remote sensing

Article

Comparison of Deep Learning Methods for Detecting and Counting Sorghum Heads in UAV Imagery

He Li, Peng Wang and Chong Huang *

State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; lih@lreis.ac.cn (H.L.); 2101190086@cugb.edu.cn (P.W.)

* Correspondence: huangch@lreis.ac.cn; Tel.: +86-10-6488-8316

Abstract: With the rapid development of remote sensing with small, lightweight unmanned aerial vehicles (UAV), efficient and accurate crop spike counting, and yield estimation methods based on deep learning (DL) methods have begun to emerge, greatly reducing labor costs and enabling fast and accurate counting of sorghum spikes. However, there has not been a systematic, comprehensive evaluation of their applicability in cereal crop spike identification in UAV images, especially in sorghum head counting. To this end, this paper conducts a comparative study of the performance of three common DL algorithms, EfficientDet, Single Shot MultiBox Detector (SSD), and You Only Look Once (YOLOv4), for sorghum head detection based on lightweight UAV remote sensing data. The paper explores the effects of overlap ratio, confidence, and intersection over union (IoU) parameters, using the evaluation metrics of precision P, recall R, average precision AP, F1 score, computational efficiency, and the number of detected positive/negative samples (Objects detected consistent/inconsistent with real samples). The experiment results show the following. (1) The detection results of the three methods under dense coverage conditions were better than those under medium and sparse conditions. YOLOv4 had the most accurate detection under different coverage conditions; on the contrary, EfficientDet was the worst. While SSD obtained better detection results under dense conditions, the number of over-detections was larger. (2) It was concluded that although EfficientDet had a good positive sample detection rate, it detected the fewest samples, had the smallest R and F1, and its actual precision was poor, while its training time, although medium, had the lowest detection efficiency, and the detection time per image was 2.82-times that of SSD. SSD had medium values for P, AP, and the number of detected samples, but had the highest training and detection efficiency. YOLOv4 detected the largest number of positive samples, and its values for R, AP, and F1 were the highest among the three methods. Although the training time was the slowest, the detection efficiency was better than EfficientDet. (3) With an increase in the overlap ratios, both positive and negative samples tended to increase, and when the threshold value was 0.3, all three methods had better detection results. With an increase in the confidence value, the number of positive and negative samples significantly decreased, and when the threshold value was 0.3, it balanced the numbers for sample detection and detection accuracy. An increase in IoU was accompanied by a gradual decrease in the number of positive samples and a gradual increase in the number of negative samples. When the threshold value was 0.3, better detection was achieved. The research findings can provide a methodological basis for accurately detecting and counting sorghum heads using UAV.

Keywords: unmanned aerial vehicle; deep learning; EfficientDet; SSD; YOLOv4

Citation: Li, H.; Wang, P.; Huang, C. Comparison of Deep Learning Methods for Detecting and Counting Sorghum Heads in UAV Imagery. *Remote Sens.* **2022**, *14*, 3143. https://doi.org/10.3390/rs14133143

Academic Editor: Abdul M. Mouazen

Received: 9 May 2022
Accepted: 27 June 2022
Published: 30 June 2022

1. Introduction

Timely and accurate information on crop production is of great significance in formulating agricultural policies, macro-regulating food prices, and ensuring food security [1,2]. The number of crop ears per unit area is an important factor of crop yield [3]. Therefore, it has important practical significance for rapidly counting the number of ears to estimate crop yield [4–7].

However, counting the number of crop ears is a complex task. The traditional ways of manual counting combined with sampling methods are time-consuming, labor-intensive, and prone to human bias as they are influenced by plant density, human visual ability, sample representativeness, and sampling methods [3,4,6,8]. They cannot meet the need for monitoring a substantial number of plots consisting of a substantial number of varieties over continuous space and time. Automated counting methods are becoming increasingly important in modern precision farming, especially in smart agriculture applications [9].

With the rapid development of remote sensing platforms and sensor technology, imagery with high spatial and temporal resolutions can be obtained to detect and quantify crop ears in a low-cost and easy-to-use manner by unmanned aerial vehicles (UAVs) [10–13], which are flexible for image acquisition and unaffected by cloud and terrain conditions. UAVs can usually be equipped with a variety of sensors, such as RGB, multispectral, hyperspectral, thermal, and Light Detection and Ranging (LiDAR) [14]. Although the latter three sensors have better performance in crop detection, they are expensive and difficult to spread over a large area [8]. To circumvent this, most of the current crop plant counting studies are based on RGB and multispectral sensors [5,12,15–18], of which RGB sensors have a lower cost and higher spatial resolution. They also meet the requirements of crop plant counting using computer vision [4,8,19–22], which mainly uses some computer algorithms to simulate human visual functions to extract feature information from images, process, understand, and finally achieve counting of crop targets.

Besides the advances in remote sensors, the rapid development of computer software and hardware has also driven rapid changes in crop plant counting methods [23]. Traditional remote sensing methods for counting crop ears are mainly based on image processing, applying high/low pass filtering or morphological operators to achieve crop counting through image transformation and segmentation [11]. These methods are based on inference and are suitable for small datasets. However, these algorithms are not scalable when applied to images of different crop stages and conditions, such as changes in light conditions, shading, crop ear morphology, flowering status, soil background, and image quality [18]. Machine learning (ML), a product in the evolution of statistical learning to artificial intelligence, is suitable for non-linear crop ear counting. These types of methods use image processing techniques to extract features such as color and texture, and then use machine learning methods such as support vector machines (SVM) [24] and random forests (RF) [25] to build regression models for crop identification and counting [18]. Compared with traditional statistical learning methods, these methods have better scalability. However, ML algorithms require human-defined features, and the performance will be saturated with an increase in data, which cannot meet increasing data processing needs. Deep learning (DL), an emerging branch of machine learning, is driven by "big data" and aims to "minimize prediction errors" by building neural networks that mimic the human brain for analytical learning and has been proven to be the most advanced in processing massive, high-dimensional complex data [26,27]. Compared to ML, DL not only has the ability to automatically learn to extract features but can also achieve a richer spatial level of feature extraction through deep networks driven by big data [26,28]. In addition, DL has the capability for desaturation and high precision in remote sensing big data, which can meet the needs of accurate crop ear counting [23].

Deep learning-based crop detection and counting methods can usually be divided into three categories: segmentation-based methods, density map-based methods and object detection-based methods [23,27]. Segmentation-based methods usually use the Fully Convolutional Network (FCN) [29] or U-net [30] algorithms to achieve crop detection and counting by combining a deep Convolutional Neural Network (CNN) [31] with semantic segmentation of the images in high resolution. These algorithms perform well but poorly discriminate crop ears that overlap and occlude each other. To this end, related studies have further proposed density map-based methods, using MCNN (Multi-Column Convolutional

Neural Network, MCNN) [32] and CSRnet [33] to estimate the target densities of different parts of the image by training regression and integration to obtain the number of crop ears [23]. These methods improve the discrimination accuracy of the overlapping and mutually occluded crop ears and show good performance, but there are still shortcomings. The ability of the models to generalize has not been verified, and the accuracy needs to be further improved. In addition to the above two methods, many studies have carried out crop ear counting using object detection methods, which usually utilize Faster R-CNN [34], EfficientDet (a new network architecture proposed in 2020 on the basis of EfficientNet) [35], SSD (an end-to-end classical one-stage target detection algorithm that directly uses a single deep neural network to achieve feature extraction, which balances detection efficiency and accuracy) [36], or YOLO (an end-to-end trained real-time target detection network) [37] to detect and count crop ears by generating multiple location frames. These methods can visualize the target information of each crop ear, but the performance of the crop detection algorithm on images with very dense crop spike distribution is still unclear. The size of the crop ears also poses a challenge in selecting detection frames with appropriate aspect ratios.

Sorghum is the fifth top cereal crop in the world, planted in more than 100 countries and regions [38]. As an important mixed crop, sorghum has multiple resistances to high temperature, drought, flooding, salinity, and barrenness, and occupies a particularly important position in arid and semi-arid regions [6]. There are many varieties of sorghum, including grain sorghum for human food, brewing sorghum for alcoholic beverages, and forage sorghum for livestock hay and fodder [6]. By counting the number of sorghum heads, growers can estimate potential final yield and it is more practical for counting rapidly and producing timely estimates.

A series of related studies have been conducted. Zhao et al. [5] developed a pipeline to derive reflectance data from raw multispectral UAV images that preserve the original high spatial and spectral resolution, using these data for sorghum plant and head feature detection. Guo et al. [22] proposed a two-step machine-based image processing method to detect and count the number of sorghum heads from high resolution images captured by UAVs. Lin et al. [6] demonstrated that the integration of image segmentation and the U-Net CNN model is an accurate and robust method for counting sorghum panicles. Ghosal et al. [21] proposed an active learning-inspired weakly supervised deep learning framework for sorghum head detection and counting, which significantly reduced the human labeling effort without compromising final model performance to perform synthetic annotation. Compared to rice, wheat, and corn, the aforementioned studies on sorghum head counting have relatively few applications of object detection-based DL methods, and there is a lack of a systematic evaluation of the applicability of multiple object detection DL methods for sorghum.

To fill these research gaps, this study: (1) selected three object detection DL methods, namely EfficientDet, SSD, and YOLOv4, for sorghum head detection and counting based on UAV remote sensing imagery; (2) systematically evaluated the adaptability of the three DL methods and model parameters for sorghum head counting; and (3) obtained the most optimal method for sorghum head detection and counting.

2. Materials and Methods

2.1. Study Area

The Yellow River Delta (YRD) is located in Dongying City, northeastern Shandong Province, China, adjacent to Bohai Bay and Laizhou Bay [39]. The YRD has a temperate continental monsoon climate with four distinct seasons, an average annual temperature of 12.8 °C, and a frost-free period of 206 days. The average annual precipitation is 530 to 630 mm, of which 70% is concentrated in summer. Since the study area is in a coastal area, the soil type is mainly coastal alluvial soil, and the soil salinization is serious (the salt content is generally in the range of 0.40–4.00 g/kg, and the pH value is about 8.5) [40]. Sorghum has a high salinity tolerance and is one of the main crops grown in the area. The growing period of sorghum is mainly from mid-April to mid-October.

The study area of this experiment is located in the eastern part of the YRD at the Yellow River Delta Research Center experimental station, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (37°40′26′′N, 118°54′59′′E). A field was planted with sorghum, with controlled fertilizer and irrigation levels, allowing for accurate sorghum spike detection and counting studies. The total area of the experimental field was 2 ha, divided equally into six plots of 0.33 ha each (Figure 1).

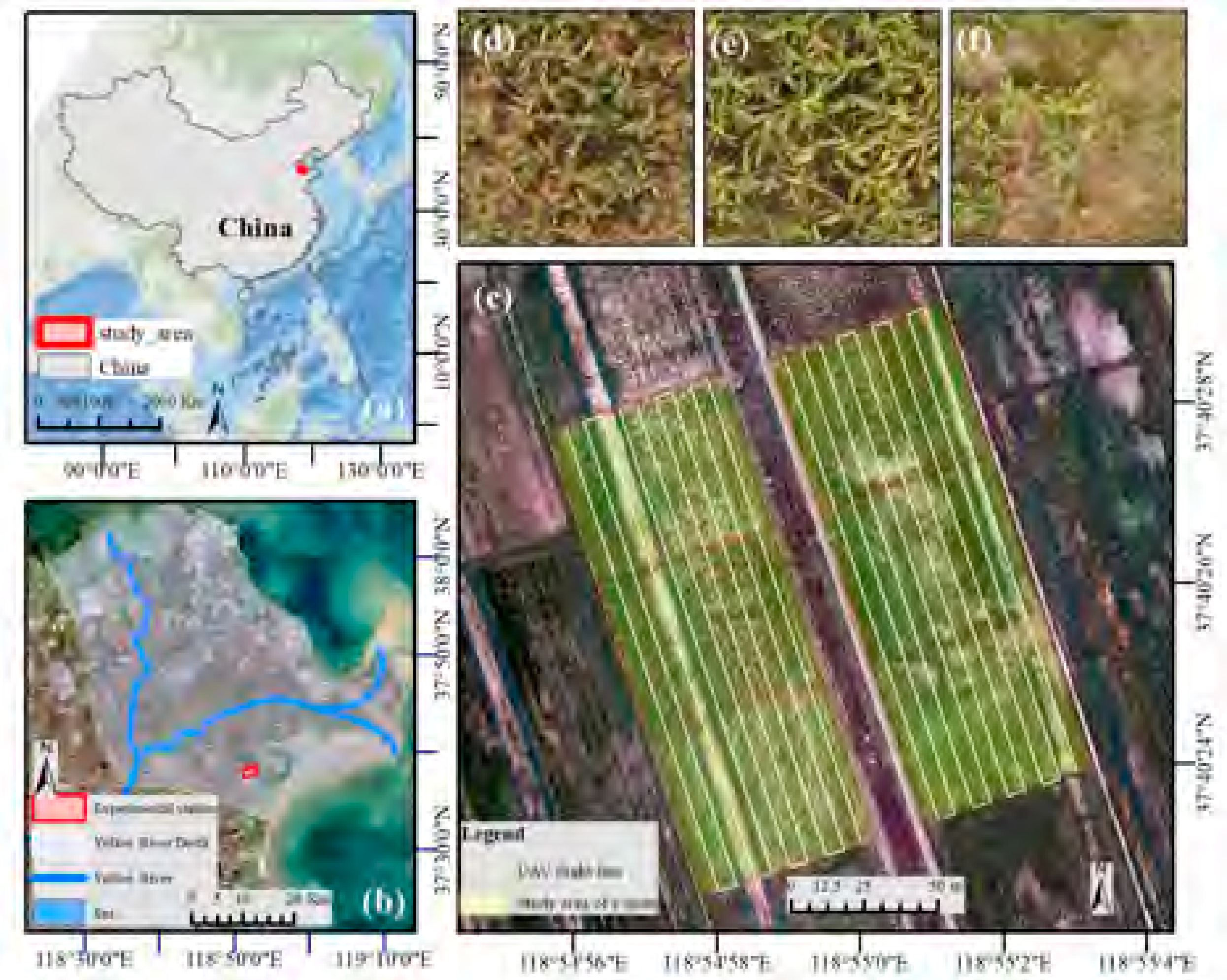

Figure 1. Location of the study area ((**a**) shows the location of the study area in China; (**b**) is the location of the experimental station in the Yellow River Delta; (**c**) is the test plots for this study, and (**d**–**f**) are the examples of UAV images of sorghum heads at different density conditions).

2.2. Data Collection and Preparation

2.2.1. UAV Image Collection

In this study, a DJI Phantom 4 Multispectral UAV (DJI, Shenzhen, China) was used for remote sensing image acquisition in the experimental field. This UAV integrated with a real-time kinematic (RTK) network, which enhances image positioning accuracy by calibrating the position in real time through the network while performing UAV image acquisition. The drone carries a multispectral sensor, containing one visible channel and five monochromatic channels for multispectral imaging (blue, green, red, red edge, and near infrared). Due to the high resolution of the visible sensor, the collected visible images at the millimeter or centimeter level were used for sorghum head detection and counting in this study. This drone has an integrated multispectral light intensity sensor at the top of the fuselage. When post-processing the images, the solar irradiance data will be used to compensate the illumination of images, eliminate the interference of ambient light on the data acquisition, and significantly improve the accuracy and consistency of the data collected at different times.

The flight parameters of the UAV were set using the processing software of DJI GS PRO (DJI, Shenzhen, China), which accompanies the DJI UAV, and the detailed aerial photography parameters are shown in Table 1. After several trials of aerial photography with the UAV in the field, the image acquisition mode was set to waypoint hover photography. The parameters of flight altitude, longitudinal and lateral overlap ratios were set to 20 m, 75 and 60%, respectively. The final ground resolution was 1.1 cm.

Table 1. Camera and aerial parameters of the DJI Phantom 4 Pro unmanned aerial vehicle.

Camera Parameters		**Aerial Photography Parameters**	
Parameters	**Value**	**Parameters**	**Value**
Effective pixels (RGB)	2.08 million	Flight altitude (m)	20
Field of view (FOV, °)	62.7	Longitudinal overlap ratio (%)	75
resolution	1600 × 1300	Lateral overlap ratio (%)	60
		Ground resolution (cm)	1.1

The UAV imagery was collected on 3 October 2020. At this time, the sorghum was fully in the milk-ripening stage and the images were collected to facilitate sorghum head counts. On the day of image acquisition, the weather conditions were favorable, with a southwesterly wind speed of 1.954 m/s.

Since the six plots were divided into two sides by a drainage ditch and road, separate routes were set for each side of the three plots to reduce the impact of extraneous imagery. In the end, a total of 928 waypoints and images were obtained and 464 UAV images were taken for each side of the three plots in this experimental field.

2.2.2. UAV Image Dataset Construction

The improvement in detection accuracy of DL algorithms is no longer obvious when the number of annotations reaches a certain level. Therefore, to prevent excessive duplication of sample data, the images of poor quality from the collected UAV data were eliminated and the remaining images after screening were used as training and validation samples in this study.

In the experiment, 51 images of clear sorghum heads were selected from 928 images and nine images of 416 × 416 pixels suitable for neural network training were cropped out of each training set image. On this basis, further image screening was conducted to eliminate images with fewer sorghum head targets, and finally the remaining 384 images were used for training and 96 were used for testing.

2.2.3. Image Labeling

The images needed to be labeled before training the sorghum head detection and counting algorithm on the UAV images. LABELIMG is a widely used open-source graphical annotation tool [41] suitable for object localization or detection. Using LABELIMG, a rectangle was drawn for each identified sorghum ear in the image. Since this study was conducted only for sorghum heads detection, the image was only divided into sorghum heads and background. Only the sorghum heads needed to be labeled, while the rest of the image was automatically labeled as background by the LABELIMG software and, finally, all the labeled information was saved directly as an XML file.

According to the Microsoft Common Objects in Context (MS COCO) dataset, targets smaller than 32 × 32 pixels were considered small targets [42]. It was found that the average pixels of sorghum heads in this study occupied only 7 × 7 pixels, which is a very small target and easy to miss when labeling. In order to prevent mislabeling and improve labeling accuracy, this study first trained the labeled images and used the trained model to detect the targets and automatically generate labels; then, the automatically generated labels were further corrected manually to finally obtain standard annotations for all samples (Figure 2).

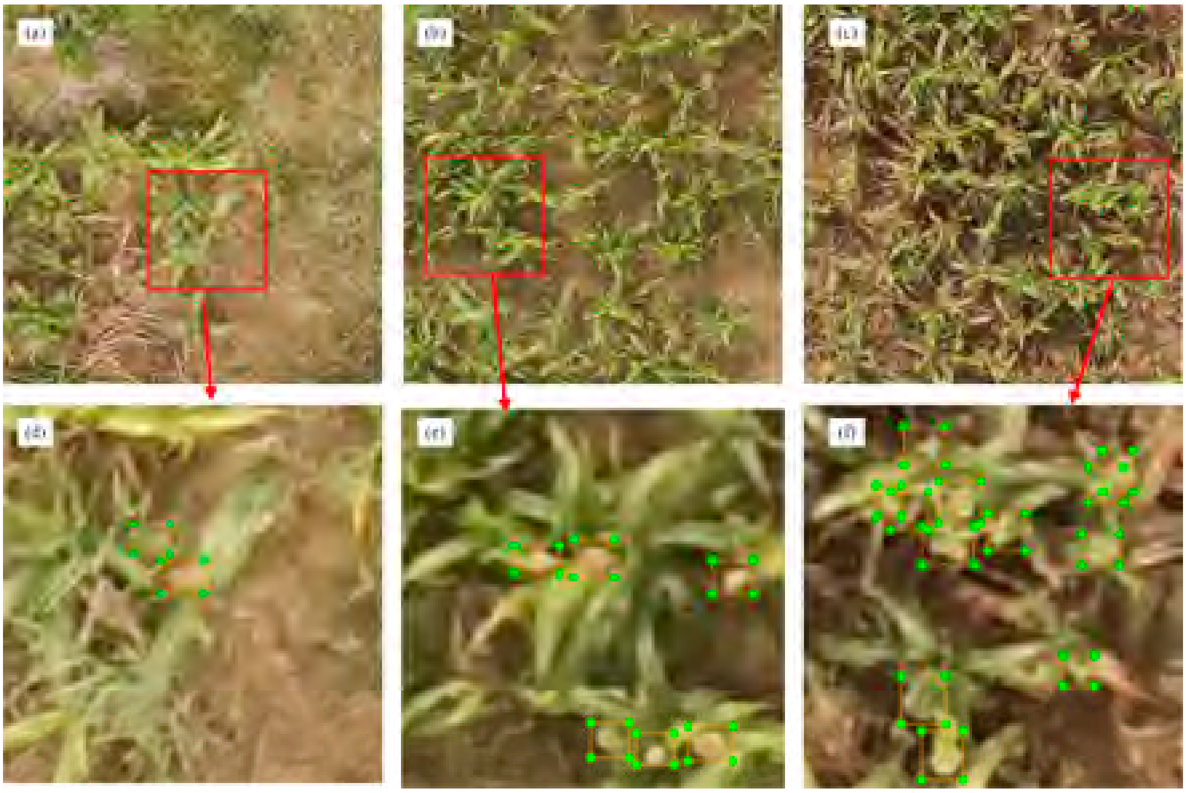

Figure 2. Example of labeling and corresponding local zoomed-in images for different densities of sorghum heads ((**a**–**c**) represents sparse, moderate, and dense conditions, and (**d**–**f**) represents the corresponding local magnified images, respectively).

3. Method

3.1. Deep Learning Algorithms

Current target detection algorithms based on DL can usually be divided into two-stage and one-stage detection algorithms. The former algorithm contains two target detection processes: candidate region extraction is the first target monitoring; candidate region classification and candidate region coordinate correction are the second target detection. These dual target detection processes improve the accuracy but also increase the model complexity and limit its computational efficiency. Furthermore, the algorithm extracts target information using the feature layer after multiple convolutions, which can easily lose small target information and is not suitable for sorghum head detection and counting.

The one-stage detection algorithm treats the target detection problem as a regression analysis problem of target location and category information and can directly output detection results through a neural network model. This algorithm contains only one target detection process, which has a simple structure, high computational efficiency, and can be easily trained end-to-end. It has great potential for application in the field of real-time target detection. Therefore, three one-stage target detection algorithms—EfficientDet, SSD, and YOLOv4—were selected to carry out the detection and counting of sorghum heads in this study.

3.1.1. EfficientDet

EfficientDet is mainly composed of three parts: backbone network, enhanced feature network, and box/class prediction network [43]. The backbone network employs EfficientNet, which continuously extracts the features from the input image with down-scaling and obtains P1–P7 [43]. Since they only shallowly down-scale and do not have sufficient semantic information, P1 and P2 are not used as inputs in the enhanced feature network. Then, a total of five effective feature layers of P3–P7 obtained by down-scaling are introduced into the enhanced feature network for further feature extraction and the proposed weighted bi-directional feature pyramid network (BiFPN) is used to repeat the operation to obtain five effective feature layers with high semantic information. Finally, these feature layers are fed into the box/class prediction network for box/class prediction or regression prediction and the prediction results are finally obtained [35].

As a mainstream one-stage detection model, EfficientDet still uses the common framework of "feature extraction, multi-scale feature fusion, and classification/regression prediction". The algorithm balances detection accuracy and efficiency by coordinating the network depth, the number of channels in each layer and the resolution of the input image, and the proposed BiFPN network structure enables the model to achieve efficient bi-directional cross-scale connectivity and weighted feature fusion, which yields better detection results on the COCO dataset [35]. However, the effectiveness of this algorithm in detecting the small targets of sorghum heads needs to be further explored. Therefore, the EfficientDet algorithm was selected for sorghum head detection and counting in this study. The overall architecture of EfficientDet is shown in Figure 3.

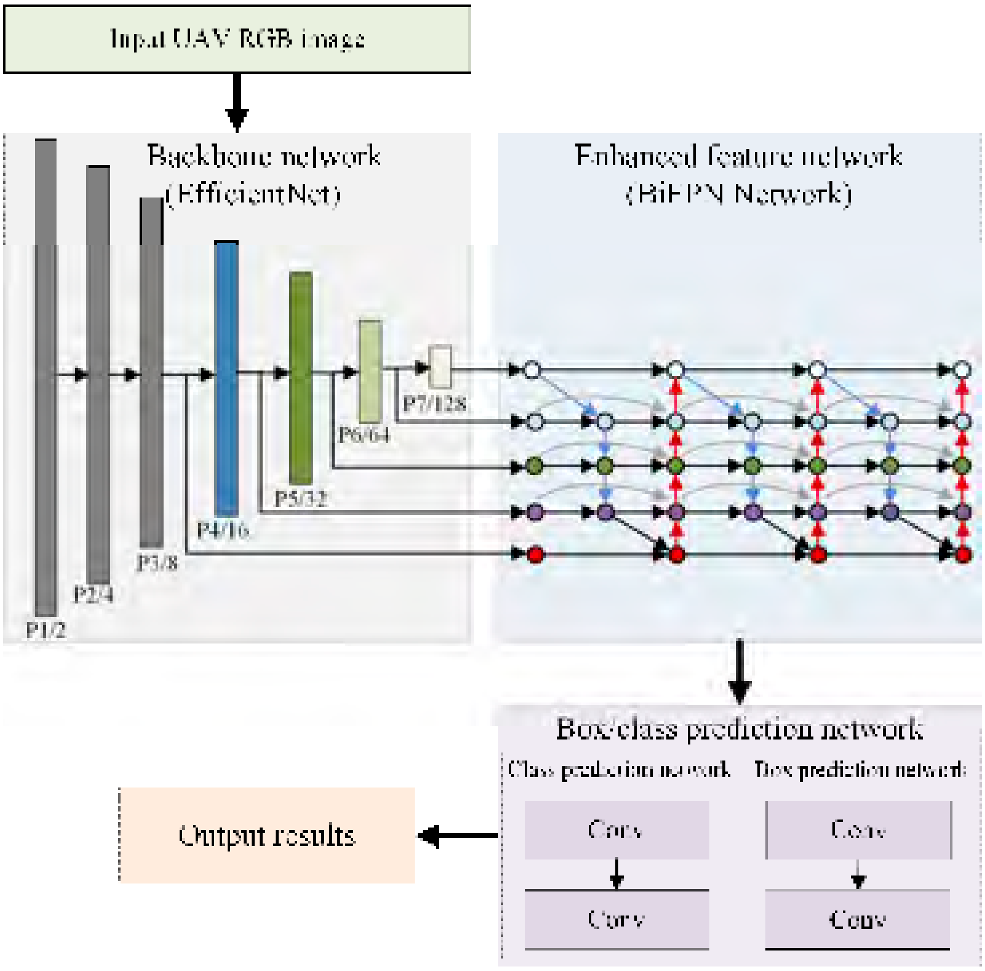

Figure 3. Architecture of EfficientDet proposed by Tan et al. [35].

3.1.2. SSD

SSD adopts the Visual Geometry Group (VGG) model [44] as its backbone network, which consists of 13 convolutional layers, three full connection layers and five pooling layers. The backbone network mainly acts on the fifth convolutional layer. Compared with AlexNet [34], VGG has stronger generalization and better performance.

The SSD algorithm has three main advantages. First, one of the core components of the algorithm is its adoption of multi-scale features for target detection. The feature pyramid structure is used to set different candidate frames on different feature layers to accommodate targets of different sizes for classification. The model has a small receptive field in the lower feature layers and mainly detects and identifies small targets, while a larger receptive field in the higher feature layers mainly detects and identifies large targets [45,46]. Secondly, the model sets up anchor boxes with different ratios of height and width. The prior box introduced by the model has a similar operation mechanism to anchor boxes in Faster R-CNN [34]. By constantly improving the box position, the target can be better matched. Third, the multiple data enhancement methods make the algorithm more robust to targets with different sizes and shapes of inputs [36].

The SSD algorithm uses a combination of shallow and deep feature information for detection, so it can have better detection accuracy for weak targets [36]. However, the sorghum heads in this study are smaller and more difficult to distinguish compared with other scenes of weak targets and the adaptability of the SSD algorithm needs to be further explored. Therefore, the SSD algorithm was selected for sorghum head detection and counting in this study. The overall architecture of SSD is shown in Figure 4.

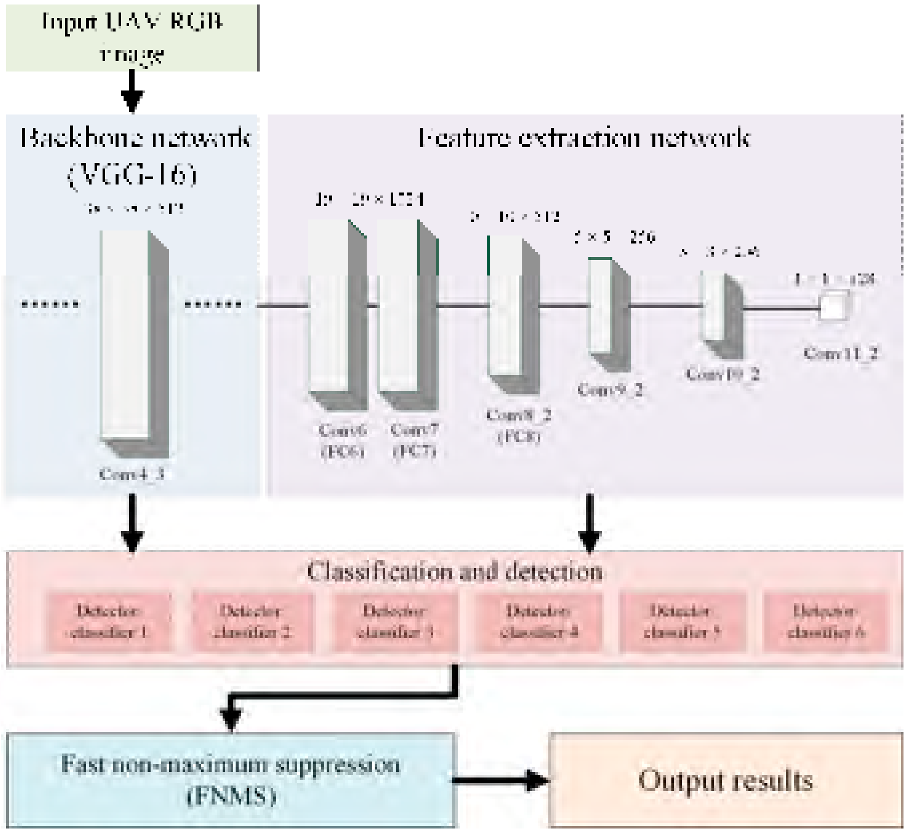

Figure 4. Architecture of SSD proposed by Liu et al. [36].

3.1.3. YOLOv4

The YOLO network is a new algorithm proposed by Joseph Redmon in 2015 [37]. Unlike region-based classification algorithms, its single convolutional network can predict the target class and bounding box of the entire image by running a single algorithm. YOLO divides the input image into S × S grids and each grid detects one object. In each grid, m bounding boxes are obtained. For each bounding box, the network provides an offset value of the bounding box and class probability, which are selected and further used to locate the object in the image with higher values than a specific threshold [37]. YOLO can obtain information from the entire image during training and testing, thus making good use of contextual information in detecting objects [37].

YOLOv4 is the better-performing target detection and recognition network in the YOLO series [47]. Compared with previous versions [48], YOLOv4 adopts a multi-scale detection algorithm and has a more complex network structure, which can effectively detect large and small targets in images. The YOLOv4 network structure mainly consists of three parts: feature extraction, feature fusion, and the detection head. The network of the feature extraction part is replaced by Cross-Stage-Partial-connections (CSPDarknet53), from Darknet53 used in YOLOv3 [49], which consists of five CSP residual resblock body modules, each of which contains a different number of residual block structures. The feature fusion part adopts spatial pyramid pooling (SPP) and a path aggregation network (PAN). SPP can integrate multi-scale perceptual field information and extract top and bottom features without any significant decrease in the network processing speed. PAN network structure enhances the feature hierarchy with precisely located signals using a bottom-up path enhancement method, shortens the information path between the bottom and topmost layers, and avoids the information loss problem, while the information obtained from the feature map after stitching contains both bottom and semantic features, realizing the two-way fusion of feature information from deep to shallow and from shallow to deep layers [50,51].

The YOLOv4 network has good recognition performance for significantly separated large and medium-sized targets [52,53], though few reports have been made on small targets for sorghum head detection. Therefore, YOLOv4 was selected for sorghum head detection in this study. The overall architecture is shown in Figure 5.

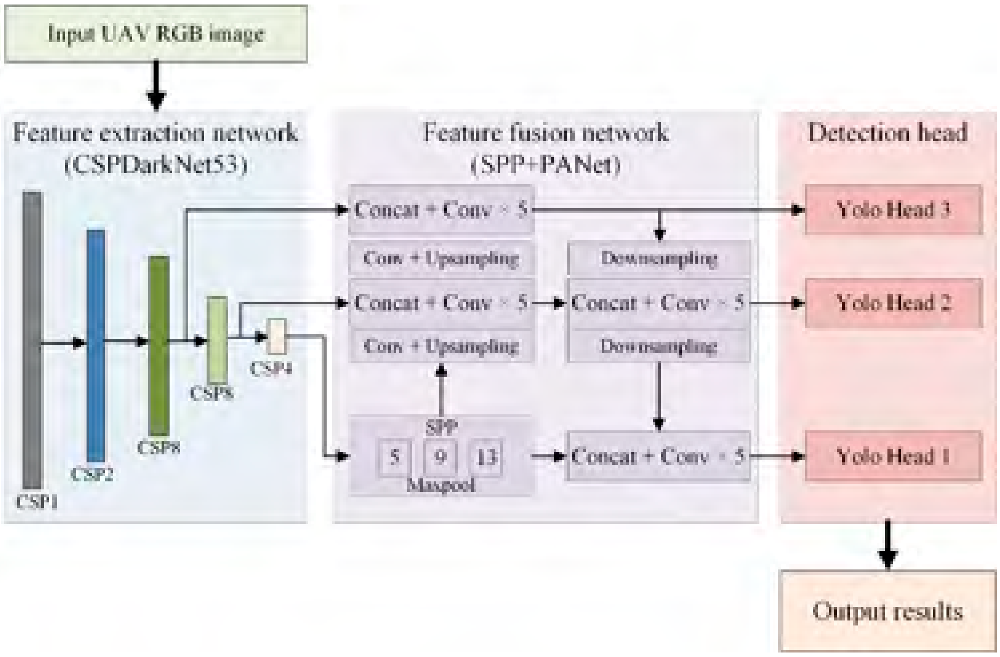

Figure 5. Architecture of YOLOv4 proposed by Bochkovskiy et al. [37].

3.2. *Programming and Model Training Environment*

3.2.1. Programming Environment

For better comparison of different deep learning methods for sorghum head detection and counting, Python was chosen as the programming language, and the TensorFlow 2.0 programming environment framework was applied for all DL models used in this study. The computer hardware configuration included 16 GB of RAM, a 2.60 GHz CPU (Intel® Core™ i7-9750H), and a 4 GB graphics card (NVIDIA GeForce GTX 1650).

3.2.2. Transfer Learning and Training

In the model training process, the shallow layers of the neural network can extract the edge, size, shape, texture, and other information of the image, which has strong transferability. Therefore, the method of transfer learning was adopted in this study [54,55].

All backbone feature extraction networks of the three algorithms involved in this study were frozen without training, and only the target feature detection networks were trained. The study was conducted using the weights trained in the Pattern Analysis, Statistical Modeling and Computational Learning (PASCAL) Visual Object Classes (VOC) dataset, combined with the labeled image dataset for the max training epoch value of 100, with the initial learning rate set to 1×10^{-3} and the batch size set to 2. Then, the final trained models were obtained.

3.3. Designing the Experiments

The study first explored the performance differences of the three deep learning algorithms, EfficientDet, SSD, and YOLOv4, in terms of sorghum heads detection and counting by training the samples. The effects of overlap ratio, confidence, and intersection over union (IoU) parameters on sorghum heads detection were then explored.

The overlap ratio is the ratio of two overlaps between the prediction frames [18,19]. It is an important indicator affecting how many prediction frames are available, which was reflected in the sorghum head detection performance. Seven overlap ratio thresholds (0.1–0.7) in steps of 0.1 were set to explore the impact of each DL method on sorghum head detection at different overlap ratio thresholds.

Confidence is used to determine whether the object in the prediction box is a positive sample or a negative sample, which are represented as the objects detected consistent or inconsistent with ground truth samples [20]. If the object is larger than the confidence threshold, it is a positive sample, while it is a negative sample, namely the background, when the object is smaller than the confidence threshold. Confidence is another important metric affecting how many prediction frames are available and also an important measure of the accuracy of detecting sorghum heads. Seven overlap ratio thresholds (0.1–0.7) in steps of 0.1 were set with other parameters held constant to explore the impact of each DL method on sorghum head detection at different confidence thresholds.

The intersection of union (IoU) is the overlap ratio between the predicted bounding box and the corresponding labeled ground truth (GT) bounding box [20,21]. IoU also affects prediction frames and measures detection accuracy. Seven IoU thresholds (0.1–0.7) in steps of 0.1 were set with other constant parameters to see the impact of each DL method on sorghum head detection at different overlap ratio thresholds.

3.4. Evaluation Metrics

Four metrics, namely the precision (P), recall (R), average precision (AP), and F1 score, were utilized in this study to evaluate the above three DL models for the purpose of sorghum head detection and counting. P is the proportion of the number of the samples whose predicted value is the true value in the total number of samples, while R is the proportion of the number of samples whose predicted value is the true value out of the total number of positive values. AP is a frequently used metric for the evaluation of object detection and can be considered the area under the P-R curve. The F1 score is the harmonic accuracy of P and R and is their weighted average.

To calculate the above four metrics, a predicted bounding box was considered a true positive (*TP*) if it overlapped more than the IoU area threshold with the corresponding labeled ground truth (GT) bounding box. Otherwise, the predicted bounding box was considered a false positive (*FP*). When the labelled GT bounding box had an IoU with a predicted bounding box lower than the threshold value, it was considered a false negative (*FN*). The formulas for calculating the four evaluation metrics are as follows:

$$\mathrm{P} = \frac{TP}{TP + FP} \quad (1)$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{2}$$

$$\mathrm{AP} = \sum_{i=1}^{k} \mathrm{P}(i)\Delta\mathrm{R}(i) \tag{3}$$

$$\mathrm{F1} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \tag{4}$$

where k represents the number of all images in the test set, which is equal to 96 in this study; P(i) represents the value of P when i images can be detected; and ΔR(i) represents the change in R when the number of detected images changes from $i-1$ to i.

4. Results

4.1. Comparison of the Detection Results with Different Model Algorithms

Due to the salinity of the coastal soil, the distribution of sorghum heads in the experimental area showed dense, moderate, and sparse conditions. Therefore, the detection results of the three deep learning methods under different sorghum head cover conditions were compared. The EfficientDet method detected only 42 of the 75 actual sorghum heads with a missing rate of 44% under dense conditions, detected 13 of the 32 actual ears with a missing rate close to 60% under moderate conditions, and detected 5 of the 12 actual ears with a missing rate of 58.33%, and over-detected one ear under sparse conditions (Figure 6a,d,g). Overall, the EfficientDet detection results were poor, and the algorithm suffered from a large number of missed detections at different densities, where the moderate and sparse conditions had worse results compared to dense conditions.

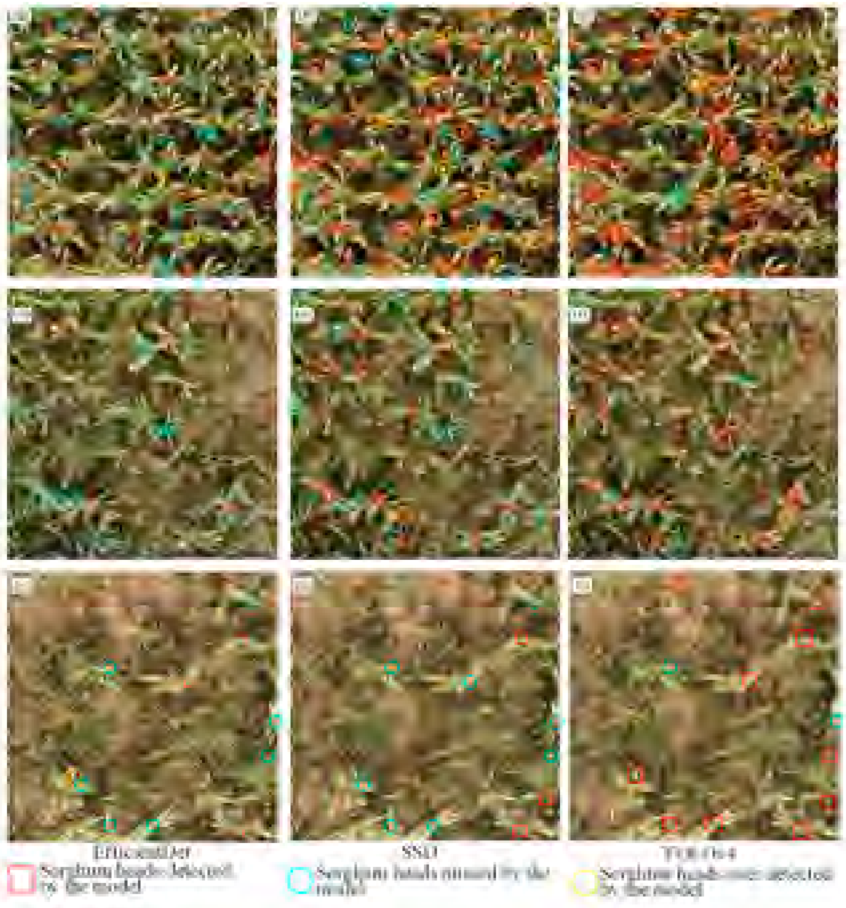
EfficientDet
SSD
YOLOV4
Sorghum heads detected by the model
Sorghum heads missed by the model
Sorghum heads over-detected by the model

Figure 6. Detection results for sorghum heads with different densities by different algorithms ((**a**,**d**,**g**), (**b**,**e**,**h**) and (**c**,**f**,**i**) represents the three DL algorithms of EfficientDet, SSD, and YOLOV4 at the dense, moderate, and sparse conditions, respectively).

For SSD, 53 sorghum heads were detected under dense conditions with a missing rate of less than 30%, 20 sorghum spikes were detected with a missing rate of only 28.57% under moderate conditions, and only three sorghum spikes were detected with a missing rate of more than 70% under sparse conditions. With the decrease in sorghum head density, the detection accuracy also decreased. In addition, SSD also had a more serious over-detection problem under the dense conditions of sorghum spikes, with 14 ears over-detected. In contrast to EfficientDet, SSD showed a greater improvement in detection results under both dense and moderate conditions, but also had more serious over-detection problems. SSD was also less effective than the EfficientDet in detection, with an actual accurate detection rate of only 25%.

YOLOv4 had good detection accuracy under different coverage conditions. There were six, seven, and two missed detections for dense, moderate, and sparse conditions, respectively, at 8.00, 21.88, and 16.67%. However, the algorithm also has an over-detection problem, but the number of over-detections was less, with four and one over-detected sorghum heads under dense and medium conditions, respectively. The overall detection was good.

In summary, the detection of sorghum heads was best under dense cover conditions, and worst under moderate conditions for these three methods. Among them, YOLOv4 had the highest detection accuracy under different coverage conditions, while EfficientDet had the worst, with a missed detection rate of more than 40%. Although SSD was relatively accurate under dense conditions, it also had certain over-detection problems. In conclusion, YOLOv4 is relatively optimal in sorghum head detection.

4.2. Performance Evaluation of Different Models for Sorghum Head Detection

4.2.1. Evaluation of Positive and Negative Samples for Different Models

Based on the comparison of the detection results, the study further evaluated the detection of positive and negative samples for the 96 test images (Figure 7). There were a total of 3961 sorghum heads, and EfficientDet, SSD, and YOLOv4 could detect 1879, 2961 and 3608 heads, accounting for 47.44, 74.75, and 91.09% of the total number of sorghum heads in the test set, respectively. For the detected sorghum heads, the number of positive samples detected by the three methods accounted for 90.69, 71.43, and 94.12% of the total number of detected samples, respectively.

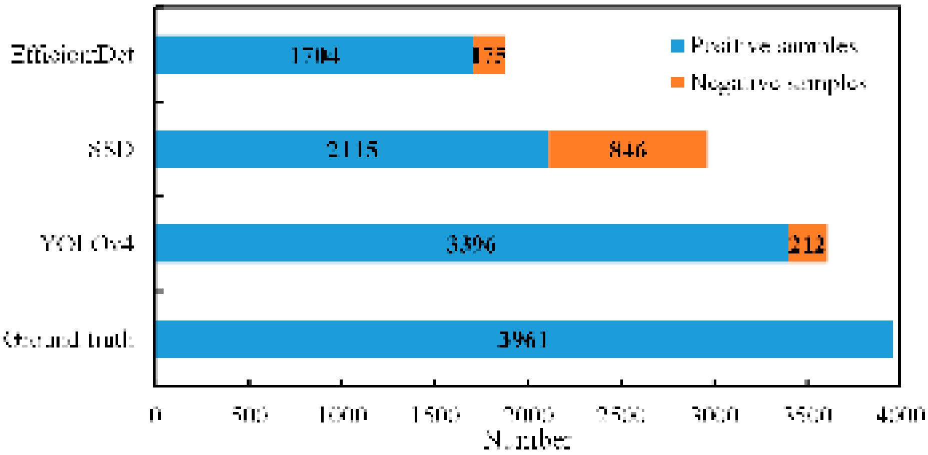

Figure 7. Comparison of the number of sorghum heads detected by different deep learning methods.

To sum up, YOLOv4 had the best detection rate of sorghum heads both in terms of the number of detections and positive samples. EfficientDet detected the fewest sorghum heads, but the accuracy of positive samples in the detected sorghum heads was also high. SSD was in the middle in the number of detections, and the worst in terms of accuracy of positive samples. Therefore, YOLOv4 could obtain the best detection results in terms of positive and negative sample evaluation.

4.2.2. Accuracy Evaluation of Different Methods

The four evaluation metrics of P, R, AP, and F1 were used to evaluate the detection performance of the DL methods (Table 2). It was found that EfficientDet had the highest precision P and the lowest recall R, and its AP and F1 score were the lowest, at 40.39% and 0.06, respectively. SSD had the lowest P and moderate R, at 84.38 and 38.05%, respectively. Its AP and F1 score were medium, at 47.30% and 0.52. The P of YOLOv4 was 97.62%, the recall rate was 64.20%, and the AP and F1 scores of YOLOv4 were 84.51% and 0.77, respectively. Among the three methods, YOLOv4 had relatively good performance in the four metrics.

Table 2. Accuracy evaluation of each model based on the whole testing dataset.

Method	P (%)	R (%)	AP (%)	F1 Scores
EfficientDet	99.20	3.13	40.39	0.06
SSD	84.38	38.05	47.30	0.52
YOLOv4	97.62	64.20	84.51	0.77

The study further compared the actual number of sorghum heads per test image in the test set with the number of detections by different methods (Figure 8). It was found that the three methods of EfficientDet, SSD and YOLOv4 all had good correlation between the detected sorghum head count and the true count, with coefficients of determination (R^2) of 0.90, 0.92, and 0.99, respectively. However, EfficientDet counting had the largest bias, with a representative slope of 0.54, while YOLOv4 had the best fit with the actual number of detections, with a representative slope of 0.99. Therefore, YOLOv4 had the best sorghum head detection and counting results, which were closest to the actual number of sorghum heads and better reflected the actual conditions.

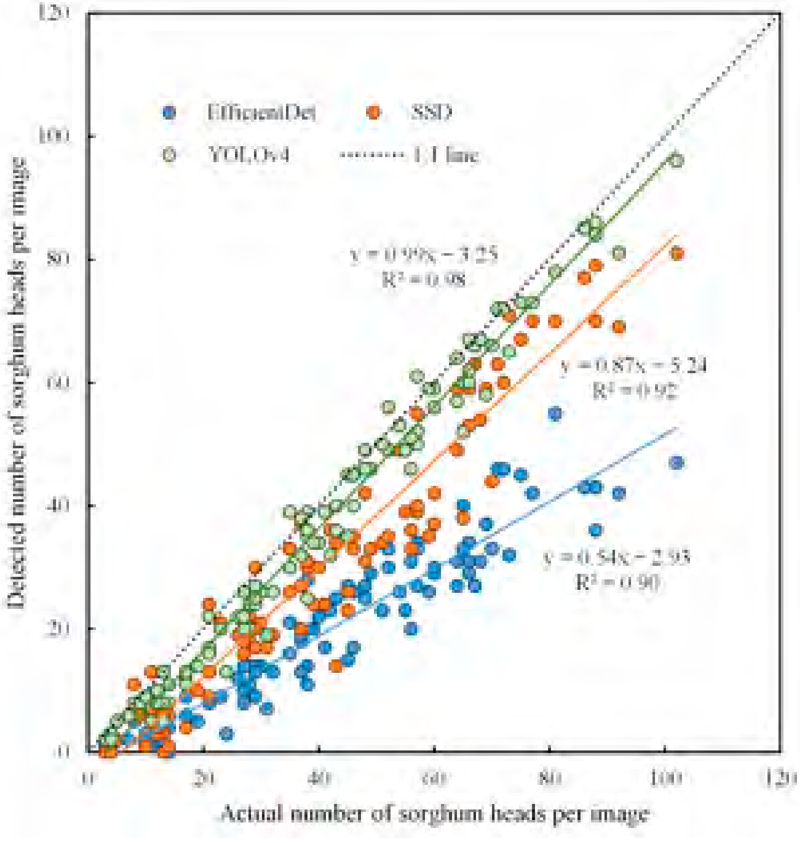

Figure 8. Comparison between the actual number and detected number of sorghum heads per test image using different deep learning methods.

4.2.3. Comparison of Computational Efficiency of Different Methods

In addition to accuracy, detection efficiency is also an important metric for DL target detection. The study further compared the computational efficiency of the three methods in sorghum head detection and counting. In Figure 9, the training times of EfficientDet, SSD, and YOLOv4 were relatively long due to the computer configuration, but SSD consumed the least time in both training time and detection time under the uniform computer configuration, with an average time of 23 s per training epoch and the most efficient average detection time of 0.0160 s per image. YOLOv4 had the slowest training time, with an average time of 32 s per training epoch, while EfficientDet was the most time-consuming in terms of image detection time, with an average detection time of 0.0451 s per image, which is 2.82-times the SSD detection time.

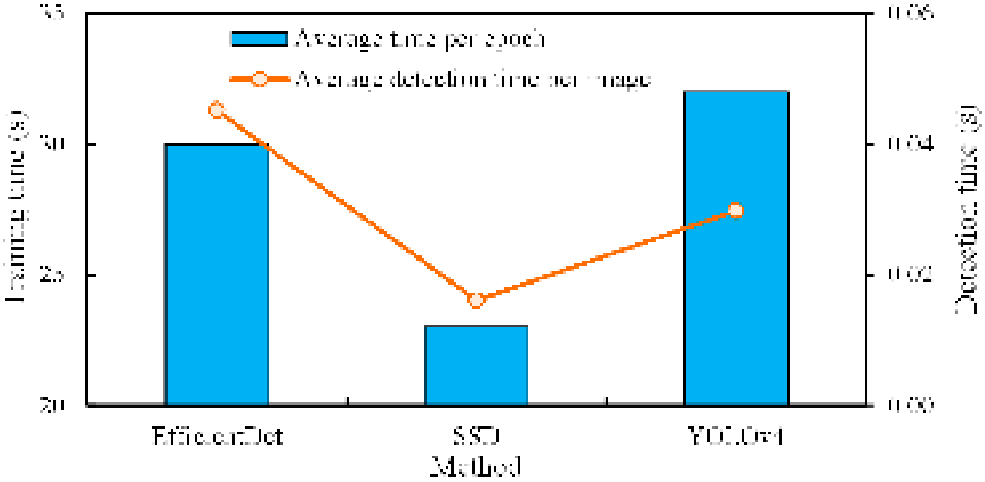

Figure 9. Comparison of the computational time of three different deep learning methods.

Therefore, when performing sorghum head detection, the DL method could be selected according to the actual needs. If we focused on detection efficiency, SSD was the best method with the shortest training and detection time. Considering the detection accuracy and efficiency, YOLOv4 had an acceptable image detection speed.

4.3. Comparison of Overlap Ratio Thresholds

It was concluded that as the overlap ratio increased, the positive and negative samples detected by each method showed an increasing trend, with the number of positive samples changing relatively slowly and the number of negative samples changing more dramatically (Figure 10 and Table 3). Especially after 0.3, the number of negative samples gradually increased significantly.

In terms of performance, P, R, AP, and F1 changed relatively slowly as the overlap ratio increased. Each method obtained relatively optimal sorghum head detection results with an overlap ratio of 0.3–0.5. The optimal overlap ratio for EfficientDet was 0.3, with AP and F1 score values of 40.39% and 0.06, respectively; for SSD, it was 0.5 (AP and F1 were 47.52% and 0.53); and for YOLOv4, the rate was 0.5 (AP and F1 were 84.56% and 0.77).

Considering the number of samples and performance evaluation metrics, an overlap ratio of 0.3 was chosen to obtain relatively optimal sorghum head detection results.

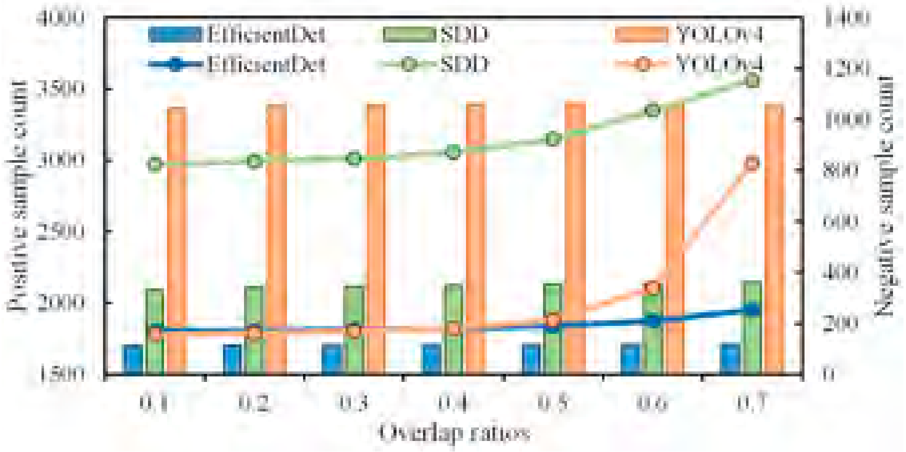

Figure 10. Statistical results of positive and negative sample numbers for each method with different overlap ratios (the histogram in the figure corresponds to the number of positive samples, the line graph corresponds to the number of negative samples; blue, green, and orange correspond to the three methods of EfficientDet, SSD, and YOLOv4, respectively).

Table 3. Performance evaluation results of each method with different overlap ratios.

Method	Evaluation Metrics	Overlap Ratios						
		0.1	0.2	0.3	0.4	0.5	0.6	0.7
EfficientDet	P (%)	99.20	99.20	99.20	99.20	99.20	99.20	99.20
	R (%)	3.13	3.13	3.13	3.13	3.13	3.13	3.13
	AP (%)	40.32	40.35	40.39	40.35	40.36	40.27	40.04
	F1	0.06	0.06	0.06	0.06	0.06	0.06	0.06
SSD	P (%)	84.36	84.39	84.38	84.39	84.31	83.99	83.49
	R (%)	37.87	37.94	38.05	38.07	38.12	38.15	38.17
	AP (%)	46.82	47.15	47.30	47.44	47.52	47.45	47.27
	F1	0.52	0.52	0.52	0.52	0.53	0.52	0.52
YOLOv4	P (%)	97.79	97.73	97.62	97.47	97.44	96.23	89.83
	R (%)	64.00	64.10	64.20	64.30	64.33	64.40	64.43
	AP (%)	83.95	84.36	84.51	84.55	84.56	83.97	80.56
	F1	0.77	0.77	0.77	0.77	0.77	0.77	0.75

4.4. Comparison of Confidence Values

It can be seen that as the confidence threshold increased, the number of positive and negative samples for all three methods tended to decrease (Figure 11 and Table 4). Especially after 0.3, the number of negative samples decreased gradually. In terms of performance, all three methods had small changes in P, R, and F1 score as the confidence increased, but AP decreased significantly. Although the optimal threshold was 0.1, the counts of negative samples were high, at 7276, 25,675, and 2200, respectively, resulting in poor performance in sorghum head detection.

Considering the number of samples and performance evaluation metrics, a confidence threshold of 0.3 was chosen to obtain relatively optimal detection results.

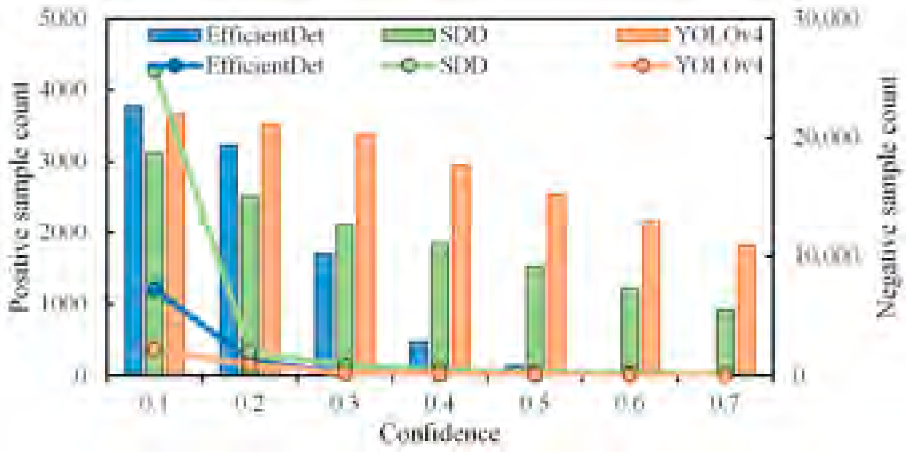

Figure 11. Statistical results of positive and negative sample numbers for each method with different confidence thresholds (the histogram in the figure corresponds to the number of positive samples, the line graph corresponds to the number of negative samples; blue, green, and orange correspond to the three methods, EfficientDet, SSD, and YOLOv4, respectively).

Table 4. Performance evaluation results of each method with different confidence thresholds.

Method	**Evaluation Metrics**	**Confidence**						
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
EfficientDet	P (%)	99.20	99.20	99.20	99.20	99.20	10.00	100.00
	R (%)	3.13	3.13	3.13	3.13	3.13	0.91	0.25
	AP (%)	80.74	72.61	40.39	11.23	3.12	0.91	0.25
	F1	0.06	0.06	0.06	0.06	0.06	0.02	0.01
SSD	P (%)	84.38	84.38	84.38	84.38	84.38	84.46	92.59
	R (%)	38.05	38.05	38.05	38.05	38.05	20.40	23.33
	AP (%)	59.41	53.80	47.30	42.10	35.18	28.58	22.19
	F1	0.52	0.52	0.52	0.52	0.52	0.45	0.37
YOLOv4	P (%)	97.58	97.62	97.62	97.62	97.58	97.64	99.40
	R (%)	64.23	64.20	64.20	64.20	64.23	54.83	46.07
	AP (%)	89.93	87.42	84.51	73.72	63.91	54.69	46.02
	F1	0.77	0.77	0.77	0.77	0.77	0.70	0.63

4.5. Comparison of IoU Thresholds

It was found that with an increase in IoU, the positive samples of the three algorithms showed a decreasing trend, but the negative samples showed an opposite increasing trend, and the inflection points of the positive and negative samples of the three methods were about 0.3 (Figure 12 and Table 5). In terms of performance, the P and R values of the three methods remained basically unchanged as IoU increased and AP and F1 scores decreased gradually and significantly, so after the IoU threshold was equal to 0.3.

Considering the number of samples and performance evaluation metrics, an IoU threshold of 0.3 was chosen to obtain relatively optimal sorghum head detection results.

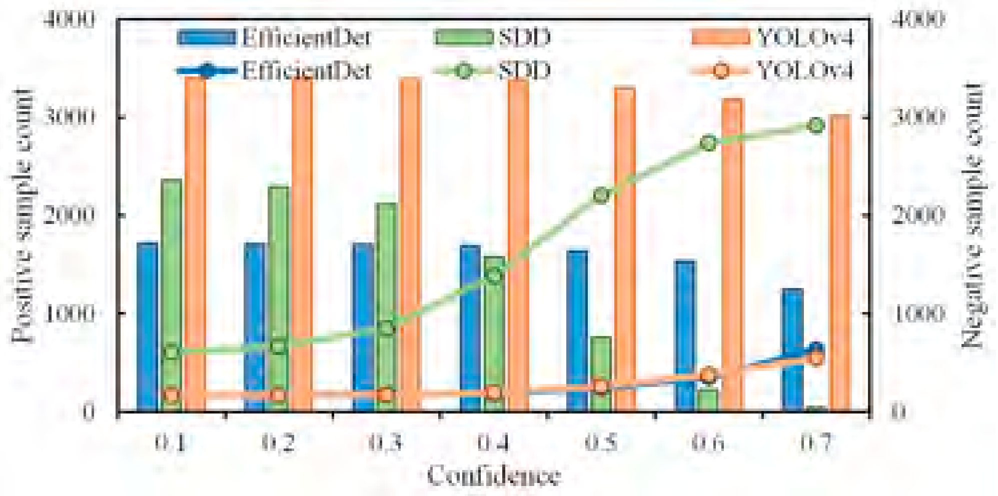

Figure 12. Statistical results of positive and negative sample numbers for each method with different IoU thresholds (the histogram in the figure corresponds to the number of positive samples, the line graph corresponds to the number of negative samples; blue, green, and orange correspond to the three methods, EfficientDet, SSD, and YOLOv4, respectively).

Table 5. Performance evaluation results of each method with different IoU thresholds.

Method	**Evaluation Metrics**	**IoU**						
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
EfficientDet	P (%)	99.20	99.20	99.20	99.20	95.20	92.00	80.00
	R (%)	3.13	3.13	3.13	3.13	3.00	2.90	2.52
	AP (%)	40.87	40.68	40.39	39.96	37.38	33.21	22.93
	F1	0.06	0.06	0.06	0.06	0.06	0.06	0.05
SSD	P (%)	91.55	90.31	84.38	63.77	29.79	8.85	1.40
	R (%)	41.28	40.72	38.05	28.76	13.43	3.99	0.63
	AP (%)	55.94	53.96	47.30	27.51	6.04	0.54	0.04
	F1	0.57	0.56	0.52	0.40	0.19	0.05	0.01
YOLOv4	P (%)	97.77	97.77	97.62	97.16	95.82	92.67	88.10
	R (%)	64.33	64.30	64.23	63.90	63.01	60.99	57.94
	AP (%)	84.79	84.74	84.51	83.62	81.16	76.48	69.29
	F1	0.78	0.78	0.77	0.77	0.76	0.74	0.70

5. Discussion

5.1. Effect of Different DL Methods on Sorghum Head Detection

It was concluded that EfficientDet was the least effective in sorghum head detection for small targets, with the lowest average accuracy AP and F1 scores. Most of the detected targets were still concentrated in the low confidence region, proving that EfficientDet required high target clarity and was suitable for large target detection, which was consistent with the conclusion of related studies focused on detecting large targets such as crop growing circles [56], ships [57], pedestrians [58], and solid waste garbage [59]. In addition, the BiFPN network with integrated bidirectional cross-scale connectivity and fast normalized fusion was largely inferior to the SSD algorithm of the VGG-16 network and the YOLOv4 algorithm of the SPP+ PANet network in terms of detection efficiency [60]. However, the EfficientDet detected very few wrongly detected high-confidence targets, i.e., among the 1879 sorghum heads detected in Figure 7, there were only 175 wrongly detected negative samples, and the number of negative samples accounted for only 9.31% of the number detected, indicating that this method could have high detection accuracy if sufficiently detailed information could be provided [57,60]. Meanwhile EfficientDet had a medium number of parameters and the time required for model training was in the middle among the three methods [56].

Compared with EfficientDet, SSD had better results on the detection of sorghum heads. The number of sorghum heads detected and the AP and F1 scores were all higher than EfficientDet, but lower than YOLOv4, mainly because SSD is based on the VGG-16 network with deeper network layers. In the process of convolution, the effective feature information of the sorghum heads extracted by the shallow feature layer was less, causing it to be insufficient to accurately detect the sorghum head targets, generating a large number of negative samples with low confidence. The deeper detection layers only detected fewer but higher confidence targets [46,61], causing SSD to detect a higher number of sorghum heads, but the detected results had more wrongly detected sorghum heads. In terms of algorithm efficiency, SSD had the least number of parameters and the shortest training time among the three methods, and also its detection efficiency was the highest because its detection layers detected the target alone and finally only needed to be fused by FNMS to obtain the detection results, which was also consistent with the findings of Liu et al. [36], Yi et al. [46] and Aziz et al. [62].

YOLOv4 had the highest detection count and highest R, AP, and F1 scores for sorghum head detection. The main reason was that YOLOv4 used the PANet network structure to fully fuse the features of different feature layers. Compared to EfficientDet, which required higher clarity of sorghum heads, and the SSD network, which predicted targets for each feature layer individually, the YOLOv4 algorithm had better sorghum head detection results. The YOLO series of algorithms has been used for detecting corn plant seedlings [19], rice ears [10], cotton seedlings [63], cherry fruit [64], apples and apple flowers [49,51], and greenhouses [65], all of which have more extensive applications and better detection results. However, there are more convolutional layers stacked on each other in the CSPDarknet53 backbone network of YOLOv4, so that YOLOv4 had a larger number of parameters and floating-point computation, and its number of parameters was the largest among the three methods. This resulted in the longest training time, but the detection efficiency was better than EfficientDet with BiFPN network structure and was only less efficient than SSD detection.

Therefore, considering its recognition accuracy and efficiency, YOLOv4 should be preferred as the optimal method for sorghum head detection and counting in practical applications.

5.2. Effect of Model Parameters on Sorghum Head Detection

In addition to the differences in sorghum head detection caused by different network structures and depths of the deep learning methods [4,66], the differences in overlap ratio, confidence, and IoU in sorghum head detection are also discussed separately.

Overlap ratios have been less explored in previous research, and the default threshold of 0.5 has been commonly used [18,19,52,67]. Our results show that the overlap ratio was neither larger nor smaller: with a larger overlap ratio and a weakened ability to detect sorghum heads, while the detection results also contained some anchor frames of mutual inclusion relationships, causing certain over-detection problems (Figure 10). This conclusion was consistent with the findings of Ma et al. [4] on wheat ear counting. Through experiments, we found that the optimal interval of the overlap ratios on different methods was 0.3–0.5. When combining the number of positive and negative samples, it finally obtained the overlap ratios at 0.3, which balanced all factors and yielded better sorghum head detection.

For each UAV image, each DL method returned a set of prediction anchor frames with a confidence value (threshold range 0–1). By setting the confidence threshold, the NMS directly filters out prediction anchor frames smaller than this confidence threshold [20]. On the other hand, the NMS also removed images in the high overlap region with a high overlap ratio with the maximum confidence prediction frame [68] and the impact of this part was mainly controlled by setting the threshold value of the overlap ratios. Unlike previous studies where the default confidence threshold of 0.5 was commonly used [20,53,69], this study concluded that a confidence of 0.3 was the best for sorghum head recognition. The main reason was that the sorghum heads in this study were weak, small targets on the image and the confidence returned by the prediction frame was generally not high due to factors such as the UAV image, the sorghum heads themselves, and the environmental background (see Section 5.3 for details). A high confidence threshold setting would eliminate many correct prediction frames and affect the final detection results. On the contrary, a small confidence threshold would cause a geometric increase in negative samples and affect the final detection.

Most of the previous studies set IoU at the default 0.5 or 0.75 for the threshold when evaluating the performance of common algorithms [8,15,20,21,53,68,70]. Our results found that the number of positive and negative samples of the three methods rapidly decreased when IoU exceeded 0.3, while AP and F1 values rapidly decreased, and detection performance rapidly decayed in sorghum head detection. Therefore, the optimal IoU threshold was 0.3, indicating that in practical target detection studies, appropriate IoU thresholds need to be set according to the size of different detection targets. The sorghum heads in this study are small targets or have a high overlap ratio, so a smaller IoU threshold needed to be selected, and this conclusion was consistent with the results of Velumani et al. for corn planting density [69]. On the contrary, for medium and large targets or targets with a low overlap ratio, an IoU threshold of 0.5 or more was selected to obtain better detection results; for example, the average IoU detected by Malambo et al. exceeded 0.8 [71] and the IoU value of Tian et al. for the detection of apples exceeded 0.85 [49].

5.3. Effects of Other Factors

In addition to DL methods and associated model parameters, there were many factors that affected sorghum head detection and counting performance from three main aspects: the UAV image, the sorghum head itself, and the environmental context [4,5].

The main influencing factor in terms of UAV images was the spatial resolution of the images [4], which was an important indicator affecting the accuracy of sorghum head detection, and was mainly affected by the UAV flight altitude [8], which in this study was 20 m and the spatial resolution of the image was 1.1 cm. Compared to existing studies on crop spike detection and counting that commonly used millimeter-level (generally less than 5 mm) spatial resolution images [4,6,10,18,20,22,71], the sorghum head targets in this study were smaller and their boundaries were more blurred in the images used in this study at the centimeter level. The lower spatial resolution directly affected the learning performance of the DL methods [72,73], which undoubtedly hindered sorghum head detection and counting from the UAV images. Fortunately, by studying the impact of sorghum head detection with three different DL methods and related parameters in this study, we found that we could also obtain better sorghum head detection at centimeter-level spatial resolution using the YOLOv4 method combined with appropriate overlap ratios, confidence, and IoU. However, the centimeter-level spatial resolution images had geometrically increased image elements compared to previous satellite data at the meter, 100-m, or kilometer level. The collected UAV images generally needed to be cropped into many small images for DL training, detection, and counting due to the detected window size and computational performance limitations of DL methods [4,5,7,68]. In this study, the original UAV images were cropped to 416 × 416 pixels for sorghum head detection, which resulted in duplicate identification and secondary counting of sorghum heads in the cropped edge regions, resulting in overestimation of the results [4,5]. In subsequent studies, more efficient image processing algorithms and image fusion counting algorithms should be used to avoid duplicate detection and secondary counting problems.

The main indicators that affect the detection accuracy of sorghum heads themselves are planting density and growth period [3,4,74]. (1) In terms of planting density, it has been generally accepted that as planting density increases, crops overlap and shade each other, which, in severe cases, causes some duplicate detection, a large number of missed detections, underestimation of crop counts, and a reduction in the performance of DL methods [5,8,10,18]. However, the opposite conclusion was drawn in this study and it was found that the detection accuracy tended to be worse for moderate and sparse conditions than for dense conditions (Figure 6). The main reasons involved two aspects. First, the sample plot of this experiment was not planted with dense sorghum varieties, and the density of sorghum planted was only 75,000–82,500 plants/ha. There were few problems with sorghum heads overlapping and shading each other. At the same time, affected by the high salinity of the coastal soil, the overall plant growth condition was moderate, and there was a plant loss phenomenon. Under the condition of relatively dense sorghum heads, the methods had fewer missed detections and over-detections, and had their best detection accuracy. Second, for sorghum heads in sparse conditions, the soil, weeds, and other backgrounds (discussed in the next paragraph) could introduce noise, causing a significant decrease in detection accuracy. (2) In terms of growth period, the existing studies have generally argued that the texture information of crop spikes at the flowering stage is not obvious and would result in poor detection accuracy due to insufficient performance of features learned by DL methods [3,4], while the texture information of crop spikes at the filling-ripening stage (especially late filling stage) is obvious and differs significantly from the plant canopy, yielding better detection accuracy [15,74,75]. The UAV acquisition date for this study was also selected for the maturity stage (3 October) when the difference between the sorghum heads and the plant canopy is large, which was consistent with the growth stage. However, there was some immature discoloration of the sorghum head in the grouting stage due to ground conditions, which introduced detection errors. A separate study is needed for different colors of sorghum heads. In addition to these two main factors, different crop varieties also affect the detection accuracy of DL to a certain extent [3], but the sorghum samples used in this study were of the same variety, so variety differences were not a factor in detection accuracy in this study.

The factors that affect the accuracy of sorghum head detection in terms of environmental context mainly include the meteorological conditions of insolation [76], wind speed [5], and soil and weeds [6]. Previous studies have shown that strong insolation can seriously affect the quality of UAV images. Strong insolation could bring shadows to high spatial resolution UAV images and it could have a serious negative impact on crop head detection, leading to false detection [15,76]. Meanwhile, under strong direct light conditions, highlighted areas of soil, weeds, and leaves can have duplicate misdetection with sorghum heads in UAV images [6,21], so, in practice, it is best to choose cloudy or soft light conditions to reduce shadows for UAV data acquisition [6,15]. Wind speed mainly affects the image quality of the UAV imagery and thus indirectly affects the accuracy of sorghum head detection [5,77]. The UAV in this study was flown in a Force 2 wind, which had a small impact on the quality of the images but not enough to hinder the subsequent research conducted on sorghum head detection and counting. The impact of soil and weeds was mainly due to their color and texture being similar to sorghum heads [5,77], which led to errors in detection. Furthermore, the lack of effective field management at the experimental site due to the epidemic had resulted in sparse sorghum plants mixed with a many dense weeds such as reeds and alkali puffs. The presence of a large number of plant tassels and panicles impeded the counting of sorghum heads, resulting in poorer detection accuracy under sparse conditions. Subsequent studies could be conducted with the introduction of image segmentation techniques to remove background effects such as soil and weeds to enhance accuracy.

6. Conclusions

In this study, we evaluated the performance differences of three DL methods, EfficientDet, SSD, and YOLOv4, for sorghum head detection in RGB UAV imagery. Based on this, we further analyzed the effects of model parameters such as overlap ratio, confidence, and IoU and concluded the following.

(1) Among the three DL methods, YOLOv4 had the highest accuracy in sorghum head detection, with a detection rate of positive samples at 94.12%, P of 97.62%, R of 64.20%,

and AP and F1 scores of 84.51% and 0.77, respectively. Although the average elapsed time of the training epoch was 32 s, which was not as efficient as EfficientDet and SSD, the image detection time was 0.0451, which was more efficient than EfficientDet.

(2) For the analysis of the model parameters, it was concluded that the highest sorghum head detection accuracy was obtained when the overlap ratios, confidence, and IoU were each 0.3.

Although YOLOv4 is a relatively optimal method for sorghum heads detection and can obtain the best detection results when the overlap, confidence, and IoU are all set at 0.3, it is still necessary to focus on the algorithm structure, prediction frame size, training efficiency, and repetition detection issues to further improve the algorithm performance for sorghum heads detection and counting in subsequent studies for the future monitoring and yield estimation of sorghum in the field using UAV remote sensing.

Author Contributions: H.L. and C.H. designed the experiments. P.W. collected the data and completed the program code. H.L. wrote the manuscript. C.H. and H.L. reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (41801353), and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19060302 and XDA23050101).

Acknowledgments: We thank Jinyong Lou and Jinbang Peng in our institute for providing the basic meteorological information of the experiment station.

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

References

1. Lipper, L.; Thornton, P.; Campbell, B.M.; Baedeker, T.; Braimoh, A.; Bwalya, M.; Caron, P.; Cattaneo, A.; Garrity, D.; Henry, K. Climate-smart agriculture for food security. *Nat. Clim. Chang.* **2014**, *4*, 1068–1072. [CrossRef]
2. Li, H.; Chen, Z.; Liu, G.; Jiang, Z.; Huang, C. Improving Winter Wheat Yield Estimation from the CERES-Wheat Model to Assimilate Leaf Area Index with Different Assimilation Methods and Spatio-Temporal Scales. *Remote Sens.* **2017**, *9*, 190. [CrossRef]
3. Xu, X.; Li, H.; Yin, F.; Xi, L.; Qiao, H.; Ma, Z.; Shen, S.; Jiang, B.; Ma, X. Wheat ear counting using K-means clustering segmentation and convolutional neural network. *Plant Methods* **2020**, *16*, 106. [CrossRef] [PubMed]
4. Ma, J.; Li, Y.; Liu, H.; Wu, Y.; Zhang, L. Towards improved accuracy of UAV-based wheat ears counting: A transfer learning method of the ground-based fully convolutional network. *Expert Syst. Appl.* **2022**, *191*, 116226. [CrossRef]
5. Zhao, Y.; Zheng, B.; Chapman, S.C.; Laws, K.; George-Jaeggli, B.; Hammer, G.L.; Jordan, D.R.; Potgieter, A.B. Detecting Sorghum Plant and Head Features from Multispectral UAV Imagery. *Plant Phenomics* **2021**, *2021*, 9874650. [CrossRef]
6. Lin, Z.; Guo, W. Sorghum Panicle Detection and Counting Using Unmanned Aerial System Images and Deep Learning. *Front. Plant Sci.* **2020**, *11*, 534853. [CrossRef]
7. Wu, J.; Yang, G.; Yang, X.; Xu, B.; Han, L.; Zhu, Y. Automatic Counting of in situ Rice Seedlings from UAV Images Based on a Deep Fully Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 691. [CrossRef]
8. Osco, L.P.; dos Santos de Arruda, M.; 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. A CNN approach to simultaneously count plants and detect plantation-rows from UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *174*, 1–17. [CrossRef]
9. Lu, H.; Liu, L.; Li, Y.; Zhao, X.; Wang, X.; Cao, Z. TasselNetV3: Explainable Plant Counting with Guided Upsampling and Background Suppression. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4700515. [CrossRef]
10. Wei, L.; Luo, Y.; Xu, L.; Zhang, Q.; Cai, Q.; Shen, M. Deep Convolutional Neural Network for Rice Density Prescription Map at Ripening Stage Using Unmanned Aerial Vehicle-Based Remotely Sensed Images. *Remote Sens.* **2022**, *14*, 46. [CrossRef]
11. Ampatzidis, Y.; Partel, V.; Costa, L. Agroview: Cloud-based application to process, analyze and visualize UAV-collected data for precision agriculture applications utilizing artificial intelligence. *Comput. Electron. Agric.* **2020**, *174*, 105457. [CrossRef]
12. Gnädinger, F.; Schmidhalter, U. Digital Counts of Maize Plants by Unmanned Aerial Vehicles (UAVs). *Remote Sens.* **2017**, *9*, 544. [CrossRef]
13. Torres-Sánchez, J.; López-Granados, F.; Peña, J.M. An automatic object-based method for optimal thresholding in UAV images: Application for vegetation detection in herbaceous crops. *Comput. Electron. Agric.* **2015**, *114*, 43–52. [CrossRef]
14. García-Martínez, H.; Flores-Magdaleno, H.; Khalil-Gardezi, A.; Ascencio-Hernández, R.; Tijerina-Chávez, L.; Vázquez-Peña, M.A.; Mancilla-Villa, O.R. Digital Count of Corn Plants Using Images Taken by Unmanned Aerial Vehicles and Cross Correlation of Templates. *Agronomy* **2020**, *10*, 469. [CrossRef]

15. Barreto, A.; Lottes, P.; Ispizua Yamati, F.R.; Baumgarten, S.; Wolf, N.A.; Stachniss, C.; Mahlein, A.; Paulus, S. Automatic UAV-based counting of seedlings in sugar-beet field and extension to maize and strawberry. *Comput. Electron. Agric.* **2021**, *191*, 106493. [CrossRef]
16. Vong, C.N.; Conway, L.S.; Zhou, J.; Kitchen, N.R.; Sudduth, K.A. Early corn stand count of different cropping systems using UAV-imagery and deep learning. *Comput. Electron. Agric.* **2021**, *186*, 106214. [CrossRef]
17. Feng, A.; Zhou, J.; Vories, E.; Sudduth, K.A. Evaluation of Cotton Emergence Using UAV-Based Narrow-Band Spectral Imagery with Customized Image Alignment and Stitching Algorithms. *Remote Sens.* **2020**, *12*, 1764. [CrossRef]
18. Madec, S.; Jin, X.; Lu, H.; de Solan, B.; Liu, S.; Duyme, F.; Heritier, E.; Baret, F. Ear density estimation from high resolution RGB imagery using deep learning technique. *Agric. For. Meteorol.* **2019**, *264*, 225–234. [CrossRef]
19. Wang, L.; Xiang, L.; Tang, L.; Jiang, H. A Convolutional Neural Network-Based Method for Corn Stand Counting in the Field. *Sensors* **2021**, *21*, 507. [CrossRef]
20. Liu, Y.; Cen, C.; Che, Y.; Ke, R.; Ma, Y.; Ma, Y. Detection of Maize Tassels from UAV RGB Imagery with Faster R-CNN. *Remote Sens.* **2020**, *12*, 338. [CrossRef]
21. 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. A Weakly Supervised Deep Learning Framework for Sorghum Head Detection and Counting. *Plant Phenomics* **2019**, *2019*, 1525874. [CrossRef] [PubMed]
22. Guo, W.; Zheng, B.; Potgieter, A.B.; Diot, J.; Watanabe, K.; Noshita, K.; Jordan, D.R.; Wang, X.; Watson, J.; Ninomiya, S.; et al. Aerial Imagery Analysis—Quantifying Appearance and Number of Sorghum Heads for Applications in Breeding and Agronomy. *Front. Plant Sci.* **2018**, *9*, 1544. [CrossRef] [PubMed]
23. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
24. Chapelle, O.; Vapnik, V.; Bousquet, O.; Mukherjee, S. Choosing multiple parameters for support vector machines. *Mach. Learn.* **2002**, *46*, 131–159. [CrossRef]
25. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *5*, 5–32. [CrossRef]
26. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
27. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
28. Ba, L.J.; Caruana, R. Do Deep Nets Really Need to be Deep? In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; MIT Press: Cambridge, MA, USA, 2014; pp. 2654–2662.
29. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *EEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
31. Sadeghi-Tehran, P.; Virlet, N.; Ampe, E.M.; Reyns, P.; Hawkesford, M.J. DeepCount: In-Field Automatic Quantification of Wheat Spikes Using Simple Linear Iterative Clustering and Deep Convolutional Neural Networks. *Front. Plant Sci.* **2019**, *10*, 1176. [CrossRef]
32. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 589–597.
33. Li, Y.; Zhang, X.; Chen, D. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 1091–1100.
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
35. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
36. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
37. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
38. FAO. Statistical Database of the Food and Agricultural Organization of the United Nations. 2017. Available online: https://www.fao.org/faostat/en/#data (accessed on 30 May 2022).
39. Li, H.; Huang, C.; Liu, Q.; Liu, G. Accretion–Erosion Dynamics of the Yellow River Delta and the Relationships with Runoff and Sediment from 1976 to 2018. *Water* **2020**, *12*, 2992. [CrossRef]
40. Li, H.; Chen, Z.; Jiang, Z.; Sun, L.; Liu, K.; Liu, B. Temporal-spatial variation of evapotranspiration in the Yellow River Delta based on an integrated remote sensing model. *J. Appl. Remote Sens.* **2015**, *9*, 96047. [CrossRef]
41. LabelImg. 2022. Available online: https://github.com/tzutalin/labelImg (accessed on 30 May 2022).
42. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Ar, P.D.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014.

43. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning 2019, Long Beach, CA, USA, 9–15 June 2019.
44. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
45. Li, X.; Liu, C.; Dai, S.; Lian, H.; Ding, G. Scale specified single shot multibox detector. *IET Comput. Vis.* **2020**, *14*, 59–64. [CrossRef]
46. Yi, J.; Wu, P.; Metaxas, D.N. ASSD: Attentive single shot multibox detector. *Comput. Vis. Image Underst.* **2019**, *189*, 102827. [CrossRef]
47. Bochkovskiy, A.; Wang, C.; Liao, H.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
48. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
49. Tian, Y.; Yang, G.; Wang, Z.; Wang, H.; Li, E.; Liang, Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput. Electron. Agric.* **2019**, *157*, 417–426. [CrossRef]
50. Zakria, Z.; Deng, J.; Kumar, R.; Khokhar, M.S.; Cai, J.; Kumar, J. Multiscale and Direction Target Detecting in Remote Sensing Images via Modified YOLO-v4. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1039–1048. [CrossRef]
51. Wu, D.; Lv, S.; Jiang, M.; Song, H. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Comput. Electron. Agric.* **2020**, *178*, 105742. [CrossRef]
52. Jiang, J.; Fu, X.; Qin, R.; Wang, X.; Ma, Z. High-Speed Lightweight Ship Detection Algorithm Based on YOLO-V4 for Three-Channels RGB SAR Image. *Remote Sens.* **2021**, *13*, 1909. [CrossRef]
53. Shi, P.; Jiang, Q.; Shi, C.; Xi, J.; Tao, G.; Zhang, S.; Zhang, Z.; Liu, B.; Gao, X.; Wu, Q. Oil Well Detection via Large-Scale and High-Resolution Remote Sensing Images Based on Improved YOLO v4. *Remote Sens.* **2021**, *13*, 3243. [CrossRef]
54. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
55. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014.
56. Mekhalfi, M.L.; Nicolo, C.; Bazi, Y.; Rahhal, M.M.A.; Alsharif, N.A.; Maghayreh, E.A. Contrasting YOLOv5, Transformer, and EfficientDet Detectors for Crop Circle Detection in Desert. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3003205. [CrossRef]
57. Qin, P.; Cai, Y.; Liu, J.; Fan, P.; Sun, M. Multilayer Feature Extraction Network for Military Ship Detection from High-Resolution Optical Remote Sensing Images. *EEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11058–11069. [CrossRef]
58. Kim, J.; Park, I.; Kim, S. A Fusion Framework for Multi-Spectral Pedestrian Detection using EfficientDet. In Proceedings of the 2021 21st International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea, 12–15 October 2021; pp. 1111–1113.
59. Majchrowska, S.; Mikołajczyk, A.; Ferlin, M.; Klawikowska, Z.; Plantykow, M.A.; Kwasigroch, A.; Majek, K. Deep learning-based waste detection in natural and urban environments. *Waste Manag.* **2022**, *138*, 274–284. [CrossRef] [PubMed]
60. Ammar, A.; Koubaa, A.; Benjdira, B. Deep-Learning-Based Automated Palm Tree Counting and Geolocation in Large Farms from Aerial Geotagged Images. *Agronomy* **2021**, *11*, 1458. [CrossRef]
61. Jia, S.; Diao, C.; Zhang, G.; Dun, A.; Sun, Y.; Li, X.; Zhang, X. Object Detection Based on the Improved Single Shot MultiBox Detector. In Proceedings of the International Symposium on Power Electronics and Control Engineering (ISPECE), Xi'an, China, 28–30 December 2018.
62. Aziz, L.; Haji Salam, M.S.B.; Sheikh, U.U.; Ayub, S. Exploring Deep Learning-Based Architecture, Strategies, Applications and Current Trends in Generic Object Detection: A Comprehensive Review. *IEEE Access* **2020**, *8*, 170461–170495. [CrossRef]
63. Zhang, C.; Li, T.; Zhang, W. The Detection of Impurity Content in Machine-Picked Seed Cotton Based on Image Processing and Improved YOLO V4. *Agronomy* **2022**, *12*, 66. [CrossRef]
64. Gai, R.; Chen, N.; Yuan, H. A detection algorithm for cherry fruits based on the improved YOLO-v4 model. *Neural Comput. Appl.* **2021**. [CrossRef]
65. Li, M.; Zhang, Z.; Lei, L.; Wang, X.; Guo, X. Agricultural Greenhouses Detection in High-Resolution Satellite Images Based on Convolutional Neural Networks: Comparison of Faster R-CNN, YOLO v3 and SSD. *Sensors* **2020**, *20*, 4938. [CrossRef]
66. Uzal, L.C.; Grinblat, G.L.; Namías, R.; Larese, M.G.; Bianchi, J.S.; Morandi, E.N.; Granitto, P.M. Seed-per-pod estimation for plant breeding using deep learning. *Comput. Electron. Agric.* **2018**, *150*, 196–204. [CrossRef]
67. SC Chapman, M.; Olsen, P.; Ramamurthy, K.N. Counting and Segmenting Sorghum Heads. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019.
68. Chandra, A.L.; Desai, S.V.; Balasubramanian, V.N.; Ninomiya, S.; Guo, W. Active learning with point supervision for cost-effective panicle detection in cereal crops. *Plant Methods* **2020**, *16*, 34. [CrossRef] [PubMed]
69. Velumani, K.; Lopez-Lozano, R.; Madec, S.; Guo, W.; Gillet, J.; Comar, A.; Baret, F. Estimates of Maize Plant Density from UAV RGB Images Using Faster-RCNN Detection Model: Impact of the Spatial Resolution. *Plant Phenomics* **2021**, *2021*, 9824843. [CrossRef] [PubMed]
70. Pang, Y.; Shi, Y.; Gao, S.; Jiang, F.; Veeranampalayam-Sivakumar, A.; Thompson, L.; Luck, J.; Liu, C. Improved crop row detection with deep neural network for early-season maize stand count in UAV imagery. *Comput. Electron. Agric.* **2020**, *178*, 105766. [CrossRef]
71. Malambo, L.; Popescu, S.; Ku, N.; Rooney, W.; Zhou, T.; Moore, S. A Deep Learning Semantic Segmentation-Based Approach for Field-Level Sorghum Panicle Counting. *Remote Sens.* **2019**, *11*, 2939. [CrossRef]

72. Koziarski, M.; Cyganek, B. Impact of Low Resolution on Image Recognition with Deep Neural Networks: An Experimental Study. *Int. J. Appl. Math. Comput. Sci.* **2018**, *28*, 735–744. [CrossRef]
73. Dodge, S.; Karam, L. Understanding How Image Quality Affects Deep Neural Networks. In Proceedings of the 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 6–8 June 2016; pp. 1–6.
74. Fernandez Gallego, J.A.; Lootens, P.; Borra Serrano, I.; Derycke, V.; Haesaert, G.; Roldán Ruiz, I.; Araus, J.L.; Kefauver, S.C. Automatic wheat ear counting using machine learning based on RGB UAV imagery. *Plant J.* **2020**, *103*, 1603–1613. [CrossRef]
75. Wilke, N.; Siegmann, B.; Postma, J.A.; Muller, O.; Krieger, V.; Pude, R.; Rascher, U. Assessment of plant density for barley and wheat using UAV multispectral imagery for high-throughput field phenotyping. *Comput. Electron. Agric.* **2021**, *189*, 106380. [CrossRef]
76. Chopin, J.; Kumar, P.; Miklavcic, S.J. Land-based crop phenotyping by image analysis: Consistent canopy characterization from inconsistent field illumination. *Plant Methods* **2018**, *14*, 39. [CrossRef]
77. Lee, U.; Chang, S.; Putra, G.A.; Kim, H.; Kim, D.H. An automated, high-throughput plant phenotyping system using machine learning-based plant segmentation and image analysis. *PLoS ONE* **2018**, *13*, e196615. [CrossRef]