Computers and Electronics in Agriculture 215 (2023) 108425

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Review article

A survey of deep learning-based object detection methods in crop counting

Yuning Huang [a], Yurong Qian [a,b,c,d,*], Hongyang Wei [a], Yiguo Lu [a], Bowen Ling [a], Yugang Qin [a]

[a] School of Software, Xinjiang University, Urumqi, 830000, China
[b] Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Urumqi, 830000, China
[c] Key Laboratory of Software Engineering, Xinjiang University, Urumqi, 830000, China
[d] College of Information Science and Engineering, Xinjiang University, Urumqi, 830000, China

ARTICLE INFO

Keywords:
Deep learning
Object detection
Precision agriculture
Crop counting
Occlusion

ABSTRACT

Crop counting is a crucial step in crop yield estimation. By counting, crop growth status can be accurately detected and adjusted, improving crop yield and quality. In recent years, with the rapid development of convolutional neural networks, deep learning-based object detection methods have been widely used in crop counting. By summarizing the research related to crop counting, this paper reviews the development status of object detection and crop counting. It then compares deep learning-based object detection counting methods with other counting methods. The paper also introduces public datasets and evaluation metrics commonly used for algorithmic models and provides a more in-depth analysis of the application of object detection in crop counting. Finally, the current problems that need to be solved, such as the lack of datasets, difficulties in small object counting, occlusion in complex environments, and some future directions are summarized. We hope this review will encourage more researchers to use deep-learning object detection methods in agriculture.

1. Introduction

1.1. Overview

Precision agriculture is an agricultural technology allowing precise production control. Precision agriculture includes determining the yield changes in the field and their causes, determining possible solutions based on economic rationality, implementing new technologies, and repeating this process. It can be used to improve crop production's economic and environmental sustainability (Syal et al., 2013). Precision agriculture has become a significant trend and direction in the current development of modern agriculture in the world. Application areas are not only for wheat, rice, corn, and other crops but also for the cultivation process of fruits and vegetables. The tasks involved include crop counting, pest detection and yield estimation (Kamilaris and Prenafeta-Boldú, 2018).

Crop counting is an essential basis for precision agriculture. It is the core of plant density estimation, crop yield estimation, developmental stage observation, and plant phenotypic analysis. It is widely used to solve practical problems in agricultural production. Accurate crop identification and statistics are critical for monitoring crop growth, estimating yield, and analyzing plant phenotypic characteristics. They can optimize agricultural production, help effectively manage crop planting and harvesting processes, and promptly develop strategies for harvesting, packaging, transportation, and marketing (Qureshi et al., 2017). Especially crop yield estimation is crucial to address emerging food security issues. In addition to helping farmers make appropriate economic and management decisions, accurate yield estimates can also help prevent famine (Muruganantham et al., 2022).

Crop counting plays different roles in different growth stages. The pre-growth counting evaluates the seed germination rate and seedling survival rate, the mid-growth counting evaluates the plant mortality caused by potential pests and diseases (Neupane et al., 2019), and the post-growth counting accurately evaluates the crop yield. Therefore, realizing automatic, rapid, and accurate calculation of the whole growth period of crops is a critical problem that needs to be solved urgently (Zhang and Li, 2023).

The traditional manual counting methods for empirical yield prediction based on the number of spikes and grains are time-consuming and inaccurate and are far from meeting the needs of large-scale data analysis and processing in "precision agriculture"(Weng et al., 2019). Therefore, a more intelligent and efficient way should be chosen to meet the needs of large-scale genotype and phenotype analysis.

In recent years, with the continuous advancement of deep learning techniques and hardware performance, deep learning-based methods have become the most advanced techniques in image detection and classification problems. Such methods have promising applications in fields such as agriculture (Sun et al., 2020), where crops (Afonso

* Corresponding author at: School of Software, Xinjiang University, Urumqi, 830000, China.
E-mail address: qyr@xju.edu.cn (Y. Qian).

https://doi.org/10.1016/j.compag.2023.108425
Received 14 May 2023; Received in revised form 28 October 2023; Accepted 6 November 2023
Available online 11 November 2023

et al., 2020), flowers (Farjon et al., 2020), and leaves (Xie et al., 2023) can be accurately detected and counted by using deep learning techniques, which not only reduce the labor intensity of manually measuring phenotypic information but also are critical steps for automating processes such as harvesting. In these fields, they can handle significant variations in the data better than traditional computer vision methods. This includes deep learning-based object detection methods for crop counting can be a good solution to the problems in traditional methods, greatly reducing the time spent on crop counting and improving the efficiency of crop estimation.

1.2. Methodology and contributions

In order to ensure the comprehensiveness and credibility of the review content, we first specified specific keywords and search strings for the research topics in the field. We used "crop counting", "crop detection and counting", and "deep learning in crop counting" as the keywords in several academic databases and academic search engines. An extensive literature search was conducted, including Google Scholar, ScienceDirect, IEEE Xplore, etc. These databases and search engines screened papers on various object detection methods for crop counting applications from 2019 to 2023. Furthermore, by tracking and analyzing the citations of published papers, we found more literature related to our research topics, especially for those widely cited classical literature.

There have been a number of reviews related to crop counting (Darwin et al., 2021; Maheswari et al., 2021; Koirala et al., 2019b). Maheswari et al. (2021) focuses on a review of research on various fruit yield estimation using deep learning-based semantic segmentation methods. Koirala et al. (2019b) reviews methods for orchard yield estimation in response to the problem of occluded fruits in imaging. Orchard yield estimation is part of crop counting, and the review does not summarize the research on crop counting as a whole and the methods used.

Unlike these review articles, this paper mainly highlights the application of object detection in crop counting. This paper discusses the application of object detection in counting tasks and some problems and challenges. For example, object detection accuracy is often the primary indicator to measure its performance in artificial intelligence. However, for tasks such as crop counting, whether it can effectively deal with crop occlusion, density change, and other issues is more important. In addition, this paper also discusses different counting methods, including density map-based regression methods and segmentation methods. At the same time, this article also introduces the most widely used deep learning algorithms, such as YOLO, Faster R-CNN, and SSD. These algorithms have achieved remarkable results in counting, which is significant for improving counting accuracy and efficiency. In addition, this article introduces available public datasets and evaluation indicators for crop counting to help researchers better test and evaluate their algorithms.

The outline of the review is as follows: Section 2 describes in detail the manual methods for crop counting, traditional image processing methods, deep learning methods, and the advantages and disadvantages of each method. Section 3 summarizes the deep learning-based object detection methods and the two most widely used models for crop counting. Section 4 lists publicly available datasets that can be used for crop detection and counting research and standard metrics for evaluating model performance. Section 5 provides an in-depth analysis of the problems and possible solutions of object detection algorithms for crop counting. Section 6 concludes this survey with a discussion of possible future work.

2. Crop counting

Object counting has essential research value in many fields, such as agriculture, forestry, medicine, and transportation. Counting work plays a vital role in various aspects, such as animal counting (Delplanque et al., 2023), plant counting (Parico and Ahamed, 2021), cell counting (Ciampi et al., 2022), and vehicle counting (Gomaa et al., 2022) for the protection of endangered plants and animals, the prevention of abnormal cell proliferation, and the counting of traffic peaks.

Crop counting is of vital importance in all aspects of human society. It is relevant to the economic development of countries and an essential foundation for agricultural management, market dynamics, resource planning, and environmental protection, among other areas. Catastrophic weather has caused significant losses in crop production, and crop failures have had a strong impact worldwide. Monitoring crop growth and yield is the only vital technology to address food security (Kuwata and Shibasaki, 2015). Crop yields may vary yearly depending on climate, soil parameters, and fertilizer used. Crop yield estimates are essential for efficient crop management at all stages of the production cycle. This information can help growers prepare packing and storage facilities and estimate harvest logistics and expected costs. It can also help growers identify spatial yield variability on farmland, which will further help them optimize their site-specific management and increase their profits (Li et al., 2016). It also allows producers to ensure their crops are financially compensated in case of crop loss due to weather (Gongal et al., 2016). Physical counting of young fruit, flowers, or fruits at different stages of growth is labor-intensive and an expensive procedure for crop yield estimation.

In recent years, computer vision has been proposed to solve problems with counting various crops, fruits, leaves, etc. Farmers knowing the number of crops to be harvested will help better transport and decision-making. The flowering intensity and peak date can also be further estimated by counting the flowers of fruit trees to improve the yield by precise chemical thinning. Counting leaves permits precise control of plants' conditions and developmental stages (Giuffrida et al., 2018). Agricultural output estimation is essential to address emerging issues of food security. Accurate yield estimation facilitates famine prevention efforts, in addition to helping farmers make appropriate economic and management decisions. Crop counts are essential for agricultural science research, such as breeding and plant phenotype analysis, estimating grain futures prices, and ensuring national food security.

The research development of crop counting has gone through 3 stages: traditional manual counting methods, traditional image processing-based counting methods, and deep learning-based counting methods. Traditional manual counting methods are judged by human eyes, which takes a lot of time and workforce. In addition, due to the remarkable morphological similarity between different plants in the field and the subjectivity of individual observers, manual counting is very error-prone, especially in large-scale production scenarios, which is far from meeting the needs of large-scale data analysis and processing in "precision agriculture"(Madec et al., 2019; Liu et al., 2017). Therefore, more automated and efficient methods have emerged to meet the needs of large-scale variety and phenotype analysis. With the in-depth research and application of computer vision technology, the object counting methods can be divided into image processing-based counting methods and deep learning-based counting methods according to the difference between feature extraction methods.

2.1. Traditional image processing-based counting methods

Traditional image processing-based counting methods have been successfully used in practical applications, such as plant leaf and fruit counting (Saddik et al., 2023). The counting process consists of extracting the features of the object, such as color and shape, and setting them as positive samples and the background color as negative samples, then separating the object and background in the image using traditional machine learning classification methods.

Liu et al. (2019) designed a method for estimating the number of wheat ears based on the improved K-means (Redmon and Farhadi, 2017) algorithm, which establishes a direct mapping relationship from the low-level features of the image to the number of wheat ears contained in the image through color features. This method makes full use of the color features of the wheat image. It uses the area features of the sub-regions extracted within the local area as the basis for wheat ears judgment, thus outputting the number of sub-regions within the clustered area as the estimated number.

Du et al. (2019) firstly preprocessed images by Simple Linear Iterative Cluster(SLIC) (Achanta et al., 2012) for superpixel segmentation; extracted and analyzed some color features of the image, selected the appropriate color feature to train the classifier; then selected the classifier with the highest accuracy to classify the image and recognized the wheat ears.

Bao et al. (2020b) proposed a counting method for wheat images in natural scenes. This method used multi-scale and multi-direction decomposition to highlight the information of wheat spikes and reduce the interference of background information. Then, it used the threshold segmentation method to segment the wheat ear image and morphological operation to separate the connected regions containing the wheat ear information. Finally, the find maximum method was used to calculate wheat ears.

Lu et al. (2017) first transformed the image color space, extracted the color saturation, and then used the method based on concave point detection to match the linkage to achieve the sticky. They then used the method based on the matching line of concave detection to calculate the number of wheat ears. However, in the actual task, the varieties and maturity of wheat vary greatly, and the predefined positive sample color cannot represent the color of all wheat ears, so this method has some limitations.

Counting methods based on traditional image processing rely on pre-defined features, which limits their generalization capability and lowers their robustness. Moreover, these methods suffer from long latency, high sensitivity, and susceptibility to noise interference (Hassan et al., 2022). Previous crop counting methods are mainly implemented by manual counting and traditional image processing methods, which have significant room for improved accuracy and generalization. In contrast, deep learning has an inherent advantage in overcoming some disadvantages of traditional methods for counting with complex backgrounds and dense object distributions.

2.2. Deep learning-based counting methods

Segmentation, classification, and detection methods based on deep learning have been applied in agriculture, and they can better adapt to changes in data and are more effective than traditional methods based on image features. This is because deep learning methods can automatically learn features and classifiers without the need for manual feature extraction and can improve the accuracy and robustness of the model by training with large amounts of data.

There are three types of deep learning-based object counting methods: regression-based on density maps, segmentation, and detection.

2.2.1. Regression method based on the density map

The main idea of regression counting is to learn the direct mapping from the local features of the image to the number of objects, establish the correspondence between the original image and the density map, and calculate the number of objects contained in the image using the total number of pixel values in the density map (Fan et al., 2022) as shown in Fig. 1.

Lu et al. (2017) designed the local counting regression network TasselNet by counting the local images. The corresponding counting map can be obtained after merging and normalizing all the results. This was the first time plant-related problems are considered using computer vision methods in an unconstrained setting, fully considering the complex pose and size variation of maize tassels in images. However, when applying TasselNet to the counting of wheat ears, TasselNet was unable to predict accurate results. For this reason, contextual information was added to the local blocks of the convolutional neural network in TasselNetv2 (Xiong et al., 2019). The first layer of the convolutional neural network was modified using global average pooling, where context can improve the feature extraction performance without increasing the model capacity. Bao et al. (2020a) used the data preprocessing on histogram equalization with threshold segmentation algorithm to reduce the effect of complex background on wheat ear counts. Regression counting was performed on the dataset using the dense scene counting network CSRNet (Li et al., 2018).

Fig. 1. Original images and corresponding density maps obtained by convolving.

Fig. 2. Original images and the corresponding result after semantic segmentation.

Current methods of counting through regression networks, while providing as reliable as possible, do not allow for accurate analysis of crop phenotypes after counting. As crop density increases, occlusion becomes more severe, and the scene becomes more complex. Although regression-based methods can solve occlusion and complex backgrounds to a certain extent, they usually ignore the spatial information of the image. They cannot determine the location of each object in the scene, thus limiting the scope and accuracy of their practical application (Yu et al., 0000; Fan et al., 2022).

2.2.2. Segmentation-based methods

The segmentation algorithm aims to achieve pixel-level separation between crops and the background, accurately distinguishing the object region from the background and achieving detailed edge delineation of the object (Chen et al., 2017; Garcia-Garcia et al., 2017). The segmentation process involves the following steps:

(1) Pixel-wise annotation of the objects in the original image.
(2) Training of a fully convolutional network to densely predict each pixel in the input image, resulting in labeling each pixel with its corresponding object or region class.
(3) Training of a counting convolutional network to obtain segmented images and output intermediate estimates of object counts.

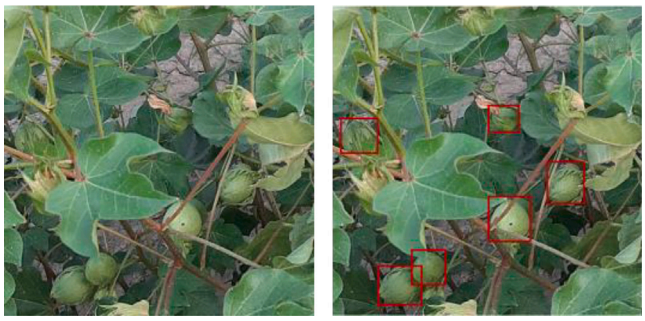

Fig. 3. Original images and the corresponding result after detection.

(4) Making of a linear regression model to map the intermediate count estimates to the final results, using the labels as ground truth.

Fig. 2 shows the original image and its result after segmentation. Many researchers have used segmentation-based methods for counting. Kestur et al. (2019) proposed a 'MangoNet' network that can process images with different sizes of inputs. Experimental results showed robustness under different illumination conditions, varying object sizes with uneven density while better comparing MangoNet performance with FCN variant architecture trained on the same data. Afonso et al. (2020) used Mask R-CNN (He et al., 2017) to segment the crop and the pixels corresponding to each object and used CNN architectures such as ResNet as backbone networks to extract feature maps. The region proposal network(RPN) was applied to the feature map to calculate the region proposal, and then the pooling operation is performed. Finally, a classifier was used to each pooled feature map to obtain the class's bounding box predictions.

Lin and Guo (2020) developed a sorghum head detection and counting network based on UAS images, integrating image segmentation and convolutional neural network models using UAV system images and deep learning to count three different sorghum varieties. Zabawa et al. (2020) used the traditional U-shaped decoder-encoder architecture to achieve the counting task for pixel-by-pixel semantic segmentation. According to the characteristics of the small size of the counting object, in order to prevent the performance loss caused by the down-sampling processing, the traditional U-shaped decoder-encoder architecture is used for pixel-by-pixel semantic segmentation. The encoder backbone network was MobileNetV2 (Sandler et al., 2018), and the concept of inverted residual was introduced to achieve more efficient and lightweight feature extraction. The decoder used refines the segmentation results for DeepLabV3+(Yurtkulu et al., 2019).

To summarize the above studies, it has been discovered that the segmentation-based approach has two difficulties. First, training a fully convolutional network requires pixel-level labeling. This annotation method incurs significant time and effort as it demands precise labeling with detailed information. Secondly, the output mask of the fully convolutional network has no direct relationship with the number of objects. It cannot obtain any available information about the location of the crop, and the processing speed is slow. Solving these problems usually requires designing multifaceted post-processing steps. These problems can be effectively avoided by using deep learning-based object detection methods for counting.

2.2.3. Object detection-based methods

The deep learning-based object detection method is a mainstream method in crop counting research, which first labels the predicted crops with prediction boxes and then counts the obtained prediction boxes. The number of prediction boxes is equal to the number of crops. Object detection methods can accurately locate objects, identify and classify plants of different forms and developmental processes, and quantify them, laying a solid foundation for subsequent agricultural research (see Fig. 3). The method based on object detection divides the task of crop counting into two parts. Firstly, the crop is detected, and then the number of crops is counted to provide indicators for further yield prediction. Counting based on object detection generally contains the following steps:

(1) Image preprocess: Use image scaling, color space change, image data normalization, and other methods to preprocess the image.
(2) Generate candidate regions: Generate possible regions containing objects at each position of the sliding window.
(3) Extract features: Extract features from the generated candidate regions for subsequent classification and regression.
(4) Calculate scores: Use a classifier to calculate the probability and confidence level of an object belonging to a specific category.
(5) Detection: Remove candidate regions with low scores using category confidence thresholds; Suppressing multiple redundant detections, such as non-maximum suppression(NMS), aims to assign a detection box to each object.
(6) Count: Calculate the number of detection boxes generated, and the result is the number of crops.

In the following sections, we describe the deep learning-based object detection methods and explore their application in counting and its problems by summarizing the available solution methods, multiple datasets, and evaluation metrics so that readers can better understand and apply these techniques to solve problems in actual production.

3. Object detection

Object detection aims to locate objects in an image and provide classification labels for the objects (Zou et al., 2023). As one of the most fundamental tasks in computer vision, its primary goal is to accurately localize objects in an image and provide them with classification labels. It is a critical component of many vision applications, including semantic segmentation (Hao et al., 2020) and human pose analysis (Zou et al., 2023). Furthermore, it also plays a crucial role in autonomous driving (Qian et al., 2022), pedestrian tracking (Wei et al., 2022), face recognition (Jiang and Learned-Miller, 2017), and other fields.

Object detection combines object localization and classification into one, and its accuracy and real-time performance are essential indicators of the performance of object detection algorithms. In complex scenarios where multiple objects must be processed in real-time, automatic extraction and recognition of objects is particularly important. This requires the object detection algorithm to process a large amount of image data quickly while ensuring accuracy to meet real-time requirements.

With the continuous development of computer vision technology, object detection algorithms have also evolved and developed. Object detection has undergone two development stages in the past decades, from traditional methods based on hand-designed features to deep learning-based approaches. The performance and effectiveness of object detection algorithms have been greatly improved.

3.1. Traditional object detection methods

Before deep learning methods were widely used, object detection algorithms were mainly based on geometric techniques, focusing on the spatial layout of specific objects, and the models were relatively simple. Based on the traditional object detection algorithm with manually designed features, the steps and processes of detection mainly include pre-processing, region selection, feature extraction, and feature classification processes. In this, pre-processing is the redundant denoising of the detected image. The commonly used methods are histogram equalization, median filtering, mean filtering, Gaussian filtering, and other spatial domain image denoising methods, high pass filtering,

low pass filtering, and other image denoising methods. Region selection is an image processing process that uses windows of different sizes to slide over the image in all directions and uses the image in the window as a candidate region. Feature extraction uses scale-invariant feature transform(IFT) (Lowe, 2004), histogram of oriented gradient(HOG) (Dalal and Triggs, 2005), and other feature extraction methods to extract the image features of candidate regions. Feature classification refers explicitly to the classification of object features using classifiers such as adaptive boosting(Ada-boost) (Korada et al., 2012), support vector machines(SVM) (Auria and Moro, 2008), and deformable parts model(DPM) (Ott and Everingham, 2011).

Hand-designed feature extraction methods convert pixels into features suitable for a specific task and require extensive experience and expertise. However, the extracted features are usually shallow and need better generality and expressiveness.

3.2. Deep learning-based methods

With the rapid development of deep learning technology, parallel computing resources continue to iterate and update, technology and data support make deep learning-based object detection models have made significant progress in terms of accuracy and efficiency and have emerged Single Shot Detector(SSD) (Liu et al., 2016), You Only Look Once(YOLO) (Redmon et al., 2016) the Detection Transformer(DETR) (Carion et al., 2020) and a series of classical research works.

Compared with the traditional method, the deep learning-based method avoids the tedious manual design and can automatically learn deeper features with more differentiation power. At the same time, deep learning-based approaches unify feature extraction and classifier learning in a single framework, enabling end-to-end learning.

The current state-of-the-art deep learning object detection algorithms are divided into two main categories: one-stage and two-stage object detection algorithms. Fig. 4 shows the difference between the implementation process of one-stage and two-stage object detection algorithms. In two-stage object detection algorithms, multiple feature regions of interest are generated in the first stage; in the second stage, the feature vectors of the feature regions generated in the first stage are encoded by a neural network to predict the object class and localize it. The one-stage object detection algorithm considers all locations in the image as regions of interest and tries to classify each region to distinguish the background from the object.

This approach is faster and more prevalent in real-time object detection applications. However, due to the absence of a separate feature region generation phase, one-stage object detection algorithms can be relatively poor in terms of accuracy and can have a high repetition rate. Two-stage object detection algorithms often report the latest results on publicly available benchmark datasets. Such algorithms usually include both feature region generation and object classification phases and thus have more advantages in terms of accuracy and robustness. However, the inference speed is relatively slow due to two stages (Wu et al., 2020).

3.2.1. One-stage object detection methods

YOLO and SSD enable end-to-end detection of images faster than two-stage object detectors. YOLO makes detection a regression problem to allow for faster detection, as shown in Fig. 5, uses the entire image as the input to the network, and regresses the location and category directly in the output layer, completing the task of object detection using only one network. The SSD uses small convolutional kernels to predict the category scores and bounding box offsets of a set of default bounding boxes on the feature map to achieve high detection accuracy, as shown in Fig. 6 generating different scales from different feature maps, using different aspect ratios for separate prediction; because the features are shared throughout the process of image classification and localization, generating a multi-scale feature map achieve improved detection of small objects.YOLO and SSD have good speed, but both have problems detecting category imbalance on small objects. The RetinaNet (Lin et al., 2017) detector improves on this problem. To address the problem of imbalance in the foreground and background category detection in the model, RetinaNet uses a Focal Loss function in the training process and a separate network for implementing the classification and the regression.

YOLOv2 (Redmon and Farhadi, 2017) improves on YOLO by adding techniques such as Batch Normalization, high-resolution classification, and anchor to create multiple bounding boxes to achieve more accurate localization of objects. YOLOv3 (Redmon and Farhadi, 2018) uses a 53-layer backbone network, a separate logical classifier, and binary cross-entropy loss to predict overlapping bounding boxes and smaller objects. YOLOv4 (Bochkovskiy et al., 2020) introduces two fundamental techniques, “bag of freebies” to design data expansion and regularization improvements in the training process, in addition to using various data enhancement methods. The “bag of specials” post-processing module allows for more accurate and faster inference. YOLOv5 (Glenn et al., 2022) further improves data expansion and loss computation, in addition to using automatically learned bounding boxes to adapt to a given dataset for greater flexibility and speed. YOLOX (Ge et al., 2021) uses the anchor-free, decoupled head technique, allowing the model to use separate networks for classification and bounding box regression tasks. Unlike the YOLOv4 and YOLOv5 models, YOLOX uses anchor-free, which eliminates the need to design anchors in advance, reduces the number of parameters, and improves testing speed. YOLOv6 (Li et al., 2022b) introduces a self-distillation strategy in classification and regression tasks to help student models learn more effectively in all training phases by dynamically adjusting the knowledge from teachers and labels. YOLOv7 (Wang et al., 2022a) focuses on optimizing the training process by designing several training-free “bag-of-freebies” methods and using a re-parameterization module with dynamic label assignment strategies that can improve detection accuracy without increasing the parameters.

Ultralytics created YOLOv8 (Jocher et al., 2023) in January 2023, and in addition to object detection, it also supports instance segmentation, enabling multi-object detection in images or videos. The Darknet-53 architecture is used to improve feature extraction, with the network divided into more miniature stages and then partially connected, allowing for better feature reuse and gradient propagation, resulting in more accurate object detection. Unlike other YOLO models, YOLOv8 combines YOLOv4, DarkNet-53, and Pseudo Ensemble(PS) to achieve faster speeds and higher accuracy than the previous YOLOv7 network (Li et al., 2023a).

3.2.2. Two-stage object detection methods

R-CNN (Girshick et al., 2015) is one of the first deep learning-based object detectors and uses an efficient Selective Search algorithm to obtain region proposals. In the R-CNN network, the image input to the CNNs network must be a fixed-size image, meaning that the region proposals extracted by the Selective Search algorithm are not uniform in size and need to undergo a warp operation to make all the region proposals uniform in size before inputting them to the CNN network. The problem with the various dimensions of images is that they look unnatural and may reduce recognition accuracy.

SPPNet (He et al., 2015) adds an SPP layer after the convolution layer to convert the features map into a fixed-length feature vector. And then input to the fully-connected layers. SPPNet not only solves the size problem of the region proposals but also solves the computation time. Fast R-CNN (Girshick, 2015) aims to improve the time-consuming and redundant repeated convolutions in feature extraction. After obtaining candidate regions, only one convolution operation on the entire image is required to obtain all the feature maps. Then, based on the obtained region boundaries, corresponding regions of interest in the feature maps are obtained by performing max pooling on fixed-size blocks. Finally, the obtained features are used for region classification and

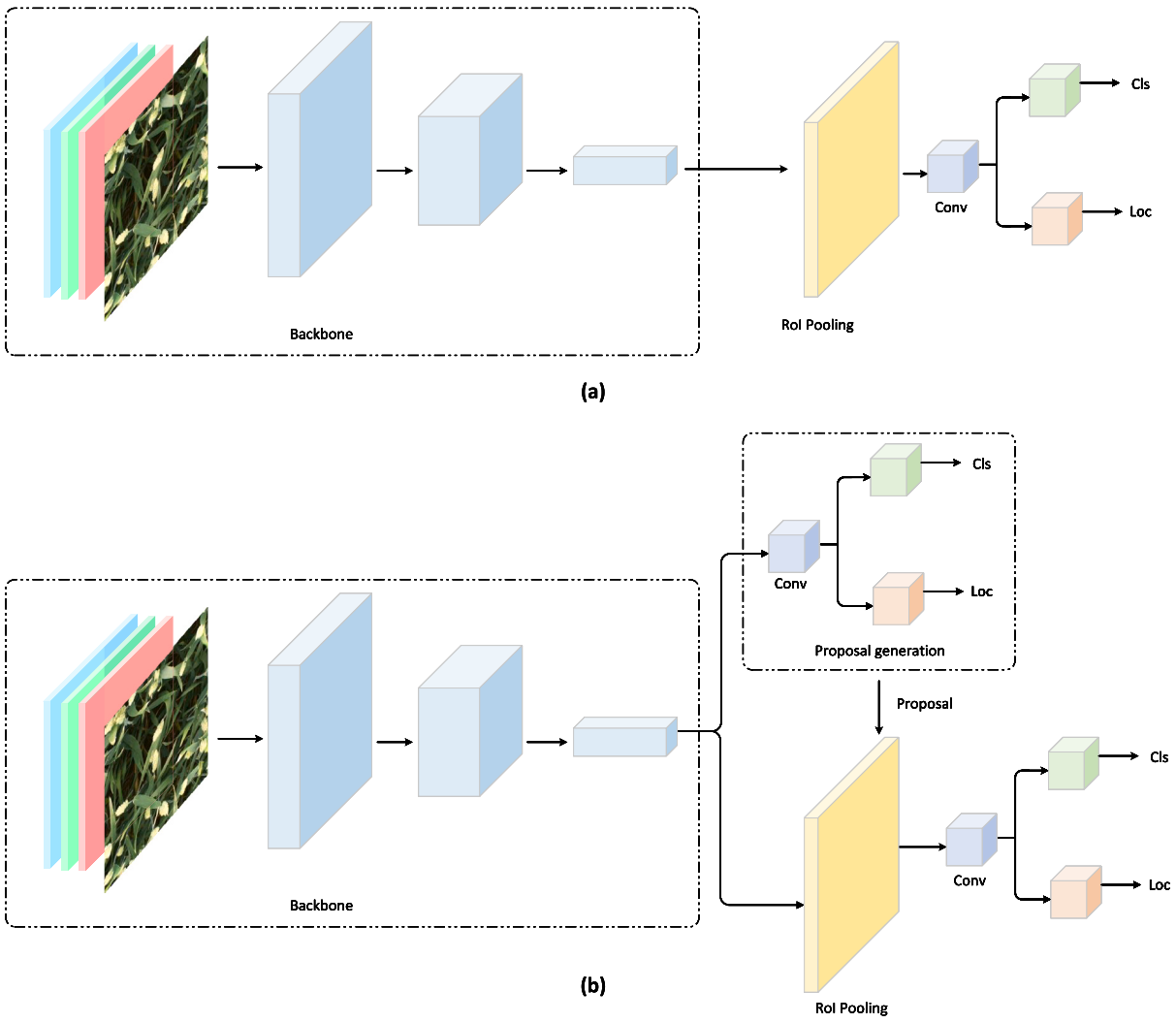

Fig. 4. The structure of one-stage methods and two-stage methods.

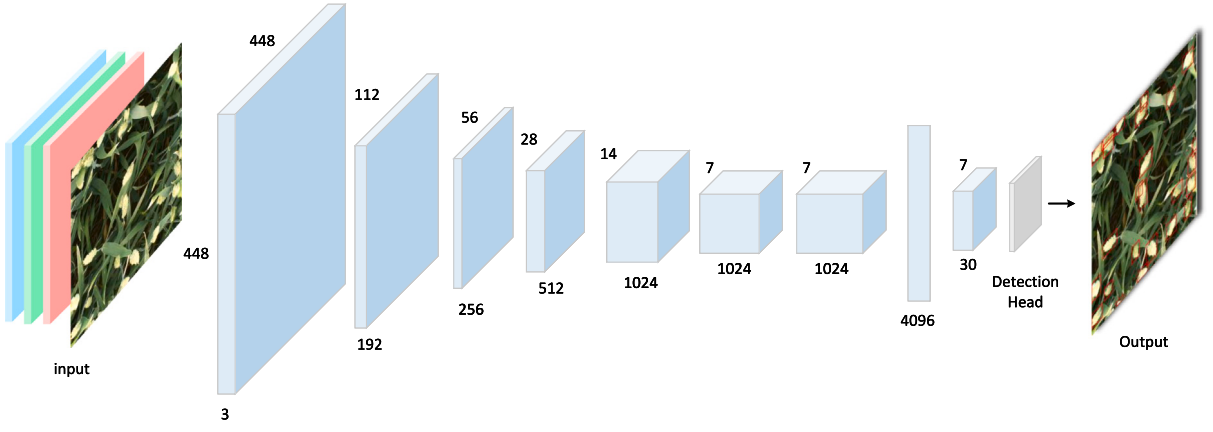

Fig. 5. The structure of YOLOv1.

bounding box regression. Faster R-CNN (Ren et al., 2015) optimizes on Fast R-CNN by not using the Selective Search (Uijlings et al., 2013) to extract proposals but by constructing an RPN structure for extraction, after which the feature Both the feature map and the RoI are input to the RoI pooling layer for reshaping for detection.

Traditional object detection algorithms suit situations with apparent features and little background information. However, in the practical application of crop counting, it is difficult for traditional object detection algorithms to complete accurate detection of objects due to the complicated background information contained in the images, the different scales of the objects to be detected, the large number of

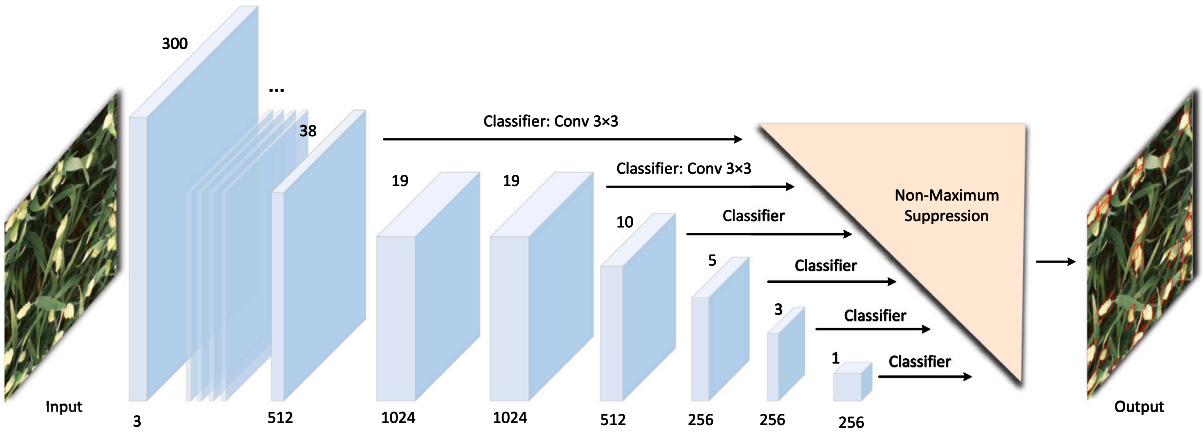

Fig. 6. The structure of SSD.

objects, and the easy overlap. Compared with traditional object detection methods, deep learning-based methods can better adapt to the complex and changing natural environment and can provide strong support for agricultural production informatization in a large field environment. In the future, with the continuous upgrading of computer hardware and the progress of deep learning technology, object detection algorithms will become more and more accurate, fast, and intelligent, providing more affluent and reliable support for more application scenarios.

3.3. Application of object detection in crop counting

There are multiple challenges in the crop counting task. Firstly, the size, color, and texture of the crop may not be similar due to factors such as variety and degree of development; the object's edge shape is irregular, and its color may be similar to the color of the leaves at a particular time; secondly, the light varies significantly under different weather conditions, and the crop posture is variable; shading is prevalent, and the cluttered background can make the probability of missing and false detection very high. These problems create obstacles to applying object detection in crop counting tasks (Qureshi et al., 2017; Li et al., 2016).

Table 1 shows the information of 45 papers collected from 2019 to 2023 using deep learning-based object detection methods to achieve crop counting tasks. The information includes the publication time, the counting target, the dataset, the backbone, and the final experimental results.

Firstly, there are 28 counting objects, including seeds, flowers, fruits, branches, and leaves of grain crops such as wheat, rice, corn, soybeans, and various fruits and vegetables. Of these, wheat ears are the most researched and have publicly available datasets. There are two possible reasons for this: (1) wheat is one of the major food crops in the world, and wheat yield is closely related to all aspects of human life, so counting wheat ears is of great significance for improving food yield and agricultural productivity; (2) many organizations jointly collected the public dataset of wheat ears around the globe, and the dataset contains a large amount of data, is rich in variety, and has a high degree of visibility.

Most of these researchers collected and processed the datasets themselves. However, only 23% used datasets that were available for public download. There may be several reasons for this phenomenon: (1) Each crop has unique characteristics, and it is challenging to develop a uniform standard and format. Data collection and processing by specialized researchers on their own can ensure the quality and completeness of the data. (2) some agriculture-related data may involve intellectual property and confidentiality issues, and it is not appropriate to make the dataset publicly available. How to solve the problem of missing datasets is an issue we focus on in the subsequent sections.

Among these 45 papers, 30 used the one-stage detection algorithm of the YOLO series, of which 15 used YOLOv5. The two-stage algorithm represented by R-CNN has ten, of which Faster R-CNN has seven. The data shows that the usage rate of YOLO-related algorithms is very high, but this does not directly conclude that the two-stage algorithms are not advantageous. They should be selected according to the crop characteristics and other factors. Next, we analyze the most used one-stage algorithm, YOLOv5, and the two-stage algorithm, Faster R-CNN.

3.3.1. YOLOv5

In this section, we analyze and summarize the recent papers on crop counting based on object detection and select the most applied YOLOv5 model for a detailed study to investigate the reasons for its excellent performance in crop counting. Uitralytics LLC publishes YOLOv5 on GitHub, continuously updated and widely used in many fields. YOLOv5 consists of several versions with essentially the same structure but different depths and widths, and the network structure consists of the backbone, neck, and head, as shown in Fig. 7.

Backbone The backbone is used to extract features of different scales of images, including low-level features extracted from shallow layers and high-level semantic features extracted from deep layers. The backbone network mainly includes the Focus module, CBS module, SPP module, and C3 module. The Focus module performs a slicing operation on the input image, expanding the input channel to four times its original size while reducing the width and height of the image to half its original size. The CBS module combines 2D convolution(Conv), batch normalization(BN), and SiLU activation function(SiLU). The C3 module consists of multiple CBS modules that form residual connections. Moreover, the input feature maps are concatenated with the results of multiple residual connections to form a Cross Stage Partial(CSP) module. This step is to extract the information for fusion on feature maps at different scales to increase the reuse of features. SPP is a spatial pyramid pooling structure, usually used to solve the problems of varying input image size and significant variation of object size in images by pooling feature maps of different sizes and stitching the pooled results together.

Neck After extracting the image features through the backbone network, the backbone and the head are connected using the neck. In this model, the neck combines Feature Pyramid Network(FPN) and Pixel Aggregation Network(PAN) to integrate a regular top-down FPN with a bottom-up PAN to fuse the extracted semantic features with fine-grained positional features. The bottom-up feature extraction network extracts feature maps from the original image at different scales. On the other hand, the top-down feature fusion network is implemented by upsampling and cross-layer connection, which can fuse the low-level feature maps with the high-level feature maps and generate a set of feature maps with different resolutions. This structure can improve the

Table 1
The application of deep learning-based object detection methods in crop counting.

Year	Object	Dataset	Method	Results	References
2019	Sorghum	Non-public	RetinaNet	R^2 is 0.88.	Ghosal et al. (2019)
2019	Cotton seedling	Public(SeedingAll)	Faster R-CNN	F1-score is 0.727, and R^2 is 0.98.	Jiang et al. (2019)
2019	Wheat ears	Public(WEDD)	Faster R-CNN	The rRMSE is 5.3%	Madec et al. (2019)
2019	Mango	Public(MangoYOLO)	YOLOv3	F1-score is 0.89.	Koirala et al. (2019a)
2020	Cotton seedling	Non-public	YOLOv3	The mAP and F1 scores are higher than 78%, and 0.77, respectively.	Oh et al. (2020)
2020	Corn plant	Non-public	Faster R-CNN	The mAP are 0.998 and 0.764 at 0.5 and 0.7 Intersection over Union(IOU), respectively.	Zhang et al. (2020)
2020	Flowers	Non-public	Fast R-CNN	The average precision score is 0.68.	Farjon et al. (2020)
2020	Tomato	Non-public	Mask R-CNN	The precision and recall values exceed the values of precision of 0.88 and recall of 0.8. The average precision is 0.82.	Afonso et al. (2020)
2020	Potato, lettuce	Non-public	Mask R-CNN	MOTA is 0.781 for potato plants and 0.918 for lettuces.	Machefer et al. (2020)
2020	Passion fruit	Non-public	Faster R-CNN	The recall, precision, and F1-score are 0.962, 0.931, and 0.946, respectively.	Tu et al. (2020)
2021	Coconuts	Non-public	Faster R-CNN	mAP@0.5 is 0.894.	Parvathi and Selvi (2021)
2021	Wheat ears	Public(WEDD, GWHD)	YOLOv4	R^2 is 0.9884.	Yang et al. (2021)
2021	Maize tassel	Non-public	CenterNet	Precision is 0.9686.	Karami et al. (2021)
2021	Pear	Non-public	YOLOv4	AP@0.5 is 0.98. F1-score is 0.8785.	Parico and Ahamed (2021)
2022	Green Citrus	Non-public	YOLOv5	mAP@0.5 is 0.9823, and the recall is 0.9766.	Lyu et al. (2022)
2022	Wheat ears	Non-public	RetinaNet	mAP50 is 0.9262, and the counting accuracy is 0.9288.	Wen et al. (2022)
2022	Tomato	Non-public	YOLOv5	Precision is 99% for green tomatoes and 85% for red tomatoes.	Egi et al. (2022)
2022	Fruit trees	Non-public	YOLOv4	The mAP is 98.21%, F1-score is 93.60% for detection, and the average overall accuracy (AOA) is 96.73% for counting.	Zhu et al. (2022)
2022	Apple	Non-public	YOLOv4	Counting accuracy is 91.49%, and R2 is 0.9875.	Gao et al. (2022a)
2022	Oranges	Non-public	YOLOv4	Precision is 91.55%, recall is 98.55%, mAP is 93.38%, and F1-score is 0.985.	Chen et al. (2022)
2022	Wheat ears	Public(GWHD)	YOLOv4	mAP is 93.7%.	Zhao et al. (2022)
2022	Wheat ears	Non-public	YOLOx	The precision, recall, average accuracy, and F1-score are 96.83%, 91.29%, 92.29%, and 93.97%, respectively.	Yang Shuqin (2022)
2022	Wheat ears	Non-public	YOLOv3	The precision is over 90%.	Gao et al. (2022b)
2022	Sugarcane Seedlings	Non-public	Faster R-CNN	The average accuracy is 93.67%, and MAE is 4.6.	Pan et al. (2022)
2022	Maize leaves	Non-public	YOLOv5	The precision is 92.0%, recall is 84.4%, the average accuracy is 89.6% for detection. And the precision is 72.9% for counting.	Xu et al. (2022a)
2022	Rice panicle	Non-public	YOLOv5	MAPE is 3.44%, and precision is 92.77%.	Wang et al. (2022b)
2022	Cotton seedling	Non-public	YOLOv4	The F1-score of the final detection model is 0.98, and the average precision is 99.12%.	Tan et al. (2022)
2022	Wheat ears	Public(GWHD, WEDD)	Faster R-CNN	The average error rate is 3.7%, AP is 95.17%, and the precision is 95.8% for counting.	Sun et al. (2022)
2022	Tomato	Non-public	YOLOv5	The mean average precision of flower, green, and red tomatoes is 93.1%, 96.4%, and 97.9%.	Ge et al. (2022)
2022	Cotton seedling	Non-public	CenterNet	F1-score is 0.982, R2 is 0.967, and RMSE is 0.394.	Yang et al. (2022a)
2022	Maize tassel	Public(MTC)	YOLOv5	R^2 is 95.72%, and AP is 86.69%.	Falahat and Karami (2022)
2022	Soybean pods	Non-public	YOLOv5	The average precision is 91.7%, MSE is 0.00865, and the average coefficient of determination R2 is 0.945.	He et al. (2022a)
2022	Camellia fruit	Non-public	YOLOv3	Average precision is 78.40%, and recall is 88.70%.	He et al. (2022b)
2022	Sorghum	Non-public	YOLO	The average precision is 0.95	Mosley et al. (2022)
2022	Cotton flowers	Non-public	DenseNet	The best mean absolute count error achieved by the model is 2.43.	Petti and Li (2022)
2023	Maize trichome	Non-public	YOLOv5	Accuracy is 92.1%.	Xu et al. (2023)
2023	Soybean pods	Public(YOLOPOD)	YOLOX	R^2 is 0.967, MAE, MAPE, RMSE are only 4.18, 10.0%, and 6.48, respectively.	Xiang et al. (2023)
2023	Rapeseed	Public(Rapeseed_Dataset)	YOLOv5	R^2 for counting and the mAP for location are over 0.96 and 92%.	Li et al. (2023b)
2023	Grape	Non-public	YOLOv5	The average counting accuracy is 84.9%, correlation coefficient with manual counting is 0.9905.	Shen et al. (2023)

(*continued on next page*)

Table 1 (*continued*).

Year	Object	Dataset	Method	Results	References
2023	Tea buds	Non-public	YOLOv5	Precision is 91.88%,R2 is 0.98.	Li et al. (2023c)
2023	Tomato	Non-public	YOLOv5	Precision is 97.9%, and mAP@0.5:0.95 is 0.748 for detection. The average precision is 95.1% for counting.	Rong et al. (2023)
2023	Bayberry trees	Public(Bayberry Tree Dataset)	YOLOv4	Precision is 97.78%, and recall is 98.16%.	Chen et al. (2023)
2023	Wheat ears	Non-public	YOLOv5	Precision, recall, and average precision are 87.2%, 84.1%, and 88.8%, respectively.	Bao et al. (2023)
2023	Lettuce	Non-public	YOLOv5	The mAP of the model is 99.42%, Recall is 99.13%, Precision is 98.24%, F1-score is 0.99.	Zhang and Li (2023)
2023	Wheat ears	Non-public	YOLOv5	R^2 in all three fertility stages is 0.87, and RMSE is 0.70.	Shi et al. (2023)

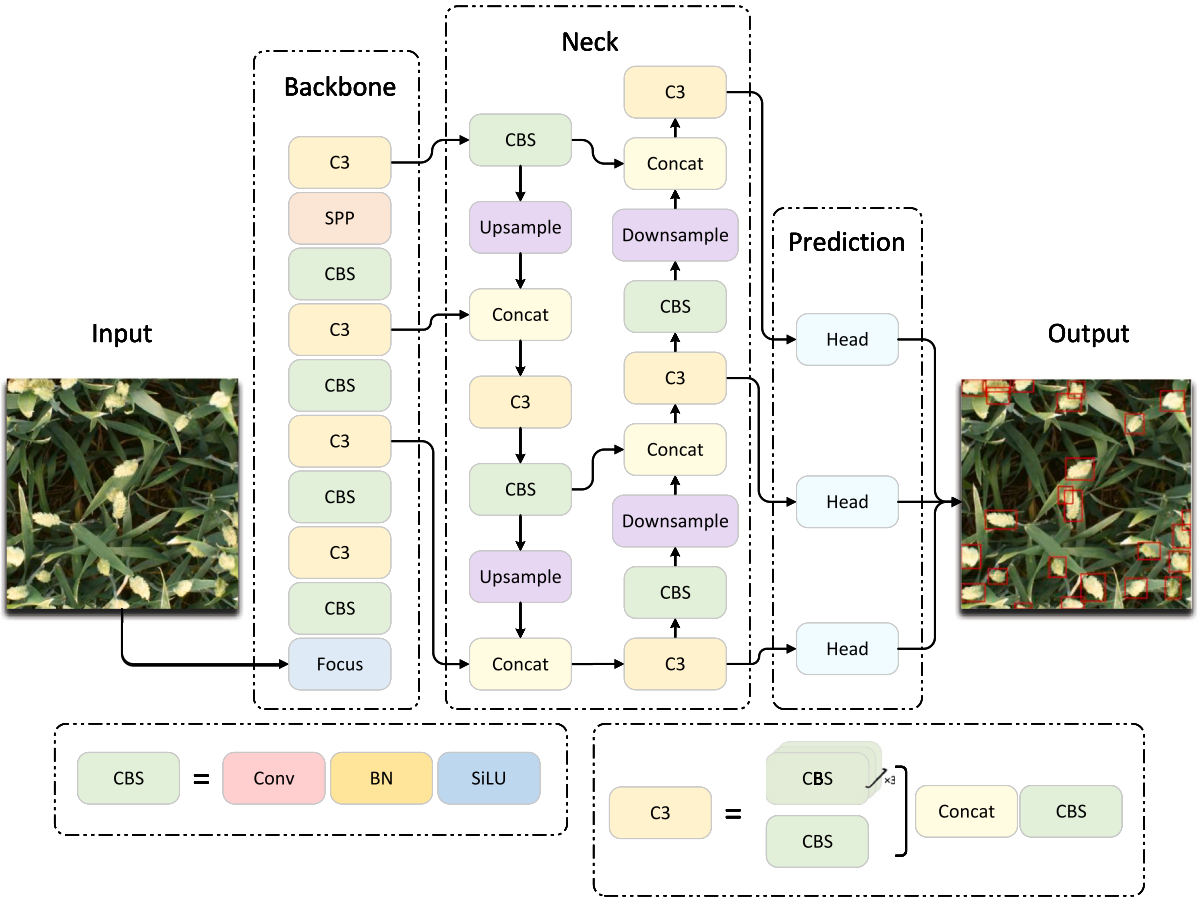

Fig. 7. YOLOv5 overall structure.

model's sensitivity to objects of different scales while improving the detection accuracy.

Head The detection head predicts the feature maps of three scales corresponding to large, medium, and small objects. It detects objects of various sizes and outputs a vector containing the object category probability, confidence, and bounding box position information. Multi-scale feature maps can improve the model's sensitivity to different scale objects and detection accuracy.

YOLOv5 also uses many data enhancement techniques. During the training process, YOLOv5 randomly scales the input image, select horizontal flip or vertical flip, randomly crop, and randomly adjust the input image's brightness, contrast, and saturation. These data augmentation techniques can effectively increase the data's diversity and improve the model's generalization ability, thereby improving the model's performance. The YOLOv5 algorithm retains more features in the feature extraction network and often optimizes the algorithm details. It can combine semantic and detailed information to improve the detection effect of small objects, which leads to the robustness and stability of YOLOv5 in dealing with complex problems.

3.3.2. Faster R-CNN

In two-stage object detection algorithms, the main task of the first stage is to generate a set of region proposals, and then these region proposals are sent to the second stage for coordinate regression and category classification; Faster R-CNN proposes that the previous two-stage object detection network relies on region proposal algorithm to reason about the location of the detected objects. Selective Search (Uijlings et al., 2013) is one of the most applied region proposal algorithms that compute pixels of similar regions based on color, texture, size, and shape merging. However, Selective Search lags significantly in speed compared to effective detection networks.

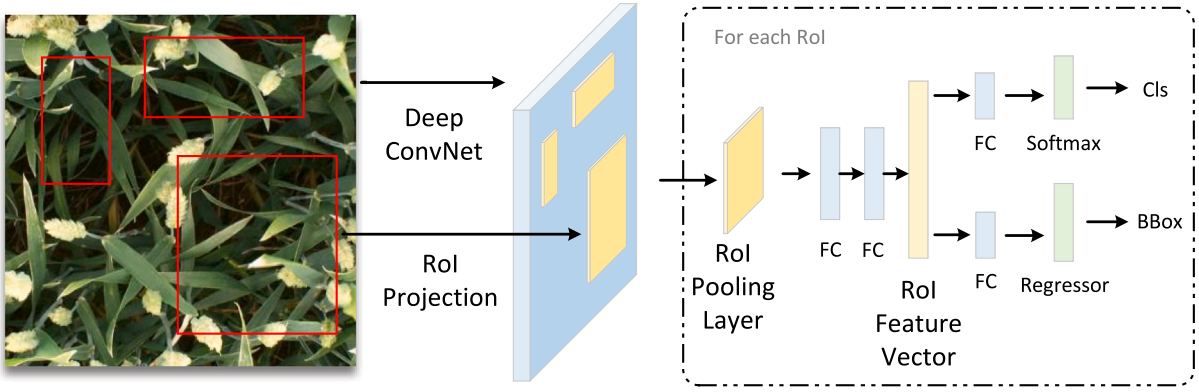

Fig. 8. Faster R-CNN overall structure.

Faster R-CNN is a further improved optimization of Fast R-CNN. Region Proposal Network(RPN) is introduced on top of Fast R-CNN to combine RPN and Fast R-CNN models for end-to-end object detection. In Fig. 8, the object detection process of Faster R-CNN is divided into two stages: (1) generating region proposals using RPN; (2) using Fast R-CNN network to classify and regress the region proposals.

Region proposals generation using RPN R-CNN and Fast R-CNN use the Selective Search algorithm to generate region proposals. The selective search algorithm is based on image segmentation, which divides an image into multiple regions and merges them based on their similarity to obtain candidate regions. Although the selective search algorithm can generate high-quality candidate regions, it is slow and unsuitable for real-time applications.

RPN is a neural network for generating candidate regions that slide a small window over the image and share fully connected layers at all spatial locations. It predicts multiple candidate regions at each location. The input of RPN is an image, and the output is a set of bounding box coordinates and scores of region proposals with different shapes and proportions and can cover objects of different sizes and shapes. RPN shares the convolution features of the entire image with the detection network, making near-zero-cost region proposals possible. It is also a fully convolutional network that can simultaneously predict object boundaries and object scores at multiple locations; after end-to-end training, Faster R-CNN can generate high-quality region proposals and use Fast R-CNN to complete detection.

Using Fast R-CNN to classify and regress In Fast R-CNN, the classification and regression of candidate regions are achieved through an RoI pooling layer. The RoI pooling layer divides each candidate region into fixed-size sub-regions and maps each sub-region to a fixed-size feature map. Then, the RoI pooling layer performs the maximum pooling operation on each sub-region to obtain a fixed-length feature vector. They then sent the feature vector to the fully connected layer network for classification and regression operations.

In general, Faster R-CNN is an efficient and accurate object detection algorithm. It aims to quickly generate candidate regions by introducing the RPN network. Faster R-CNN has achieved excellent performance on multiple object detection datasets and has become one of the essential algorithms in object detection. It has also been widely used in the field of crop counting.

4. Datasets and evaluation metrics

4.1. Public datasets

High-quality datasets can often improve the quality of model training and prediction accuracy. They are essential but easily overlooked and need more systematic introduction materials. This section first summarizes the available crop detection and counting datasets, as shown in Table 2. Afterward, because there are many kinds of crops, and the characteristics differ, we introduce the three most widely used datasets of different crop types with the most abundant data.

4.1.1. GWHD

David et al. (2020) made public the Global wheat head detection dataset(GWHD), which has 188,445 wheat ears labeled in 11 sub-datasets of 4698 RGB images collected from 7 countries and different platforms for different periods through international collaboration. All sub-datasets from Europe and North America were used as training datasets, corresponding to 3422 images, accounting for 73% of the images in the entire GWHD dataset. The test dataset includes all images from Australia, Japan, and China, representing 1276 images to evaluate the model performance, including robustness to invisible images. In 2021, they rechecked and labeled the 2020 dataset, adding nearly 2000 photos from 14 countries and 120,000 newly labeled wheat ears. GWHD_2021 contains 6500 images of 1024*1024 pixel size and 275,000 wheat ears, and the sub-dataset was increased from 18 to 47, which is larger and less noisy than the GWHD_2020 is more extensive and less noisy, and covers different growth stages with a wide range of genotypes as shown in Fig. 9.

4.1.2. MinneApple

MinneApple (H[ii]ani et al., 2020) can be used for apple detection, segmentation, and counting tasks in complex orchard environments, including 1000 images with a resolution of 1280*720 in 17 scenes and over 41,000 instances of apple annotation. Use polygons to label each object instance to assist in accurate object detection, positioning, and segmentation. The object instance is relatively tiny compared to the image size, and a single image may contain 1–120 objects. MinneApple provides two annotated datasets to train block-based counting methods; As shown in Fig. 9, one dataset contains green, and one contains red apples, totaling 13,000 image blocks. Additionally, it randomly selected 4500 patches without apples as negative examples. The test dataset consists of 4 image sequences with 2874 image blocks. Two test datasets contain red apples, one contains green, and one contains mixed colors; The fourth dataset is composed of images collected using longer shooting distances to test the algorithm's generalization ability for low-resolution fruit counting.

4.1.3. MTC

Maize Tassels Counting (Lu et al., 2017) can be used for the counting task of maize tassels, which have different resolutions and sizes of 3648*2736, 4272*2848, and 3456*2304 from 361 images, totaling 13,562 tags. MTC includes 16 independent time series image sequences collected from four experimental corn sites in China from 2010 to 2015. To avoid duplicate sampling, it selected 8–45 images from each sequence based on the specific situation presented by each sequence. The sequence with variable weather conditions selected more images. Use a CCD digital camera to capture images from a vertical perspective of 5 meters high. The training and validation sets use the same image sequence, while the testing set uses different image sequences to achieve reasonable evaluation. Among them, 186 images were used to construct the training and validation sets, and 175 were used for the testing set.

Fig. 9. Examples of images from GWHD_2021, MinneApple and MTC.

Table 2
Publicly available datasets that can be used for crop counting studies.

Name	Object	Sizes	Numbers	URL	References
Date fruit dataset	Date fruit	224*224	8079	Link	Altaheri et al. (2019)
GWHD	Wheat ears	500*500, 250*250	6515	Link	David et al. (2020)
acfr-multifruit-2016	Apple, mango, almond	202*308, 500*500, 300*300	3702	Link	Bargoti and Underwood (2017)
Bayberry Tree Dataset	Bayberry tree	5472*3648	3690	Link	Wang and Luo (2019)
PRL	Arabidopsis thaliana rosette	256*256	3285	Link	Ubbens et al. (2018)
YOLOPOD	Soybean pod	4752*3168, 5184*2916, 3960*2392	2243	Link	Xiang et al. (2023)
MrMT	Maize tassels	3648*2736	1968	Link	Yu et al. (0000)
WSC	Wheat ears	1216*912	1764	Link	Xiong et al. (2019)
MangoYOLO	Mango	500*500, 612*512, 2448*2048, 6000*4000, 1920*1080	1730	Link	Koirala et al. (2019a)
OrangeSort	Orange	1920*1080	1465	Link	Zhang et al. (2022)
MinneApple	Apple	1280*720	1000	Link	H¨ani et al. (2020)
wGrapeUNIPD-DL	White grape bunch	4288*2848, 4608*3456, 4032*3024,	373	Link	Sozzi et al. (2022)
MTC	Tassel	3648*2736, 4272*2848, 3456*2304	361	Link	Lu et al. (2017)
WEDD	Wheat ears	6000*4000	235	Link	Madec et al. (2019)
SHC	Sorghum heads	1154*1731	92	Link	Lu and Cao (2020)
Grapes dataset	Grapes	2592*2048	29(videos)	Link	Ariza-Sentís et al. (2023)

4.2. Metrics

The most critical part of the research based on deep learning is the evaluation of the model. Only by training the learned model and objectively evaluating the performance of the model can we make better practical decisions. It is crucial to select appropriate metrics to evaluate the performance of the object detection algorithm used in the crop counting task. This section presents the evaluation metrics separately into detection and counting metrics.

4.2.1. Detection metrics

Precision, Recall are widely used to evaluate the performance of detection models on training and test sets. Precision is used to evaluate the correctness of the prediction results. Recall is used to evaluate the number of correct samples in the prediction results. The performance of the model is excellent when both metrics are high. F1-score is used to balance precision and Recall, and detectors with higher F1-score are now proven to have good Recall and accuracy (Padilla et al., 2020). The specific formula is as follows:

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

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

$$F1-\ \text{score}\ = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}\ +\ \text{Recall}} \times 100\% \tag{3}$$

TP indicates the number of positive samples correctly detected as positive samples, FP indicates the number of negative samples incorrectly detected as positive samples, and FN indicates the number of positive samples incorrectly detected as negative samples. Higher F1-Score values are associated with better results. (0≤ F1-score≤ 1)

4.2.2. Counting metrics

The results of crop counting were analyzed mainly from counting error and goodness of fit, including mean absolute error(MAE), root mean square error(RMSE), and coefficient of determination(R^2). MAE indicates the average absolute error between the predicted and actual values; the smaller the error, the smaller the value. RMSE is the square root of the ratio of the square of the deviation of the predicted value to the actual value to the number of observations, which measures the deviation between the predicted value and the actual value and is more sensitive to outliers in the data. R^2 is a relative indicator of the fit of the regression line to the actual value. It takes values in the range of (0, 1) and can be expressed as a percentage. If R^2 is closer to 1, it means the better fit of the experimental results. The specific formula is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{n}\left(t_k - c_k\right)^2} \tag{4}$$

$$\mathrm{rRMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{n}\left(\frac{t_k - c_k}{t_k}\right)^2} \tag{5}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{k=1}^{n}\left|t_k - c_k\right| \tag{6}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{k=1}^{n}\left(t_k - c_k\right)^2}{\sum_{k=1}^{n}\left(t_k - \overline{t_k}\right)^2} \tag{7}$$

Where M denotes the number of images in the test set used for the experiment; t_k denotes the predicted number of crops in the kth image; and c_k denotes the actual number of object crops in the kth image. RMSE can reflect the model's accuracy, and the smaller the value, the higher the accuracy, while MAE can reflect the error of the predicted value. The performance of the counting models can be evaluated by comparing multiple evaluation metrics of each object counting model.

5. Problems and improvement strategies

Crop counting faces many challenges. With the growth of crops, the shape and size change significantly, the appearance change caused by different varieties, the light change caused by different weather conditions, the occlusion between leaves and crops, and the clutter background caused by weeds. These problems have significantly increased the difficulty of crop counting tasks. This paper classifies these problems into data scarcity, multi-scale detection and counting difficulty, and severe occlusion in complex scenes. This section analyzes the causes of these three problems and summarizes the corresponding solutions based on deep-learning object detection.

5.1. Data scarcity

There are few publicly available crop datasets in crop detection and counting tasks, and the crop categories are also single. The data for crop detection tasks are also very insufficient. In the current development of deep learning technology, supervised learning remains mainstream. The accuracy of the algorithm model obtained by this network training method is greatly affected by the quality and reliability of the training data. In contrast, the acquisition of the data itself and manual labeling often require a lot of time and labor costs. In many scenarios, a large amount of data is not available.

Deep learning has shown remarkable capabilities in learning image features and offers many opportunities for agricultural automation. Deep neural networks typically require extensive and diverse training datasets to learn generalizable models. However, due to the diversity of crops, growing seasons, and climate change, collecting and annotating large training samples from field crops and greenhouses is expensive and complex, making this requirement challenging for applying agricultural automation systems. Data augmentation and transfer learning are the most frequently used methods to alleviate data scarcity.

5.1.1. Data augmentation

This method artificially expands the dataset using label transformation with strong generalization abilities to expand the training data. Usually, different transformation combinations are used as data enhancement to make the model achieve the desired invariance and robustness. Data augmentation, which allows limited data to generate more data value without substantially increasing data, has become the mainstream method for optimizing network training strategies (Shorten and Khoshgoftaar, 2019; Yang et al., 2022b; Naveed et al., 2021; Zhong et al., 2020). Data augmentation methods can be divided into the following categories:

Basic image processing. The operations include geometric transformation, flipping, color channel, clipping, rotation, translation, adding noise, and color space conversion. Most of these operations are directly operated on the image, which is simple and easy to implement.

Image erase. The specific operation is to replace the pixel values of one or more sub-regions in the image with constant or random values, which can ensure that the network focuses on the entire image rather than one part of it.

Image mixing. Mixing two or more images or sub-regions of the image into one image makes the representation of low-level features, such as lines and edges in the training data, more robust.

In crop counting applications, data enhancement methods are essential. In counting sorghum leaves, Xie et al. (2023) adopted flipping, cutting, rotating, translating, and other operations to ensure the richness and diversity of training data. In image acquisition, the intensity and angle of sunlight on the object change significantly. Li et al. (2021b) used the data denoising method to remove the noise points in the acquired image and reduce the influence of noise on the recognition results.

Parvathi and Selvi (2021) used three randomly selected values for brightness adjustment in addition to the primary image processing method to improve the quality of the experimental data set. Then they added the newly adjusted image to the training data set to ensure that a clear view of the object edge could be obtained during manual labeling, improving the detection model's robustness. To improve the robustness of the detection model, Yang et al. (2021) not only converted the image to different degrees of brightness but also increased the contrast of the wheat image by 1.2 times and weakened it by 0.8 times so that the clarity, grayscale, and texture details of the wheat images were better expressed.

Considering that illumination, color, shading, and exposure can vary greatly depending on the acquisition time, method, and hardware conditions, Brown and Sukkarieh (2021) acquired daytime and nighttime crop images to create two datasets. Li and Wu (2022) and Liu et al. (2023) used Mosaic-8, the main idea of which is to crop, arrange randomly, and randomly scale eight images and then combine them into one image, with the benefit of enriching the image background and increasing the amount of data in the sample. At the same time, some random noise is reasonably introduced to improve the recognition ability of the network model for small samples in images and enhance the model's generalization ability.

5.1.2. Transfer learning

The performance of CNN models depends significantly on the size of the dataset. The richer the data, the better the CNN model's ability to extract image features. However, not everyone can obtain enough data. In this case, training a model from scratch is tough. Transfer learning

offers a more straightforward and faster approach. Before starting training, the backbone of the CNN model is pre-trained on a vast dataset. ImageNet (Deng et al., 2009) is a commonly used dataset for transfer learning. It contains over 14 million familiar images, which can provide ample material for CNN training. The pre-trained backbone is more sensitive to the features of the images. The trained backbone network is then transferred back to the model, and all parts of the model are fine-tuned using experimental data. In this way, a good CNN model is trained with a small amount of data. Many image datasets share data, which allows better learning of low-level features and makes transfer learning methods effective (Suh et al., 2018; Espejo-Garcia et al., 2020).

There is a considerable feature difference between ImageNet and wheat datasets. Due to these differences, the backbone network pre-trained on the ImageNet dataset does not perceive wheat ear features strongly. This direct transfer learning method cannot give good results in counting. So Wang et al. (2021) specially constructed a dataset for the pre-training backbone. For a better description, the dataset is defined as D1, and the wheat ear dataset is defined as D0. Dataset D1 consists of D0 and non-wheat data. They use the D1 to train the classification task of the EffcientDet-D0 backbone with a fully connected layer and a classification layer.

The goal of classification is to distinguish whether the image is wheat. It not only requires the trunk to be sensitive to simple features but also needs to have a strong perception of the high-level semantic features of wheat ears. Li et al. (2021b) considered many types of wheat ears in different growth environments and proposed models Faster R-CNN and RetinaNet for regional wheat recognition. The idea of transfer learning is integrated into the proposed model to explore its performance in different training samples.

5.1.3. Active learning

In supervised learning, researchers first collect many labeled samples and then train models. The effort and cost of manually labeling samples are very high. Active learning, on the other hand, is a method dedicated to augmenting the network by selecting the least possible number of labeled samples from unlabeled data, reducing the number of labeled samples, and speeding up the model training process while obtaining the highest possible detection results (Li et al., 2021a). Only a small amount of labeled data, a large amount of unlabeled data, and an Oracle (e.g., human) are used. Active learning has been applied in many counting scenarios based on object detection (Kellenberger et al., 2019; Duporge et al., 2021; Norouzzadeh et al., 2019). The basic process of the active learning algorithm is as follows (Norouzzadeh et al., 2019; Kellenberger et al., 2019):

(1) Training initial model: Use labeled data to train the initial model.
(2) Select unlabeled samples: Develop a strategy to select the N samples with the highest value from the unlabeled sample set. The strategy of selecting samples is crucial.
(3) Sample labeling: Annotate the selected N samples by Oracle.
(4) Updating the model: Train and update the model using Labeled samples.
(5) Repeat: Repeat the above process until the model meets the predetermined training objectives.

The yield of cereal such as sorghum depends on the distribution of spikes, and it is even more time-consuming and labor-intensive to label. In order to reduce the labeling effort, (Ghosal et al., 2019) designed an active learning method for counting of sorghum heads from UAV-based images. Unlike the method of selecting the most data samples for labeling in general active learning, they use the improved semi-training model of RetinaNet to label the completely random data samples and then manually verify the labels. Experiments show that the workload of manual labeling can be significantly reduced without affecting the performance of the final model. In order to further reduce manual participation, (Petti and Li, 2022) used the open-source CVAT to verify and correct the labeled samples. The number of images annotated per cycle was doubled.

5.1.4. Synthetic datasets

Deep learning requires many datasets, and collecting and annotating these datasets is very time-consuming. Synthetic data is a meaningful way to solve data problems. It can generate artificial data from scratch and use advanced data manipulation techniques to generate novel and diverse training samples. Simulation data can be safely per the requirements, customized data generation and annotation, and from the massive simulation scenarios to obtain any data volume. The production cycle is short and low-cost, avoiding the time-consuming and labor-intensive manual annotation. The simulation data can be used to train the model, and then the trained model is tested on the actual data (Nowruzi et al., 2019; Nikolenko, 2019). At present, the simulation datasets are widely used in deep learning (Li et al., 2022a; Arnold et al., 2022; Xu et al., 2022b).

Rahnemoonfar and Sheppard (2017) used green and brown solid circles to fill the blank image to simulate the background and tomato plant, then blurred with Gaussian filters and draw many random-sized circles at random locations on the image to represent the tomato. The proposed algorithm is first trained on synthetic data and tested on actual data. The experimental results show that the model can achieve more than 90% accuracy on both real and synthetic images and effectively reduce the interference of occlusion, light, and fruit size.

Rahim and Mineno (2020) proposed a new method of expanding the training dataset using synthetic images that preserve the data objects' background context and texture. A synthetic dataset of 1800 images was generated using a reference dataset and applying image processing techniques.

5.2. Difficulties in small object counting

The definition of small objects varies under each type of task, and its definition usually depends on the application scenario's requirements. The MS COCO (Lin et al., 2014) dataset proposed by Microsoft defines an absolute scale for small objects, which is considered small when the object area is less than 32*32 pixels; the other is a relative scale definition, which means that an object is small when the length and width of the object size are 0.1 of the original image size.

Small object resolution is low, and the proportion of pixels is small, limiting the adequate feature information extracted during detection and counting. This is the root cause of poor detection of small objects (Lim et al., 2019). At the same time, detecting and counting small objects makes it easier to receive the influence of environmental noise. The convolutional neural network has a bottleneck in the feature extraction of small objects. In order to increase the receptive field, the current object detection models will undergo several down-sampling operations. However, due to the fuzzy edge information and less semantic information of the small object, the information of the small object is lost seriously. It even cannot be passed into the object detector. In addition, compared with the large object, the small object is more sensitive to the offset of the prediction boxes and prone to false detection; when the scale of the prediction boxes is too large due to the small area of small objects, even if the small object is in the box, it may still cause missed detection because the IoU does not reach the threshold. Furthermore, when the preset scales of the prediction boxes are too close, the spatial difference after down-sampling cannot be guaranteed, resulting in the neglect of small crops (Chen et al., 2020; Liu et al., 2021; Mudassar and Mukhopadhyay, 2019).

Therefore, detecting crop flowers and fruits is very challenging for small object detection in the specific application of crop detection and counting. Xu et al. (2020) designed a simple and effective method called MHW-PD, which allows the identification and counting of rice panicles without depending on the number of panicles in the scene. Based on quantitative analysis of the relationship between receptive field, input image size, and average panicle dimension, MHW-PD gives a dynamic strategy of selecting an appropriate feature learning network and constructing an adaptive multi-scale hybrid window(MHW), which

maximizes the richness of features. In addition, a fusion algorithm was introduced to remove the repeated object to obtain the final number.

Yang and Deng (2020) used YOLOv4 as the baseline. The backbone is enhanced by adding a dual-space pyramidal pooling(SPP) network to improve feature learning and increase the perceptual field of the convolutional network. Multilevel features are obtained from the multipath neck using top-down and bottom-up strategies. Finally, they used the head structure of YOLOv3 to get the prediction boxes. And Zhu et al. (2022) proposed an improved YOLOv4, combined with the Mobilenetv3 network, which can reduce the model and improve the detection speed. They used adaptive spatial feature fusion(ASFF), which can enhance the multi-scale feature fusion ability of the network. In addition, the K-means algorithm and linear scale scaling are used to optimize the generation of pre-selected boxes, accelerate the model's training speed, and improve detection accuracy. The results show that the improved YOLOv4 model can effectively overcome the noise in the orchard environment and achieve the effect of efficient and accurate detection of fruit trees.

Li and Wu (2022) proposed an improved YOLOv5 algorithm based on shallow feature layers. Furthermore, it adds quadruple downsampling to the feature pyramid to increase the perceptual domain and improve small object detection performance. Tu et al. (2020) found that when Faster R-CNN detects small objects because RPN only processes the feature map obtained by the last convolutional layer, it omits the critical low-resolution feature map for small object detection, resulting in poor performance of small objects. In order to improve the robustness of small objects, they designed a multi-scale feature extractor to extract low-level features and high-level features from the image to enhance global context and local information.

Liu et al. (2023) used a variety of ways to improve the performance of multi-scale detection and counting. First, the filling operation is used to randomly select the image, cut the object, and then randomly paste the object onto other images. Select the number of pastes to be 3 to ensure the paste process does not overlap with the currently labeled objects. In this way, the number of positive samples of small objects in a single image is increased during the training process, which improves the problem of poor detection accuracy in scenes with few small objects. The data augmentation strategy dramatically improves the model's universality, significantly reduces the problem of insufficient sample size and unbalanced sample distribution, and significantly improves the detection effect of small crops. Secondly, in the process of extracting features, deep separable convolution is combined with increasing the small object detection layer, which improves the extraction effect of PANet on small object features.

5.3. Occlusion in complex scenes

Previous studies have shown that most crop detection methods are based on shallow features such as color and texture extracted by machine learning methods and have achieved good results. However, these features lack robustness, and the above methods are difficult to meet the crop detection and counting in natural scenes. Counting the number of crops in images under natural light is a critical way to evaluate crop yield, which is significant to modern intelligent agriculture.

Although object detection and counting technology have made significant progress in recent years, occlusion is still one of the most critical challenges in crop detection and counting due to the complexity of the field scene. Generally speaking, occlusion can be divided into inter-class occlusion and intra-class occlusion. Inter-class occlusion occurs when other classes of objects occlude the object; intra-class occlusion occurs when objects of the same category (Wang et al., 2018).

In the case of occlusion, the main reasons for the missed detection and false detection of the occluded object are:

(a) The object is incomplete, and too few features can be extracted.
(b) Overlapping objects usually have highly similar characteristics, and the model is difficult to determine whether they belong to different individuals.
(c) The post-processing method of non-maximum suppression(NMS) directly discards objects with low confidence in the overlapping area, which can easily lead to missed detection of occluded objects.

Because of the serious occlusion problem in complex scenes, there are the following improvement methods:

5.3.1. Data augmentation

Wang et al. (2021) proposed a Random-Cutout method that can select some rectangles and erase them according to the number and size of the crop in the image to simulate occlusion in real counting images.

5.3.2. Attention module

The added attention focuses more on the region of interest so that the network pays more attention to the local features of the image. CBAM is a lightweight attention module proposed by Woo et al. (2018), which can perform attention operations in the spatial and channel dimensions. It consists of two independent sub-modules, the channel attention module(CAM) and the spatial attention module(SAM). CAM learns the weights of different channels and then multiplies them with the weights to enhance the attention to the critical channel domain. SAM focuses on the position of the object. As shown in Fig. 10, the input features first pass through the channel attention module, and the weighted results are then passed through the spatial attention module and finally weighted to obtain the results.

Wang et al. (2021) made an improved EfficientDet-D0 object detection model for wheat spike counting, focusing on solving the occlusion problem. The convolution block attention module(CBAM) is introduced into the EfficientDet-D0 model after the backbone network so that the model pays more attention to the wheat ear while refining the features and suppressing other useless background information. Yang et al. (2021) designed an improved YOLOv4 based on CBAM, taking into account not only spatial attention but also channel attention, to enhance the feature extraction capability of the network by adding a perceptual field module. Different from other studies, Li et al. (2023b) proposed a UAV RGB image-based enhancement method using YOLOv5 in combination with CBAM. This method successfully solved the problem of difficulty in extracting sufficient features due to sticking and mutual occlusion.

5.3.3. Improved loss function

Loss functions are used to optimize models, but the classification and localization loss functions used in most detection models are insufficient for accurate localization. By designing a better loss function without changing the model structure, the localization capability can be greatly improved and help to reduce the error due to occlusion (Wu et al., 2022).

Wang et al. (2018) proposed a new type of bounding box regression loss called the rejection loss. This loss comprises two components: an attraction term to the detection object and a repulsion term to other surroundings. The repulsion term prevents the proposal from moving toward the surroundings, leading to more robust object localization. Detectors trained by rejection loss outperform state-of-the-art methods, with significant performance improvements in the presence of occlusions. Zhang et al. (2018) designed a new aggregation loss to force the region to be closely approached and localized with the corresponding object. A new partially occlusion-aware region of interest(PORoI) pooling unit is also designed to replace the RoI pooling layer. The detector is trained in an end-to-end manner. Xiang et al. (2023) proposed the YOLO POD method based on the YOLOX framework. A block for predicting the number of pods was added to YOLOX. The loss function was modified to construct a multi-task model.

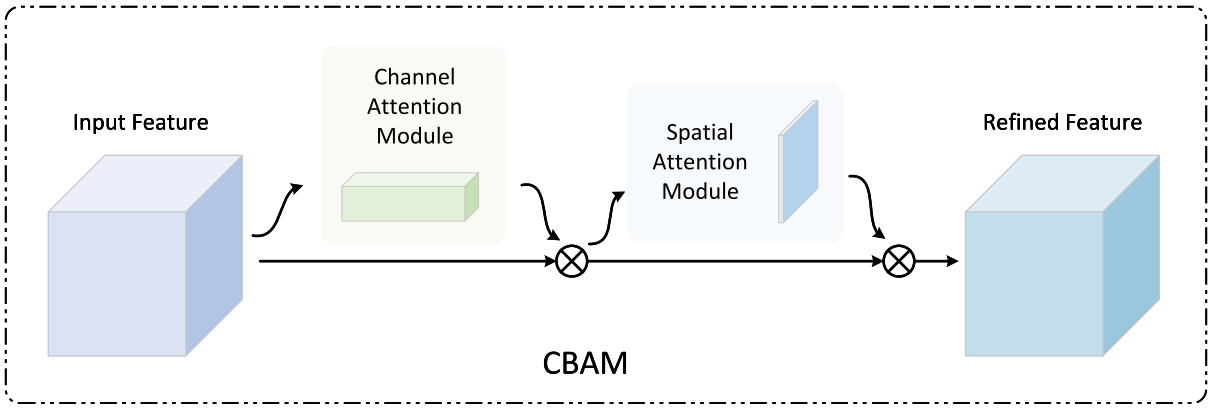

Fig. 10. The structure of CBAM.

5.3.4. Add contextual information

When the object occupies only a tiny part of the image, the information obtained directly from the fine-grained local area is minimal. General object detectors usually ignore many contextual features outside these local regions. Each object always exists in a specific environment or coexists with other objects (Chen et al., 2020). Many studies have proposed to increase context information to reduce the impact of occlusion.

Xiong et al. (2019) added context extension to TasselNet and established the TasselNetv2 model. By adding visual context to local blocks, the counting performance of local regression networks can be significantly improved. At the same time, such a context can be part of the Receptive field without increasing the model capacity. Implementing the TasselNetv2 as full convolution can significantly accelerate training and inference speed by reducing redundant calculations. Osco et al. (2021) used global and local context modules called PPM (Han et al., 2022) to make them scale invariant and help the network process multiple canopy sizes. The PPM module receives the generated feature map as input and applies four parallel pooling layers. It is a module with hierarchical global priority and different scale information between sub-regions.

Yang and Deng (2020) proposed a object detection algorithm for GC-YOLOv3 based on YOLO. Firstly, a cascade model with learnable semantic fusion is designed between the feature extraction network and the feature pyramid network to improve the detection accuracy using global contextual blocks. Secondly, the feature maps of three scales are combined to filter out the information to be retained. Finally, it used the global self-attention mechanism to highlight the useful information of the feature maps while suppressing irrelevant information. Instead of using multiple convolutional layers to extract spatial and channel features, the global context block uses only one convolutional layer and one ReLU activation function to make the attention more nonlinear. In this way, the global context block reduces a sizeable computational burden relative to previous self-attentive modules and makes classification and regression more accurate.

5.3.5. Others

Based on the YOLOv3 model, Liu et al. (2020) used the YOLO-Tomato detector that replaces the rectangular detection boxes with circular detection boxes for tomato detection. Using circular detection boxes can better match the shape of the tomato, provide more accurate IoU for the NMS process, and reduce the predicted coordinates.

6. Summary and prospect

With the continuous development of computer vision technology, the application of object detection in agriculture is becoming more and more extensive, especially in crop counting. Object detection technology has become a very effective method. This paper introduces the application of object detection methods based on deep learning in crop counting tasks, discusses the relevant datasets and performance metrics, and summarizes the challenges and feasible solutions in crop counting. The research of crop counting is still constantly enriched and developed. Here we discuss some future trends and possible research directions:

Establish high-quality datasets. This can be solved by developing new data collection methods and creating large-scale labeled datasets that can be used to train deep learning models. It can also process and expand existing data through data enhancement and other measures and use data synthesis (Barth et al., 2018; Di Cicco et al., 2017; Dyrmann et al., 2018) to establish datasets. Active learning is a machine learning method that drastically reduces the amount of training data, improves the learning efficiency and accuracy of the model, and reduces the cost of labeled data by allowing the algorithm to select the most representative and critical samples for labeling actively (Kao et al., 2019; Yuan et al., 2021; Yu et al., 2022). Active learning also has a wide range of application possibilities in counting problems.

Improve the universality and generalization of the model. How to extend the same algorithm to different objects and still maintain excellent performance is a concern. The current object detection algorithm needs to train different crops. The future development direction is to develop further adaptive object detection technology (Marsden et al., 2018), which can automatically adapt to different crop types and improve the application range. Combined with the practical application scenarios of crop detection and counting, a more comprehensive object detection and counting model is established to improve detection accuracy and counting efficiency.

Improve the accuracy and efficiency of counting. As shown in Table 1, the evaluation metrics used in the studies are all related to accuracy. Current studies have focused on the accuracy of crop counting models, while other factors, such as the efficiency of the algorithms and the actual execution time, may have yet to be fully considered. Object detection algorithms can already achieve high counting accuracy, but there are still some errors and slower speed problems in practical applications. The speed cannot be improved at the expense of accuracy, nor can the pursuit of accuracy give up speed. The future development direction is establishing models with balanced accuracy and speed. Therefore, in addition to focusing on the model's accuracy, the real-time performance and other relevant factors need to be considered to better meet the needs of practical applications (Koirala et al., 2019b).

Better use of feature fusion. Complex features during crop growth are the focus of our research. The same object presents complex and diverse features at different growth stages and from different angles. During crop growth, light, weather, terrain, and ground objects can affect the count. Therefore, more and better crop growth features can be built based on these factors by fusing multi-source information such as terrain and drawing on the methods of classical algorithms

such as ReLU and Faster R-CNN. Its integration of multiple data is critical for improving the performance of deep learning models in crop counting (Albahar, 2023).

Detect and count the video datasets. Image-based counting has achieved good results, but it requires human control of the shooting area to avoid overlapping areas in crop images. The acquisition of video is more accessible than the acquisition of images (Ge et al., 2022; Yang et al., 2022a; He et al., 2022b), and detecting video can help quickly realize crop counting.

CRediT authorship contribution statement

Yuning Huang: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft. **Yurong Qian:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. **Hongyang Wei:** Supervision, Writing – review & editing. **Yiguo Lu:** Data curation, Visualization. **Bowen Ling:** Data curation, Visualization, Investigation. **Yugang Qin:** Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., S¨usstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.

Afonso, M., Fonteijn, H., Fiorentin, F.S., Lensink, D., Mooij, M., Faber, N., Polder, G., Wehrens, R., 2020. Tomato fruit detection and counting in greenhouses using deep learning. Front. Plant Sci. 11, 571299.

Albahar, M., 2023. A survey on deep learning and its impact on agriculture: Challenges and opportunities. Agriculture 13 (3), 540.

Altaheri, H., Alsulaiman, M., Muhammad, G., Amin, S.U., Bencherif, M., Mekhtiche, M., 2019. Date fruit dataset for intelligent harvesting. Data in brief 26, 104514.

Ariza-Sentís, M., Vélez, S., Valente, J., 2023. Dataset on UAV RGB videos acquired over a vineyard including bunch labels for object detection and tracking. Data in Brief 46, 108848.

Arnold, E., Dianati, M., de Temple, R., Fallah, S., 2022. Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors. IEEE Trans. Intell. Transp. Syst..

Auria, L., Moro, R.A., 2008. Support vector machines (SVM) as a technique for solvency analysis. DIW Berlin discussion paper.

Bao, W., Xie, W., Hu, G., 2023. Wheat ear counting method in UAV images based on TPH-YOLO. Trans. Chin. Soc. Agric. Eng. 39 (155–161), http://dx.doi.org/10.11975/j.issn.1002-6819.202210020.

Bao, W., Zhang, X., Hu, G., et al., 2020a. Estimation and counting of wheat ears density in field based on deep convolutional neural network. Trans. Chin. Soc. Agric. Eng. 36 (186-193+323).

Bao, W., Zhang, T., Hu, G., et al., 2020b. Wheat ears counting in natural scenes based on multi-scale and multi-direction decomposition. J. Anhui Univ.(Nat. Sci. Ed.) 44 (20–27).

Bargoti, S., Underwood, J., 2017. Deep fruit detection in orchards. In: 2017 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 3626–3633.

Barth, R., IJsselmuiden, J., Hemming, J., Van Henten, E.J., 2018. Data synthesis methods for semantic segmentation in agriculture: A capsicum annuum dataset. Comput. Electron. Agric. 144, 284–296.

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Brown, J., Sukkarieh, S., 2021. Dataset and performance comparison of deep learning architectures for plum detection and robotic harvesting. arXiv preprint arXiv:2105.03832.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer, pp. 213–229.

Chen, S.W., Shivakumar, S.S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C.J., Kumar, V., 2017. Counting apples and oranges with deep learning: A data-driven approach. IEEE Robot. Autom. Lett. 2 (2), 781–788.

Chen, G., Wang, H., Chen, K., Li, Z., Song, Z., Liu, Y., Chen, W., Knoll, A., 2020. A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal. IEEE Trans. Syst. Man Cybern. Syst. 52 (2), 936–953.

Chen, Y., Xu, H., Zhang, X., Gao, P., Xu, Z., Huang, X., 2023. An object detection method for bayberry trees based on an improved yolo algorithm. Int. J. Digit. Earth 16 (1), 781–805.

Chen, Y., Zhang, X., Chen, X., 2022. Identification of navel orange trees based on deep learning algorithm YOLOv4. Sci. Surv. Mapp. 47 (135-144+191).

Ciampi, L., Carrara, F., Amato, G., Gennaro, C., 2022. Counting or localizing? Evaluating cell counting and detection in microscopy images. In: VISIGRAPP (4: VISAPP). pp. 887–897.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1. CVPR'05, Ieee, pp. 886–893.

Darwin, B., Dharmaraj, P., Prince, S., Popescu, D.E., Hemanth, D.J., 2021. Recognition of bloom/yield in crop images using deep learning models for smart agriculture: a review. Agronomy 11 (4), 646.

David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgessner, N., Ishikawa, G., Nagasawa, K., Badhon, M.A., et al., 2020. Global wheat head detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods. Plant Phenomics.

Delplanque, A., Foucher, S., Théau, J., Bussière, E., Vermeulen, C., Lejeune, P., 2023. From crowd to herd counting: How to precisely detect and count african mammals using aerial imagery and deep learning? ISPRS J. Photogramm. Remote Sens. 197, 167–180.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Di Cicco, M., Potena, C., Grisetti, G., Pretto, A., 2017. Automatic model based dataset generation for fast and accurate crop and weeds detection. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 5188–5195.

Du, Y., YiCheng, C., Tan, C., et al., 2019. Field wheat ears counting based on superpixel segmentation method. Sci. Agricul. Sinica 52 (21–33).

Duporge, I., Isupova, O., Reece, S., Macdonald, D.W., Wang, T., 2021. Using very-high-resolution satellite imagery and deep learning to detect and count african elephants in heterogeneous landscapes. In: Pettorelli, N., Buchanan, G. (Eds.), Remote Sen. Ecol. Conserv. 7 (3), 369–381.

Dyrmann, M., Christiansen, P., Midtiby, H.S., 2018. Estimation of plant species by classifying plants and leaves in combination. J. Field Robotics 35 (2), 202–212.

Egi, Y., Hajyzadeh, M., Eyceyurt, E., 2022. Drone-computer communication based tomato generative organ counting model using YOLO V5 and deep-sort. Agriculture 12 (9), 1290.

Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Fountas, S., Vasilakoglou, I., 2020. Towards weeds identification assistance through transfer learning. Comput. Electron. Agric. 171, 105306.

Falahat, S., Karami, A., 2022. Maize tassel detection and counting using a YOLOv5-based model. Multimedia Tools Appl. 1–18.

Fan, Z., Zhang, H., Zhang, Z., Lu, G., Zhang, Y., Wang, Y., 2022. A survey of crowd counting and density estimation based on convolutional neural network. Neurocomputing 472, 224–251.

Farjon, G., Krikeb, O., Hillel, A.B., Alchanatis, V., 2020. Detection and counting of flowers on apple trees for better chemical thinning decisions. Precis. Agric. 21, 503–521.

Gao, F., Fang, W., Sun, X., Wu, Z., Zhao, G., Li, G., Li, R., Fu, L., Zhang, Q., 2022a. A novel apple fruit detection and counting methodology based on deep learning and trunk tracking in modern orchard. Comput. Electron. Agric. 197, 107000.

Gao, Y., Sun, Y., Li, B., 2022b. Estimating of wheat ears number in field based on RGB images using unmanned aerial vehicle. J. Agric. Sci. Technol. 24 (103–110).

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., 2017. A review on deep learning techniques applied to semantic segmentation. arXiv:1704.06857.

Ge, Y., Lin, S., Zhang, Y., Li, Z., Cheng, H., Dong, J., Shao, S., Zhang, J., Qi, X., Wu, Z., 2022. Tracking and counting of tomato at different growth period using an improving YOLO-deepsort network for inspection robot. Machines 10 (6), 489.

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.

Ghosal, S., Zheng, B., Chapman, S.C., Potgieter, A.B., Jordan, D.R., Wang, X., Singh, A.K., Singh, A., Hirafuji, M., Ninomiya, S., et al., 2019. A weakly supervised deep learning framework for sorghum head detection and counting. Plant Phenomics.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2015. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 38 (1), 142–158.

Giuffrida, M.V., Doerner, P., Tsaftaris, S.A., 2018. Pheno-deep counter: A unified and versatile deep learning architecture for leaf counting. Plant J. 96 (4), 880–890.
Glenn, J., Ayush, C., Alex, S., et al., 2022. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. Zenodo, http://dx.doi.org/10.5281/zenodo.7347926.
Gomaa, A., Minematsu, T., Abdelwahab, M.M., Abo-Zahhad, M., Taniguchi, R.-i., 2022. Faster CNN-based vehicle detection and counting strategy for fixed camera scenes. Multimedia Tools Appl. 81 (18), 25443–25471.
Gongal, A., Silwal, A., Amatya, S., Karkee, M., Zhang, Q., Lewis, K., 2016. Apple crop-load estimation with over-the-row machine vision system. Comput. Electron. Agric. 120, 26–35.
Han, G., Huang, S., Ma, J., He, Y., Chang, S.-F., 2022. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 1. pp. 780–789.
H¨ani, N., Roy, P., Isler, V., 2020. MinneApple: a benchmark dataset for apple detection and segmentation. IEEE Robot. Autom. Lett. 5 (2), 852–858.
Hao, S., Zhou, Y., Guo, Y., 2020. A brief survey on semantic segmentation with deep learning. Neurocomputing 406, 302–321.
Hassan, S.I., Alam, M.M., Zia, M.Y.I., Rashid, M., Illahi, U., Su'ud, M.M., 2022. Rice crop counting using aerial imagery and GIS for the assessment of soil health to increase crop yield. Sensors 22 (21), 8567.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.
He, H., Ma, X., Guan, H., Wang, F., Shen, P., 2022a. Recognition of soybean pods and yield prediction based on improved deep learning model. Front. Plant Sci. 13.
He, L., Wu, F., Du, X., Zhang, G., 2022b. Cascade-SORT: A robust fruit counting approach using multiple features cascade matching. Comput. Electron. Agric. 200, 107223.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916.
Jiang, H., Learned-Miller, E., 2017. Face detection with the faster R-CNN. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition. FG 2017, IEEE, pp. 650–657.
Jiang, Y., Li, C., Paterson, A.H., Robertson, J.S., 2019. DeepSeedling: Deep convolutional network and Kalman filter for plant seedling detection and counting in the field. Plant Methods 15 (1), 141.
Jocher, G., Chaurasia, A., Qiu, J., 2023. YOLO by Ultralytics. URL https://github.com/ultralytics/ultralytics.
Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. Comput. Electron. Agric. 147, 70–90.
Kao, C.-C., Lee, T.-Y., Sen, P., Liu, M.-Y., 2019. Localization-aware active learning for object detection. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14. Springer, pp. 506–522.
Karami, A., Quijano, K., Crawford, M., 2021. Advancing tassel detection and counting: annotation and algorithms. Remote Sens. 13 (15), 2881.
Kellenberger, B., Marcos, D., Lobry, S., Tuia, D., 2019. Half a percent of labels is enough: Efficient animal detection in UAV imagery using deep CNNs and active learning. IEEE Trans. Geosci. Remote Sens. 57 (12), 9524–9533, arXiv:1907.07319.
Kestur, R., Meduri, A., Narasipura, O., 2019. MangoNet: A deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. Eng. Appl. Artif. Intell. 77, 59–69.
Koirala, A., Walsh, K., Wang, Z., McCarthy, C., 2019a. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'mangoyolo'. Precision Agric. 20, 1107–1135.
Koirala, A., Walsh, K.B., Wang, Z., McCarthy, C., 2019b. Deep learning – method overview and review of use for fruit detection and yield estimation. Comput. Electron. Agric. 162, 219–234.
Korada, N.K., Kuma, N., Deekshitulu, Y., 2012. Implementation of na¨ıve Bayesian classifier and ada-boost algorithm using maize expert system. Int. J. Inf. Sci. Techn. (IJIST) Vol 2.
Kuwata, K., Shibasaki, R., 2015. Estimating crop yields with deep learning and remotely sensed data. In: 2015 IEEE International Geoscience and Remote Sensing Symposium. IGARSS, IEEE, pp. 858–861.
Li, Y., An, Z., Wang, Z., Zhong, Y., Chen, S., Feng, C., 2022a. V2x-sim: A virtual collaborative perception dataset for autonomous driving. arXiv preprint arXiv:2202.08449.
Li, Y., Fan, Q., Huang, H., Han, Z., Gu, Q., 2023a. A modified YOLOv8 detection network for UAV aerial image recognition. Drones 7 (5), 304.
Li, Y., Fan, B., Zhang, W., Ding, W., Yin, J., 2021a. Deep active learning for object detection. Inform. Sci. 579, 418–433.
Li, H., Lee, W.S., Wang, K., 2016. Immature green citrus fruit detection and counting based on fast normalized cross correlation (FNCC) using natural outdoor colour images. Precision Agriculture 17 (6), 678–697.
Li, J., Li, C., Fei, S., et al., 2021b. Wheat ear recognition based on RetinaNet and transfer learning. Sensors 21 (14), 4845.
Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al., 2022b. YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976.
Li, J., Li, Y., Qiao, J., Li, L., Wang, X., Yao, J., Liao, G., 2023b. Automatic counting of rapeseed inflorescences using deep learning method and UAV RGB imagery. Front. Plant Sci. 14.
Li, Y., Ma, R., Zhang, R., Cheng, Y., Dong, C., 2023c. A tea buds counting method based on YOLOV5 and Kalman filter tracking algorithm. Plant Phenomics 5, 0030.
Li, R., Wu, Y., 2022. Improved YOLO v5 wheat ear detection algorithm based on attention mechanism. Electronics 11 (11), 1673.
Li, Y., Zhang, X., Chen, D., 2018. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1091–1100.
Lim, J.-S., Astrid, M., Yoon, H.-J., Lee, S.-I., 2019. Small object detection using context and attention. arXiv:1912.06319.
Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.
Lin, Z., Guo, W., 2020. Sorghum panicle detection and counting using unmanned aerial system images and deep learning. Front. Plant Sci. 11, 534853.
Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.
Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 21–37.
Liu, T., Chen, W., Wang, Y., Wu, W., Sun, C., Ding, J., Guo, W., 2017. Rice and wheat grain counting method and software development based on android system. Comput. Electron. Agric. 141, 302–309.
Liu, Z., Huang, W., Wang, L., 2019. Field wheat ear counting automatically based on improved K-means clustering algorithm. Trans. Chin. Soc. Agric. Eng. 35 (174–181).
Liu, G., Nouaze, J.C., Touko Mbouembe, P.L., Kim, J.H., 2020. YOLO-tomato: A robust algorithm for tomato detection based on YOLOv3. Sensors 20 (7), 2145.
Liu, Y., Sun, P., Wergeles, N., Shang, Y., 2021. A survey and performance evaluation of deep learning methods for small object detection. Expert Syst. Appl. 172, 114602.
Liu, Q., Zhang, Y., Yang, G., 2023. Small unopened cotton boll counting by detection with MRF-YOLO in the wild. Comput. Electron. Agric. 204, 107576.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.
Lu, H., Cao, Z., 2020. TasselNetV2+: A fast implementation for high-throughput plant counting from high-resolution RGB imagery. Front. Plant Sci. 11, 541960.
Lu, H., Cao, Z., Xiao, Y., Zhuang, B., Shen, C., 2017. TasselNet: counting maize tassels in the wild via local counts regression network. Plant Methods 13 (1), 1–17.
Lyu, S., Li, R., Zhao, Y., Li, Z., Fan, R., Liu, S., 2022. Green citrus detection and counting in orchards based on YOLOv5-CS and AI edge system. Sensors 22 (2), 576.
Machefer, M., Lemarchand, F., Bonnefond, V., Hitchins, A., Sidiropoulos, P., 2020. Mask R-CNN refitting strategy for plant counting and sizing in UAV imagery. Remote Sens. 12 (18), 3015.
Madec, S., Jin, X., Lu, H., De Solan, B., Liu, S., Duyme, F., Heritier, E., Baret, F., 2019. Ear density estimation from high resolution RGB imagery using deep learning technique. Agricult. Forest Meterol. 264, 225–234.
Maheswari, P., Raja, P., Apolo-Apolo, O.E., Pérez-Ruiz, M., 2021. Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques—a review. Front. Plant Sci. 12, 684328.
Marsden, M., McGuinness, K., Little, S., Keogh, C.E., O'Connor, N.E., 2018. People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8070–8079.
Mosley, L., Pham, H., Bansal, Y., Hare, E., 2022. Image-based sorghum head counting when you only look once. arXiv:2009.11929.
Mudassar, B.A., Mukhopadhyay, S., 2019. Rethinking convolutional feature extraction for small object detection. In: BMVC. 234.
Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N.H., Islam, N., 2022. A systematic literature review on crop yield prediction with deep learning and remote sensing. Remote Sens. 14 (9), 1990.
Naveed, H., Anwar, S., Hayat, M., Javed, K., Mian, A., 2021. Survey: Image mixing and deleting for data augmentation. arXiv preprint arXiv:2106.07085.
Neupane, B., Horanont, T., Hung, N.D., 2019. Deep learning based banana plant detection and counting using high-resolution red-green-blue (RGB) images collected from unmanned aerial vehicle (UAV). In: Vadrevu, K.P. (Ed.), PLoS One 14 (10), e0223906.
Nikolenko, S.I., 2019. Synthetic data for deep learning. arXiv:1909.11512.
Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N., Clune, J., 2019. A deep active learning system for species identification and counting in camera trap images. arXiv:1910.09716.
Nowruzi, F.E., Kapoor, P., Kolhatkar, D., Hassanat, F.A., Laganiere, R., Rebut, J., 2019. How much real data do we actually need: Analyzing object detection performance using synthetic and real data. arXiv:1907.07061.
Oh, S., Chang, A., Ashapure, A., Jung, J., Dube, N., Maeda, M., Gonzalez, D., Landivar, J., 2020. Plant counting of cotton from UAS imagery using deep learning-based object detection framework. Remote Sens. 12 (18), 2981.

Osco, L.P., de Arruda, M.d.S., Gonçalves, D.N., Dias, A., Batistoti, J., de Souza, M., Gomes, F.D.G., Ramos, A.P.M., de Castro Jorge, L.A., Liesenberg, V., et al., 2021. A CNN approach to simultaneously count plants and detect plantation-rows from UAV imagery. ISPRS J. Photogramm. Remote Sens. 174, 1–17.

Ott, P., Everingham, M., 2011. Shared parts for deformable part-based models. In: CVPR 2011. IEEE, pp. 1513–1520.

Padilla, R., Netto, S.L., Da Silva, E.A., 2020. A survey on performance metrics for object-detection algorithms. In: 2020 International Conference on Systems, Signals and Image Processing. IWSSIP, IEEE, pp. 237–242.

Pan, Y., Zhu, N., Ding, L., Li, X., Goh, H.-H., Han, C., Zhang, M., 2022. Identification and counting of sugarcane seedlings in the field using improved faster R-CNN. Remote Sens. 14 (22), 5846.

Parico, A.I.B., Ahamed, T., 2021. Real time pear fruit detection and counting using YOLOv4 models and deep SORT. Sensors 21 (14), 4803.

Parvathi, S., Selvi, S.T., 2021. Detection of maturity stages of coconuts in complex background using faster R-CNN model. Biosyst. Eng. 202, 119–132.

Petti, D., Li, C., 2022. Weakly-supervised learning to automatically count cotton flowers from aerial imagery. Comput. Electron. Agric. 194, 106734.

Qian, R., Lai, X., Li, X., 2022. 3D object detection for autonomous driving: A survey. Pattern Recognit..

Qureshi, W.S., Payne, A., Walsh, K.B., Linker, R., Cohen, O., Dailey, M.N., 2017. Machine vision for counting fruit on mango tree canopies. Precis. Agric. 18 (2), 224–244.

Rahim, U.F., Mineno, H., 2020. Data augmentation method for strawberry flower detection in non-structured environment using convolutional object detection networks. J. Agric. Crop Res. 8 (11), 260–271.

Rahnemoonfar, M., Sheppard, C., 2017. Deep count: fruit counting based on deep simulated learning. Sensors 17 (4), 905.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7263–7271.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Rong, J., Zhou, H., Zhang, F., Yuan, T., Wang, P., 2023. Tomato cluster detection and counting using improved YOLOv5 based on RGB-D fusion. Comput. Electron. Agric. 207, 107741.

Saddik, A., Latif, R., Abualkishik, A.Z., El Ouardi, A., Elhoseny, M., 2023. Sustainable yield prediction in agricultural areas based on fruit counting approach. Sustainability 15 (3), 2707.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520.

Shen, L., Su, J., He, R., Song, L., Huang, R., Fang, Y., Song, Y., Su, B., 2023. Real-time tracking and counting of grape clusters in the field based on channel pruning with YOLOv5s. Comput. Electron. Agric. 206, 107662.

Shi, L., Sun, J., Dang, Y., Zhang, S., Sun, X., Xi, L., Wang, J., 2023. YOLOv5s-t: A lightweight small object detection method for wheat spikelet counting. Agriculture 13 (4), 872.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 1–48.

Sozzi, M., Cantalamessa, S., Cogato, A., Kayad, A., Marinello, F., 2022. wGrapeUNIPD-DL: An open dataset for white grape bunch detection. Data Brief 43, 108466.

Suh, H.K., IJsselmuiden, J., Hofstee, J.W., van Henten, E.J., 2018. Transfer learning for the classification of sugar beet and volunteer potato under field conditions. Biosyst. Eng. 174, 50–65.

Sun, H., Li, S., Li, M., et al., 2020. Research progress of image sensing and deep learning in agriculture. Trans. Chin. Soc. Agric. Mach. 51 (1–17).

Sun, J., Yang, K., Chen, C., Shen, J., Yang, Y., Wu, X., Norton, T., 2022. Wheat head counting in the wild by an augmented feature pyramid networks-based convolutional neural network. Comput. Electron. Agric. 193, 106705.

Syal, A., Garg, D., Sharma, S., 2013. A survey of computer vision methods for counting fruits and yield prediction. Int. J. Comput. Sci. Eng. 2 (6), 346–350.

Tan, C., Li, C., He, D., Song, H., 2022. Towards real-time tracking and counting of seedlings with a one-stage detector and optical flow. Comput. Electron. Agric. 193, 106683.

Tu, S., Pang, J., Liu, H., Zhuang, N., Chen, Y., Zheng, C., Wan, H., Xue, Y., 2020. Passion fruit detection and counting based on multiple scale faster R-CNN using RGB-d images. Precis. Agric. 21, 1072–1091.

Ubbens, J., Cieslak, M., Prusinkiewicz, P., Stavness, I., 2018. The use of plant models in deep learning: an application to leaf counting in rosette plants. Plant Methods 14, 1–10.

Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W., 2013. Selective search for object recognition. Int. J. Comput. Vis. 104, 154–171.

Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M., 2022a. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696.

Wang, D., Luo, W., 2019. Bayberry tree recognition dataset based on the aerial photos and deep learning model. J. Global Change Data Discover 3 (3), 290–296.

Wang, Y., Qin, Y., Cui, J., 2021. Occlusion robust wheat ear counting algorithm based on deep learning. Front. Plant Sci. 12, 645899.

Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C., 2018. Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7774–7783.

Wang, X., Yang, W., Lv, Q., Huang, C., Liang, X., Chen, G., Xiong, L., Duan, L., 2022b. Field rice panicle detection and counting based on deep learning. Front. Plant Sci. 2921.

Wei, H., Zhang, Q., Han, J., Fan, Y., Qian, Y., 2022. SARNet: Spatial attention residual network for pedestrian and vehicle detection in large scenes. Appl. Intell. 52 (15), 17718–17733.

Wen, C., Wu, J., Chen, H., Su, H., Chen, X., Li, Z., Yang, C., 2022. Wheat spike detection and counting in the field based on SpikeRetinaNet. Front. Plant Sci. 13.

Weng, Y., Zeng, R., Wu, C., et al., 2019. A survey on deep-learning-based plant phenotype research in agriculture. Sci. Sinica Vitae 49 (698–716).

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Wu, X., Sahoo, D., Hoi, S.C., 2020. Recent advances in deep learning for object detection. Neurocomputing 396, 39–64.

Wu, S., Yang, J., Wang, X., Li, X., 2022. Iou-balanced loss functions for single-stage object detection. Pattern Recognit. Lett. 156, 96–103.

Xiang, S., Wang, S., Xu, M., Wang, W., Liu, W., 2023. YOLO POD: a fast and accurate multi-task model for dense soybean pod counting. Plant Methods 19 (1), 8.

Xie, X., Ge, Y., Walia, H., Yang, J., Yu, H., 2023. Leaf-counting in monocot plants using deep regression models. Sensors 23 (4), 1890.

Xiong, H., Cao, Z., Lu, H., Madec, S., Liu, L., Shen, C., 2019. TasselNetv2: in-field counting of wheat spikes with context-augmented local regression networks. Plant Methods 15 (1), 1–14.

Xu, C., Jiang, H., Yuen, P., Ahmad, K.Z., Chen, Y., 2020. MHW-PD: A robust rice panicles counting algorithm based on deep learning and multi-scale hybrid window. Comput. Electron. Agric. 173, 105375.

Xu, X., Wang, L., Shu, M., Liang, X., Ghafoor, A.Z., Liu, Y., Ma, Y., Zhu, J., 2022a. Detection and counting of maize leaves based on two-stage deep learning with UAV-based RGB image. Remote Sens. 14 (21), 5388.

Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.-H., Ma, J., 2022b. V2X-vit: Vehicle-to-everything cooperative perception with vision transformer. arXiv:2203.10638.

Xu, J., Yao, J., Zhai, H., Li, Q., Xu, Q., Xiang, Y., Liu, Y., Liu, T., Ma, H., Mao, Y., et al., 2023. Trichomeyolo: A neural network for automatic maize trichome counting. Plant Phenomics 5, 0024.

Yang, H., Chang, F., Huang, Y., Xu, M., Zhao, Y., Ma, L., Su, H., 2022a. Multi-object tracking using deep SORT and modified CenterNet in cotton seedling counting. Comput. Electron. Agric. 202, 107339.

Yang, Y., Deng, H., 2020. GC-YOLOv3: You only look once with global context block. Electronics 9 (8), 1235.

Yang, B., Gao, Z., Gao, Y., Zhu, Y., 2021. Rapid detection and counting of wheat ears in the field using YOLOv4 with attention module. Agronomy 11 (6), 1202.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F., 2022b. Image data augmentation for deep learning: A survey. arXiv preprint arXiv:2204.08610.

Yang Shuqin, W.P., 2022. Detecting wheat ears per unit area using an improved YOLOX. Trans. Chin. Soc. Agric. Eng. 38 (143–149).

Yu, Z., Ye, J., Li, C., Zhou, H., Li, X., 0000, TasselLFANet: A Novel Lightweight Multi-Branch Feature Aggregation Neural Network for High-throughput Image-based Maize Tassels Detection and Counting, Front. Plant Sci., Vol. 14, 1291.

Yu, W., Zhu, S., Yang, T., Chen, C., 2022. Consistency-based active learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3951–3960.

Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q., 2021. Multiple instance active learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5330–5339.

Yurtkulu, S.C., Şahin, Y.H., Unal, G., 2019. Semantic segmentation with extended DeepLabv3 architecture. In: 2019 27th Signal Processing and Communications Applications Conference. SIU, IEEE, pp. 1–4.

Zabawa, L., Kicherer, A., Klingbeil, L., T¨opfer, R., Kuhlmann, H., Roscher, R., 2020. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 164, 73–83.

Zhang, Z., Kayacan, E., Thompson, B., Chowdhary, G., 2020. High precision control and deep learning-based corn stand counting algorithms for agricultural robot. Auton. Robots 44 (7), 1289–1302.

Zhang, P., Li, D., 2023. Automatic counting of lettuce using an improved YOLOv5s with multiple lightweight strategies. Expert Syst. Appl. 226, 120220.

Zhang, W., Wang, J., Liu, Y., Chen, K., Li, H., Duan, Y., Wu, W., Shi, Y., Guo, W., 2022. Deep-learning-based in-field citrus fruit detection and tracking. Hortic. Res. 9.

Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z., 2018. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 637–653.

Zhao, Y., Wei, Y., Shan, H., et al., 2022. Wheat ear detection method based on deep learning. J. Agric. Sci. Technol. 24 (96–105), http://dx.doi.org/10.13304/j.nykjdb.2021.0612.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 7. pp. 13001–13008.

Zhu, Y., Zhou, J., Yang, Y., Liu, L., Liu, F., Kong, W., 2022. Rapid target detection of fruit trees using UAV imaging and improved light YOLOv4 algorithm. Remote Sens. 14 (17), 4324.

Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J., 2023. Object detection in 20 years: A survey. Proc. IEEE.