S.; Jiang, M.; Song, H. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Comput. Electron. Agric.* **2020**, *178*, 105742. [CrossRef]
56. Tang, Y.; Zhou, H.; Wang, H.; Zhang, Y. Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. *Expert Syst. Appl.* **2023**, *211*, 118573. [CrossRef]
57. Zhang, Y.; Yu, J.; Chen, Y.; Yang, W.; Zhang, W.; He, Y. Real-time strawberry detection using deep neural networks on embedded system (rtsd-net): An edge AI application. *Comput. Electron. Agric.* **2022**, *192*, 106586. [CrossRef]
58. Qi, X.; Dong, J.; Lan, Y.; Zhu, H. Method for identifying litchi picking position based on YOLOv5 and PSPNet. *Remote Sens.* **2022**, *14*, 2004. [CrossRef]
59. Chen, S.; Xiong, J.; Jiao, J.; Xie, Z.; Huo, Z.; Hu, W. Citrus fruits maturity detection in natural environments based on convolutional neural networks and visual saliency map. *Precis. Agric.* **2022**, *23*, 1515–1531. [CrossRef]
60. Wang, Z.; Jin, L.; Wang, S.; Xu, H. Apple stem/calyx real-time recognition using YOLO-v5 algorithm for fruit automatic loading system. *Postharvest Biol. Technol.* **2022**, *185*, 111808. [CrossRef]
61. Lyu, S.; Li, R.; Zhao, Y.; Li, Z.; Fan, R.; Liu, S. Green citrus detection and counting in orchards based on YOLOv5-CS and AI edge system. *Sensors* **2022**, *22*, 576. [CrossRef]
62. 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]
63. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Cham, Switzerland, 2016; pp. 21–37.
64. Peng, H.; Huang, B.; Shao, Y.; Li, Z.; Zhang, C.; Chen, Y.; Xiong, J. General improved SSD model for picking object recognition of multiple fruits in natural environment. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 155–162.
65. Wang, Y.; Xue, J. Lightweight object detection method for Lingwu long jujube images based on improved SSD. *Trans. Chin. Soc. Agric. Eng* **2021**, *37*, 173–182.
66. Li, Z.; Zhou, F. FSSD: Feature fusion single shot multibox detector. *arXiv* **2017**, arXiv:1712.00960.
67. Peng, H.; Li, J.; Xu, H.; Chen, H.; Xing, Z.; He, H.; Xiong, J. Litchi detection based on multiple feature enhancement and feature fusion SSD. *Trans. Chin. Soc. Agric. Eng* **2022**, *38*, 169–177.
68. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
69. Zhong-hua, Z.; Wei-kuan, J.; Wen-jing, S.; Su-juan, H.; Ze, J.; Yuan-jie, Z. Green Apple Detection Based on Optimized FCOS in Orchards. *Spectrosc. Spectr. Anal.* **2022**, *42*, 647–653.
70. Sun, J.; Qian, L.; Zhu, W.; Zhou, X.; Dai, C.; Wu, X. Apple detection in complex orchard environment based on improved RetinaNet. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 314–322.
71. Zhao, H.; Qiao, Y.; Wang, H.; Yue, Y. Apple fruit recognition in complex orchard environment based on improved YOLOv3. *Trans. Chin. Soc. Agric. Eng. Trans. CSAE* **2021**, *37*, 127–135.
72. yan, J.; Zhang, L.; Zhao, Y.; Zhang, F. Image recognition of Rosa roxburghii fruit by imoroved RetinaNet. *J. Chin. Agric. Mech.* **2021**, *42*, 78–83.
73. Song, H.; Wang, Y.; Wang, Y.; Lu, S.; Jiang, M. Camellia oleifera Fruit Detection in Natural Scene Based on YOLO v5s. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 234–242.
74. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
75. 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]

76. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
77. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
78. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
79. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
80. Peng, H.; Xue, C.; Shao, Y.; Chen, K.; Xiong, J.; Xie, Z.; Zhang, L. Semantic segmentation of litchi branches using DeepLabV3+ model. *IEEE Access* **2020**, *8*, 164546–164555. [CrossRef]
81. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
82. Majeed, Y.; Zhang, J.; Zhang, X.; Fu, L.; Karkee, M.; Zhang, Q.; Whiting, M.D. Apple tree trunk and branch segmentation for automatic trellis training using convolutional neural network based semantic segmentation. *IFAC-PapersOnLine* **2018**, *51*, 75–80. [CrossRef]
83. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
84. Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [CrossRef]
85. Jia, W.; Tian, Y.; Luo, R.; Zhang, Z.; Lian, J.; Zheng, Y. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [CrossRef]
86. Wang, D.; He, D. Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background. *Comput. Electron. Agric.* **2022**, *196*, 106864. [CrossRef]
87. Xu, P.; Fang, N.; Liu, N.; Lin, F.; Yang, S.; Ning, J. Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation. *Comput. Electron. Agric.* **2022**, *197*, 106991. [CrossRef]
88. Li, T.; Fang, W.; Zhao, G.; Gao, F.; Wu, Z.; Li, R.; Fu, L.; Dhupia, J. An improved binocular localization method for apple based on fruit detection using deep learning. *Inf. Process. Agric.* **2021**. [CrossRef]
89. Wang, C.; Zou, X.; Tang, Y.; Luo, L.; Feng, W. Localisation of litchi in an unstructured environment using binocular stereo vision. *Biosyst. Eng.* **2016**, *145*, 39–51. [CrossRef]
90. Tao, Y.; Zhou, J. Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking. *Comput. Electron. Agric.* **2017**, *142*, 388–396. [CrossRef]
91. Lin, G.; Tang, Y.; Zou, X.; Wang, C. Three-dimensional reconstruction of guava fruits and branches using instance segmentation and geometry analysis. *Comput. Electron. Agric.* **2021**, *184*, 106107. [CrossRef]
92. Li, T.; Feng, Q.; Qiu, Q.; Xie, F.; Zhao, C. Occluded Apple Fruit Detection and localization with a frustum-based point-cloud-processing approach for robotic harvesting. *Remote Sens.* **2022**, *14*, 482. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.