forests

MDPI

Article

Accurate Recognition of Jujube Tree Trunks Based on Contrast Limited Adaptive Histogram Equalization Image Enhancement and Improved YOLOv8

Shunkang Ling [1,†], Nianyi Wang [2,†], Jingbin Li [1,*] and Longpeng Ding [1]

[1] College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832003, China; 20212009037@stu.shzu.edu.cn (S.L.); dy2016@shzu.edu.cn (L.D.)
[2] College of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China; ny.wang@stu.xjtu.edu.cn
* Correspondence: lijingbin@shzu.edu.cn
† These authors contributed equally to this work.

Abstract: The accurate recognition of tree trunks is a prerequisite for precision orchard yield estimation. Facing the practical problems of complex orchard environment and large data flow, the existing object detection schemes suffer from key issues such as poor data quality, low timeliness and accuracy, and weak generalization ability. In this paper, an improved YOLOv8 is designed on the basis of data flow screening and enhancement for lightweight jujube tree trunk accurate detection. Firstly, the key frame extraction algorithm was proposed and utilized to efficiently screen the effective data. Secondly, the CLAHE image data enhancement method was proposed and used to enhance the data quality. Finally, the backbone of the YOLOv8 model was replaced with a GhostNetv2 structure for lightweight transformation, also introducing the improved CA_H attention mechanism. Extensive comparison and ablation results show that the average precision of the quality-enhanced dataset over that of the original dataset increases from 81.2% to 90.1%, and the YOLOv8s-GhostNetv2-CA_H model proposed in this paper reduces the model size by 19.5% compared to that of the YOLOv8s base model, with precision increasing by 2.4% to 92.3%, recall increasing by 1.4%, mAP@0.5 increasing by 1.8%, and FPS being 17.1% faster.

Keywords: orchard trunk extraction; image enhancement; attention mechanism; keyframe extraction

Citation: Ling, S.; Wang, N.; Li, J.; Ding, L. Accurate Recognition of Jujube Tree Trunks Based on Contrast Limited Adaptive Histogram Equalization Image Enhancement and Improved YOLOv8. *Forests* **2024**, *15*, 625. https://doi.org/10.3390/f15040625

Received: 21 February 2024
Revised: 20 March 2024
Accepted: 27 March 2024
Published: 29 March 2024

1. Introduction

In recent years, technological advances in the fields of machine vision, data mining, image processing, and computer communication have greatly promoted the development of many applications and fields such as intelligent agriculture [1,2]. Agricultural object detection and image acquisition is one of the important technical means to realize agricultural intelligence and precision agriculture [3–5]. The forest and fruit industry plays an important role in agriculture, which is not only vital to economic development but also contributes to ecological protection and the rural economy. The development and management of the jujube garden, as the main forest fruit industry in Xinjiang, is of strategic significance and has a positive impact on improving yield and quality, resource conservation and environmental protection, and farmers' income and economic benefits [6]. The overall picture of jujube trees provides the necessary database for the digital monitoring of the jujube garden [7,8]. The research site is located in the large-scale jujube garden of 224 regiments in South Xinjiang, and its planting area is as high as 103.3 km^2, from which the efficient and accurate collection of jujube tree pictures is the focus of this paper. In order to realize the automatic identification and picture collection of the whole growth cycle of jujube trees, it is found that only the trunk part of trees hardly changes morphologically in different growth stages, so the target recognition object is selected as the trunk part of jujube trees. This paper chooses March as the test period, when the tree crown is partly

ungrown and the branches greatly interfere with the recognition of the trunk by the target detection model, so as to fully verify the detection performance of the model algorithm.

In order to fully and efficiently acquire jujube garden data, this paper utilizes a UAV-mounted panoramic camera to capture a video dataset of the jujube tree. It is worth noting that we want to quickly get the overall picture of each tree, i.e., the key frames from the video. In video sequences from field operations, there may exist a large amount of low-quality, redundant information between consecutive frames in the video sequences, which will surely challenge computer communication and algorithmic efficiency if we choose to process all the information [9]. However, boosting computational resources is not a preferred option for field computing platforms. In this paper, we start from the algorithm, discarding the redundant information [10], adopting the key frame extraction technique to quickly extract effective crop sample information, and combining it with a suitable image quality enhancement technique to efficiently and accurately train an object detection model [11–14]. By improving the efficiency of the computerized information transfer, we can achieve refined management and assisted decision-making, ultimately improving the efficiency and accuracy of agricultural production [15–17].

Target recognition models are categorized into two types: single-phase recognition and dual-phase recognition, with YOLO and Faster R-CNN standing out as leading exemplars within each category [18]. Faster R-CNN excels in accuracy due to its two-stage process and can adeptly handle small or deformed targets, but it requires more computational resources. In contrast, YOLO is faster and more efficient, achieving recognition with just a single forward pass, making it highly suitable for real-time applications [19,20]. We conducted comparative experiments using these two mainstream models. For this study, given the large scale of the jujube garden, we aim to achieve rapid identification of jujube trees; therefore, the YOLO recognition algorithm was chosen as the base model for improvement.

Due to the energy and space constraints of field equipment, it is often impossible to provide high-performance computer communication capabilities. The larger workloads and low data quality are a challenge to hardware resources; the jujube garden environment information is complicated; the image size, resolution, and frame rate are high; and the requirements for hardware resources are extremely high [21,22], so the object detection model needs to undergo lightweight processing. Lightweight technology, which is used in the field to reduce field equipment computing resource requirements at the same time, can achieve fast detection in the face of large-scale planting detection tasks, and it can have a higher operational efficiency for a moving object due to its fast detection ability that effectively prevents the loss of an object [23,24]. However, lightweighting somewhat reduces the accuracy and generalization of the object detection model, which can cause considerable losses in real production activities, so we introduced and improved the Channel Attention (CA) attention module to help the model capture spatial correlations and focus more attention on the region where the trunk of the date palm tree appears to achieve accurate detection [25–27].

In summary, this article makes the following contributions:

- An automatic key frame extraction algorithm based on the object detection model is designed, which eliminates a large amount of redundant data and realizes efficient batch sample jujube acquisition from the front and side views of jujube trees in large-scale gardens.
- Focusing on the situation whereby the low quality of data collected in the jujube garden environment with light affects model detection, the CLAHE method was utilized to enhance the image quality of the dataset, which effectively improves the problem of low data quality caused by the lack of light in a dark-side image and the loss of details in a bright-side image's overexposure.
- A new YOLOv8 network structure based on the GhostNetv2 module is incorporated, which not only improves the speed of object detection for jujube tree trunks but also achieves a better balance between accuracy and efficiency than other existing methods.

- The CA attention mechanism is introduced and improved for the CA-H attention mechanism, which is reasonably integrated into the YOLOv8 trunk network, which helps the model to more accurately locate and identify the object detection region that needs more attention and improves the detection accuracy.

2. Dataset Construction

2.1. Dataset Image Acquisition

This study was carried out in the jujube cultivation demonstration garden of the 224th Regiment of Kunyu City, the 14th Division of the Xinjiang Production and Construction Corps (37°12′–37°21′ N, 79°15′–79°20′ E), located in the southwestern part of the Tarim Basin of the Hotan Region, which is part of the hinterland of the Taklamakan Desert, and the topography of which is characterized by sand dunes, with a flat terrain throughout the region. The experimental jujube garden adopts a large-scale "dwarf and close planting" pattern. Our team conducted research in August 2022, finding that the trees stand approximately 1.7 m tall. They are arranged in groups of four rows, with a spacing of 2 m within each group, and a 4 m wide mechanized operation corridor between each group. The arrangement of the jujube trees is uniform, exhibiting high levels of flatness and straightness, which create advantageous conditions for the planning of drone flight paths and data collection, as shown in Figure 1, the red line indicates the straightness of the tree rows. The data collection time for the jujube tree recognition test was from 10:00 to 13:00 on 20 March 2023. The data from the local meteorological bureau indicated that the sunshine time on that day was from 8:44 to 20:51, and the noon time was 14:47, with the highest solar radiation amounting to 800 w/m^2 and the highest light intensity being more than 18,000 lx.

Figure 1. Schematic diagram of straightness of standardized jujube trees.

Jujube tree aerial survey data are collected by the UAV-mounted panoramic camera along the route set up in the jujube garden. The UAV selects DJI's Mavic3 as the mounting platform for the panoramic camera, which has a vertical and horizontal accuracy of ±0.5 m when working normally with satellite positioning; below the platform, Insta's ONE X2 panoramic camera is mounted as the jujube tree image acquisition equipment, which weighs 149 g, has an aperture of F2.0, has an equivalent focal length of 7.2 mm, and has an image setting of 30 fps 5.7 K. The height of the UAV route is set at 1.4 m above the center of the jujube tree crown, and the path is set as a uniform linear flight along the center of the

jujube tree rows, and the route as a whole becomes an "S" shape; the route plan is shown in Figure 2.

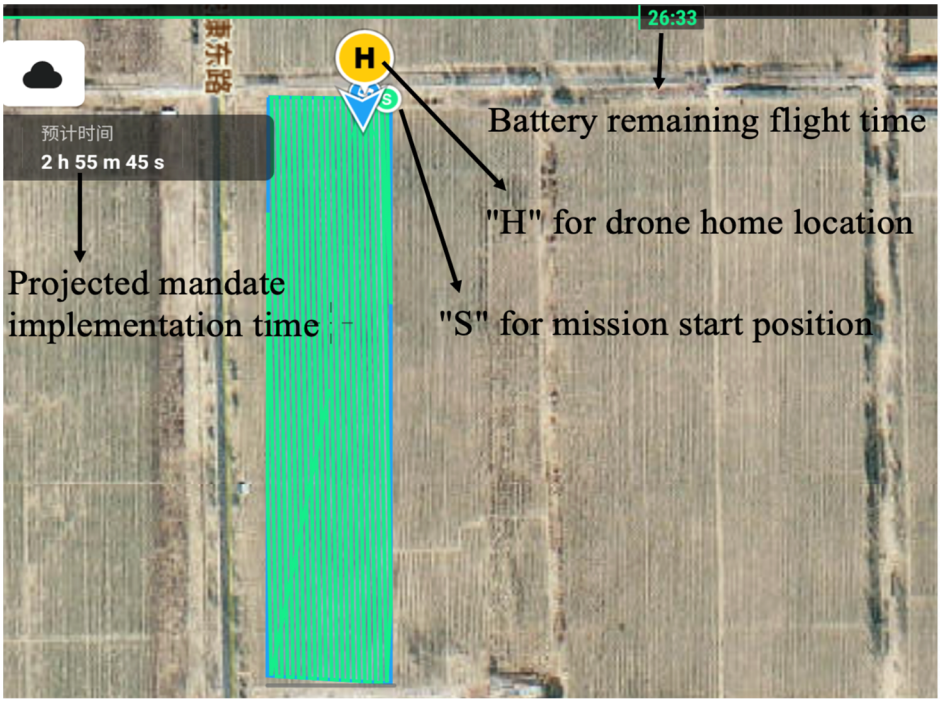

Figure 2. Drone route planning schematic.

In the process of video data acquisition, the panoramic camera has its unique advantages and characteristics, which can both capture a larger and more comprehensive image of the environment and also simultaneously capture videos from all angles; this reduces test errors and improves the reliability of the test for scientific research and also provides more accurate and detailed data materials, which greatly improve the efficiency of the work. However, the amount of raw data collected is huge and there is too much redundant and useless information.

A polar coordinate mapping model is used to remap the image into a form free of aberrations since the fisheye lens that captures the image inevitably produces aberrations in physical imaging, which distort the position and shape of the object in space, causing the image to appear convex and distorted, which can lead to failure in object detection [28]. In polar coordinate mapping, the center point of the image is the origin in the coordinate system. A line extending outward from the center point intersects a circle, and the individual arcs can be divided into small blocks, each corresponding to an uncorrected pixel point. By calculating the distance from the uncorrected pixel point to the center point, the pixel value at the correct position can be calculated, and thus correction can be achieved.

The initial data of the jujube tree aerial survey is a panoramic image, and the image pre-processing process is as follows: Firstly, the panoramic image is divided into two 180° wide-angle fisheye images according to the direction perpendicular to the rows of jujube trees; then, the polar coordinate mapping is used to correct the distortion of the two segments of the image, and the deformed and stretched parts of the surrounding area are cropped. Finally, the image scale is set to 4:3 and the field of view is adjusted so that the height of the screen can completely accommodate the standardized jujube tree, and the front and side view image data of jujube tree are obtained, and the whole data acquisition process is shown in Figure 3.

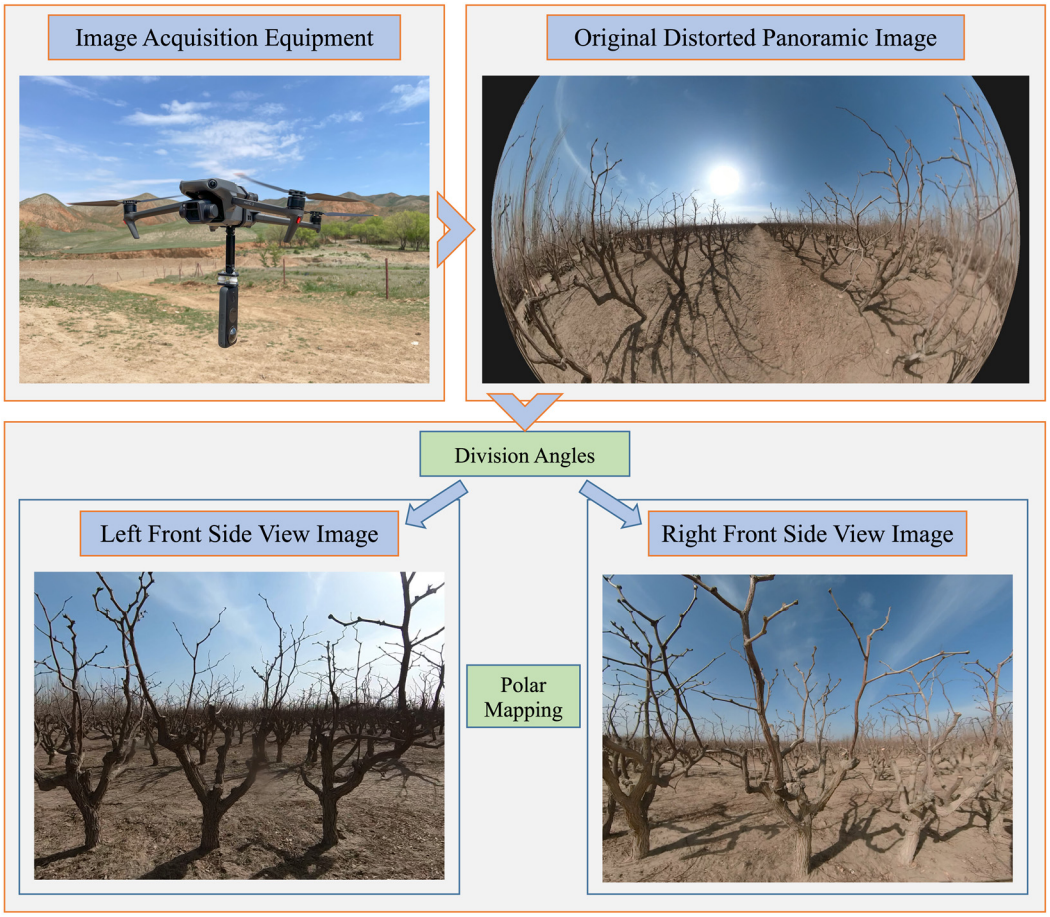

Figure 3. Data acquisition process.

2.2. Design of Automatic Key Frame Extraction Algorithm

Object detection technology cannot be separated from much high-quality dataset training; in order to efficiently realize the construction of a dataset in the field, the efficient and accurate extraction of the required key frame information from large-scale video sequences under limited computational resources is an urgent problem to be solved. There is unequal spacing of the jujube trees in the jujube tree orchard, and, at the same time, the UAV cannot realize the complete fixed speed during the image acquisition process, so it cannot use the conventional extraction strategies, such as the Fixed Frame Rate, Time Interval method, etc. [29]. It is necessary to use the object detection model as the core of automatic key frame extraction, and the extraction of the dataset can be constructed to continue to train the object detection model with high accuracy so as to realize a benign closed loop. The specific program of the automatic extraction algorithm for the key frames of jujube tree front and side views designed in this paper is as follows:

- Initial environment configuration: build the algorithm development environment based on the Pytorch11.0 framework; then, import the toolkits numpy, tqdm, and supervision, etc., in order to realize the functions of data analysis, image processing, and data visualization, etc., of which the supervision toolkit is used as a keyframe that identifies the jujube tree reaching the centerline of the field of view and implements the cross-line extraction function.
- Install the object detection model: the first object detection model needs to utilize the traditional training method, i.e., by manually obtaining a certain amount of jujube tree trunk pictures, labeling, and training an object detection model to have initial trunk detection capability.
- Frame-by-frame detection: Design the path of the video to be detected, and utilize the pre-trained object detection model to detect the trunk. Encapsulate the frame-by-frame

detection process into the "Peocess_frame" function, and output the visualized and configured image.

- Cross-line counting: run the frame-by-frame detection function on the video, use supervision to parse the prediction results, traverse all the objects on the screen, and draw the visualization effect of object detection, which is combined with the detection line in the center of the screen, to determine whether the object is over the line and to count the number of objects to be displayed in the visualization.
- Key frame extraction: Firstly, when the trunk object is detected in the image data, use the bounding box function to obtain the position information of the trunk, and then calculate the coordinates of the center point of the area where the trunk is located. Secondly, set up the trigger conditions, and when the coordinates of the center of the trunk are detected to have passed through the vertical line of the screen, then intercept the front and side view of the jujube tree. The key frame image is also labeled with the corresponding counting number and the size of image is cropped in order to make the image complete retention of the corresponding jujube tree at the same time to prevent excessive interference with the information, with the size of the cropped image being 1:1, as shown in Figure 4.

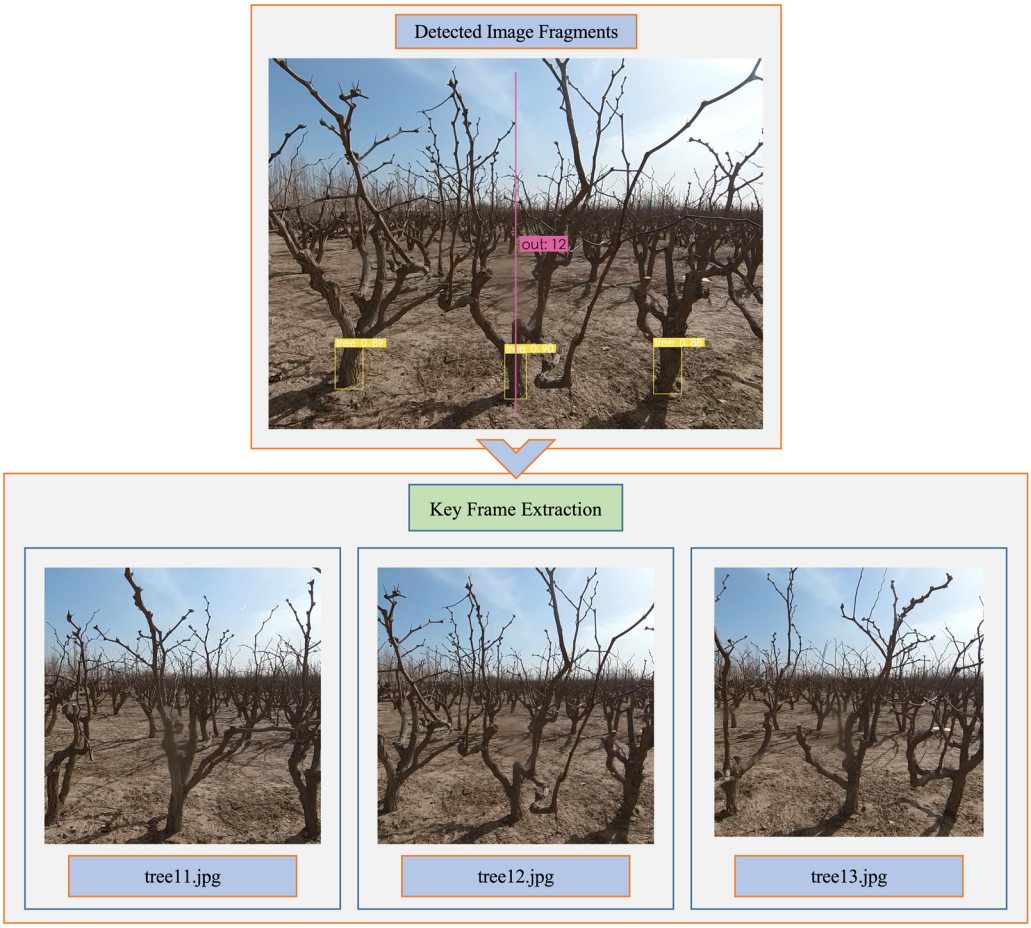

Figure 4. Extract keyframe images.

The whole process can be carried out directly in the code, which realizes the automated extraction algorithm. It can be seen that the key frame images with small data volume are mined from the panoramic image data with huge data volume, which greatly improves the dataset's construction efficiency and training quality.

2.3. Dataset Image Enhancement

Due to the high light intensity of the environment in the area where the jujube garden is located, the collected image data of the bright side and the dark side of the jujube trees have the problem of large differences in contrast, brightness, and texture detail information. Excessively low data quality will lead to the subsequent object detection model having difficulty in adequately learning the object features during the training process and thus will create difficulties for detection accuracy, so image data enhancement is an important aspect [30]. In this paper, according to the characteristics of the problems existing in the actual task, the Contrast Limited Adaptive Histogram Equalization (CLAHE) method is used to enhance the image data, and its core idea is to equalize the local histogram of the image and limit the change of the contrast so as to enhance the contrast of the image and protect the details of the image information [31]. The specific operation steps are as follows:

- Image division: First divide the image into small, non-overlapping rectangular regions; the size of these sub-regions is usually 8×8, 16×16, etc. The larger the number of pixels, the more obvious the enhancement effect is, but more information about the corresponding image details is lost. In OpenCV, the default tile size is 8×8.
- Local histogram equalization: Convert the RGB color space to grayscale HSV space, which is more suitable for brightness and contrast processing, for each small block; then, calculate its grayscale histogram, calculate the mapping function with this histogram, and apply this function to each region. And further calculate the cumulative distribution function (CDF) of the histogram.
- Contrast limitation: In order to prevent over-enhancement (resulting in noise being amplified) caused by too many values of certain pixels, the frequency of pixels exceeding a predetermined threshold T (contrast limiting parameter) in the original block histogram Figure 5a is "truncated" and the "truncated" portion is evenly distributed among other pixels to obtain the modified histogram, as shown in Figure 5b, where A denotes the pixels equally distributed in each gray level and M denotes the gray value. The principle process is shown in Figure 5.
- Pixel mapping: using the mapping relationship between the image pixels and the transformation function of the gray level of the partitioned region, an interpolation operation is applied to solve the gray level value of the corresponding pixel in order to eliminate the "blocky" image according to the number of neighboring points; the change function is 4, so bilinear interpolation is carried out between the partitioned sub-regions.
- Interpolation Smoothing: Since images are divided into multiple small sub-regions for processing, the direct application of histogram equalization may produce significant boundary effects between adjacent sub-regions [32]. To solve this problem, we use CLAHE with bilinear interpolation to smoothen the transition between neighboring subregions to ensure the continuity and smoothness of the image.
- Merging results: all the processed sub-regions are recombined into a complete image, the processed image is converted back to the RGB color space to complete the image data enhancement process, and finally the effect after enhancement by the CLAHE method is shown in Figure 6.

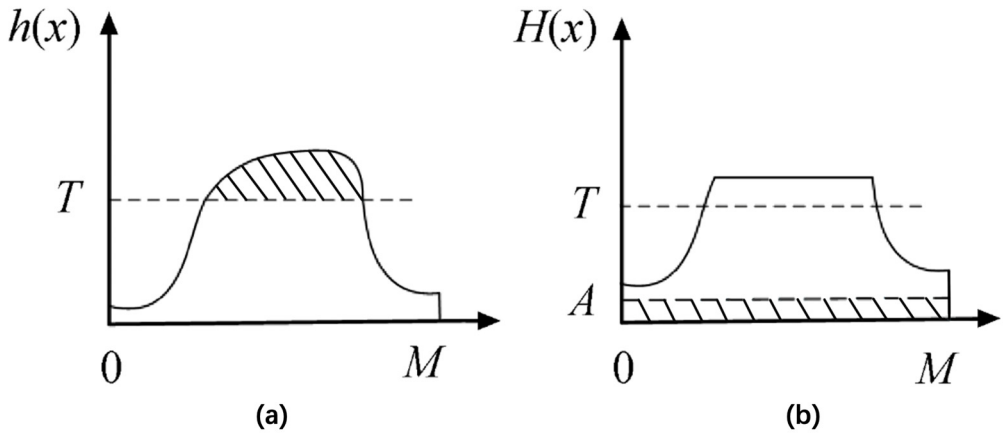

Figure 5. CLAHE truncated distribution principle ((**a**) denotes the original block histogram, (**b**) denotes the modified block histogram).

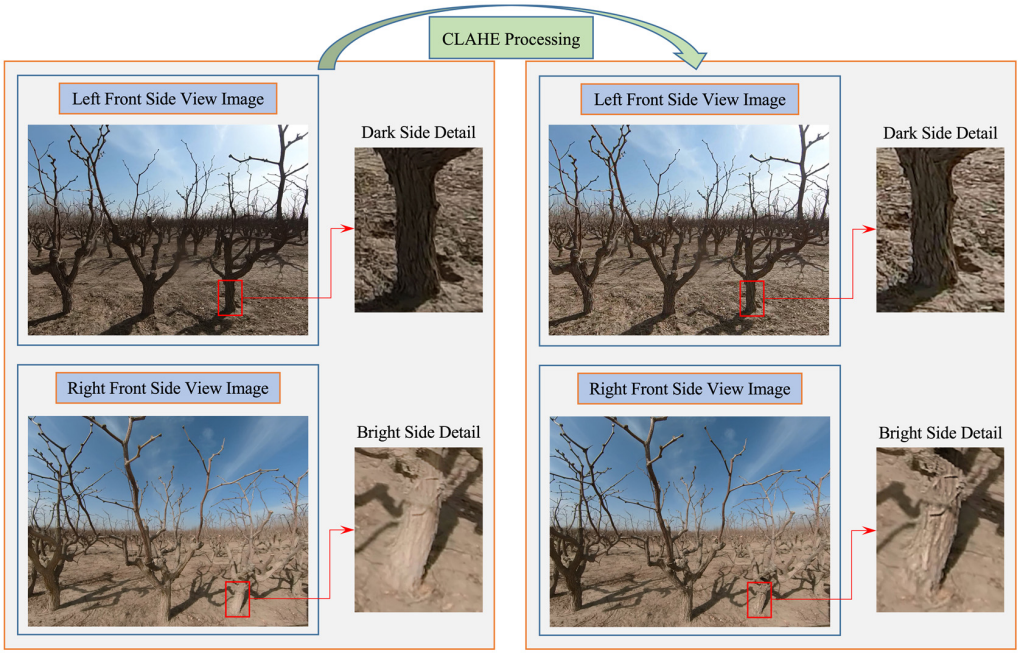

Figure 6. Comparison of CLAHE enhancement effects.

3. Methods

3.1. YOLOv8 Algorithm Structure

YOLOv8 is an efficient and accurate object detection model, which combines advanced techniques such as feature extraction networks, multi-scale fusion, and attention mechanisms. This enables YOLOv8 to achieve fast and accurate object detection while dealing with a large number of objects [33]. For this study, efficiency is the core consideration, and the model with the minimum depth and width of the network based on YOLOv8s is chosen in this paper.

3.2. YOLOv8 Improvement of Backbone Network GhostNetv2

The traditional convolutional neural network has the problems of redundancy of feature information, large model parameters, and expensive training computation cost, etc. The backbone network of YOLOv8 is based on the CSPDarknet53 network structure, which employs a large number of convolutional layers for extracting features from the input image. Although this structure has a strong ability to extract features, the image size, resolution, and frame rate are high in the practical application of the jujube garden, and, at

the same time, because of the limitation of the equipment, the traditional network structure often can only run at a very low batch size in the practical application and the speed cannot meet the requirements.

The feature maps in traditional deep neural networks usually contain rich or even redundant information, while they reduce the computational complexity of deep neural networks by introducing the Ghost module to generate more features using fewer parameters [34]. The ordinary convolution and Ghost module convolution process are shown in Figure 7, and the specific operation of the Ghost module is to divide the original convolution layer into two parts: the first part consists of ordinary convolution, the second part generates more feature maps from the first part by cheap linear transformation represented as "φ" in the figure, and, when finally spliced, the parameters and computational complexity are thus reduced without changing the size of the output feature maps.

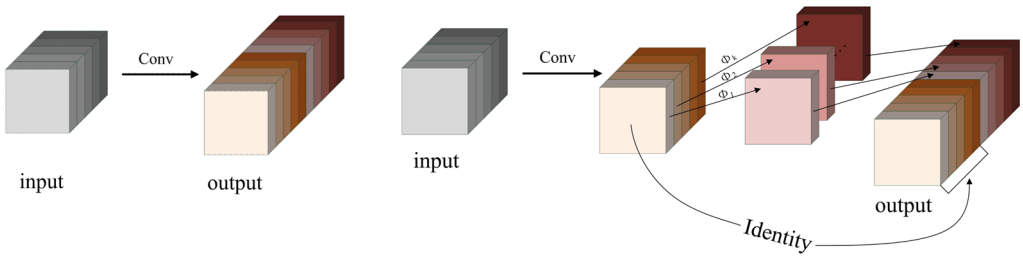

Figure 7. Ordinary convolution process and Ghost module convolution process.

There are problems in the lightweight network structure of GhostNet, for example, although the Ghost module achieves a reduction in computational cost, the representation capability is also necessarily reduced. GhostNetv2 overall uses the Ghost module and the DFC attention module and extracts the information from different viewpoints in parallel. The Ghost module reduces the computation by decreasing the size of the features [35], while the DFC attention module reduces the computation by decreasing the size of the features, as shown in Figure 8.

Figure 8. DFC attention module.

The GhostNetv2 bottleneck structure is shown in Figure 9, which is composed of a Ghost module and a DFC attention module. The DFC attention module is used to augment the output features of the Ghost module to capture the long-range dependencies between pixels in different spaces. Compared with the Self-Attention model [36], the computation process of the DFC attention module is more intuitive and simpler, in which the FC layer directly generates the attention graph and is computed as follows:

$$a_{hw} = \sum_{h',w'} F_{hw,h'w'} \odot Z_{h'w'}, \quad (1)$$

where F is the learnable weights in the FC layer and $\odot$ is the elemental multiplication. Given the feature $Z \in R^{H \times W \times C}$, it can be seen as HW tokens $Z_i \in R^C$. $\{a_{11}, a_{12}, \ldots, a_{hw}\}$ as in the generated attention map.

The DFC attention module significantly improves the expressive power of GhostNet by decomposing the fully connected layer into horizontal and vertical fully connected layers, i.e., the input features are aggregated along the horizontal and vertical directions, respectively, in order to capture the long-range dependencies along these two directions. The characteristic expressions for the horizontal direction F^H and vertical direction F^W in the calculation process of DFC attention are as follows:

$$a'_{hw} = \sum_{h'=1}^{H} F^H_{h,h'w} \odot z_{h'w}, \; h = 1, 2, \ldots, H, \; w = 1, 2, \ldots, W, \quad (2)$$

$$a_{hw} = \sum_{w'=1}^{W} F_{w,hw'}^{W} \odot a'_{hw'},\ h = 1,2,\ldots,H,\ w = 1,2,\ldots,W, \quad (3)$$

In this paper, the network structure of YOLOv8 is improved through replacement. In the fully connected layer, each pixel has a direct connection to all other pixels, resulting in a computational complexity of $O(H^2W^2)$, which is very high on high-resolution images. In contrast, DFC attention decomposes the fully connected layer into horizontal and vertical fully connected layers, which reduces the computational complexity to $O(H^2W + HW^2)$ and greatly reduces the computational effort, and, finally, the features are up-sampled by average pooling and bilinear interpolation to recover the original size of the features. This design allows GhostNetv2 to have a larger sensory field, which better captures the long-distance dependencies between different locations in the image and improves the expressive power of the model.

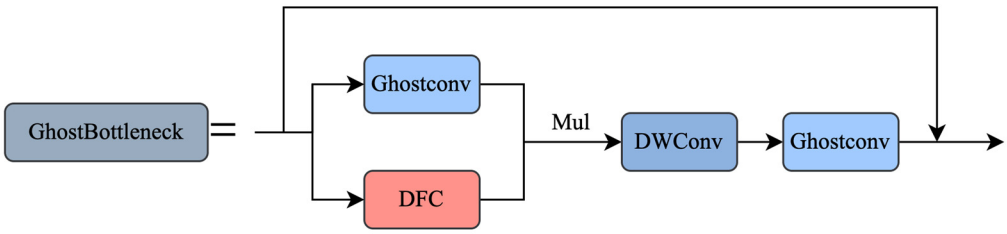

Figure 9. GhostNetv2 bottleneck structure.

Part of the Conv module is replaced with GhostConv; Conv and Bottleneck in C2f are replaced with GhostConv and GhostBottleneck, and the improved C2fGhost structure is shown in Figure 10. Not changing the size of the output feature map reduces the parameters and computational complexity improves the detection speed.

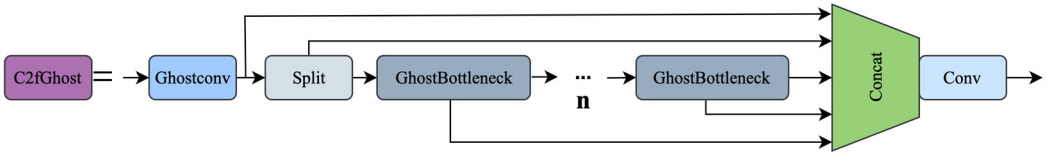

Figure 10. Improved C2fGhost structure.

3.3. YOLOv8 Improvement of the CA_H Attention Mechanism

In this paper, the object detection area of interest is the trunk area of jujube trees, and we tried to keep the height and speed of the UAV as stable as possible during the process of the image acquisition of the trunks, so in the acquired images, the object area is characterized by the following: the position of the object area in the image is the lower area and more fixed at the same time, and, secondly, the object keeps the uniform speed movement in the image. From the above considerations, it is necessary to increase the attention mechanism in conjunction with the characteristics and to suppress the accuracy degradation caused by lightweighting.

The base attention mechanism chosen in this paper is coordinate attention (CA), and the most significant feature is that it embeds the position information in the design of the mobile network. The core algorithm of the CA attention mechanism is divided into two steps: coordinate information embedding and coordinate information embedding. The core algorithm of the CA attention mechanism is divided into two steps: coordinate information embedding and coordinate attention generation.

Coordinate information embedding comprises the X Avg Pool and Y Avg Pool operations. The specific operation is as follows: first of all, the input feature maps are subjected to two one-dimensional global pooling operations, which require the use of pooling kernels with dimensions of (h,1) and (1,w) to aggregate the features along the vertical and horizon-

tal directions to obtain the two directionally aware feature maps, and the computational formulae are as follows, respectively:

$$z_c^h(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h,i), \tag{4}$$

$$z_c^w(w) = \frac{1}{H}\sum_{0\le i<H} x_c(j,w), \tag{5}$$

where given the input x, c denotes the channel, z_c^h denotes the output of the cth channel at height h, and z_c^w denotes the output of the cth channel at width w.

Coordinate attention generation is the remaining process in the structure picture, and the specific operation is as follows: firstly, the two feature maps inputted from the Concat segment are transformed using a shared 1×1 convolution to reduce the dimensions to the original C/r and to generate the feature maps F_h and F_w, where r denotes the undersampling ratio [37]; secondly, the intermediate feature maps generated after convolution are sliced, normalized with other operations, and the feature maps F are passed through the nonlinear activation function δ to obtain the intermediate feature maps f with spatial information in the horizontal and vertical directions, which are calculated as follows:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right), \tag{6}$$

Then, the intermediate feature map f is convolved to generate a feature map with the same channel as the input F. The CA module finally uses the Sigmoid activation function, which is improved in this paper by adopting the H-Sigmoid activation function to replace and form a new attention mechanism CA_H, and the structure is shown in Figure 10; the computation using the new improved method is performed to obtain the attentional weight g^h in the height direction and the attentional weight g^w in the width direction of the feature map as follows:

$$g^h = \sigma\left(F_h\left(f^h\right)\right), \tag{7}$$

$$g^w = \sigma(F_w(f^w)), \tag{8}$$

Finally, the output y_c with the weights of the attention mechanism in the height and width directions is obtained by multiplicative weighting with the following equation:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{9}$$

This attention mechanism decomposes the channel attention into two one-dimensional features encoding processes that aggregate features along the horizontal and vertical directions as shown in Figure 11. This captures long-range dependencies and simultaneously preserves precise positional information. This helps the model to more accurately localize and identify object detection regions that require more attention.

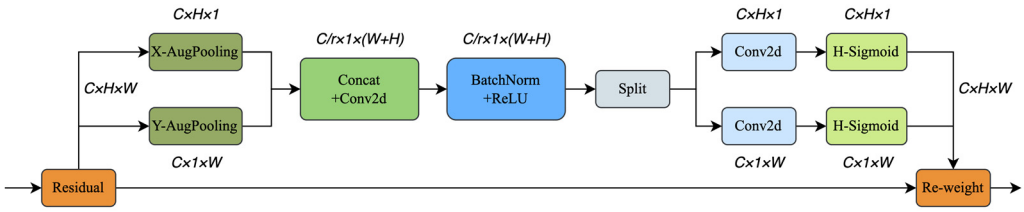

Figure 11. Improved CA_H structure.

The position information is embedded into the channel attention, taking both spatial and channel attention into account, and the improved CA_H attention mechanism is added to the end of the Neck part of the YOLOv8 network architecture, i.e., the intermediate position after the Concat segment of each up-sampling and before the Head detection layer after the multi-scale feature fusion of the three effective feature layers of the FPN and the

PAN, and so the overall network structure of the trunk detection model based on improved yolov8 is shown in Figure 12. This attentional embedding design can filter and analyze a large amount of input feature information in both channel and spatial dimensions to enhance the impact of important features and obtain performance gains [38].

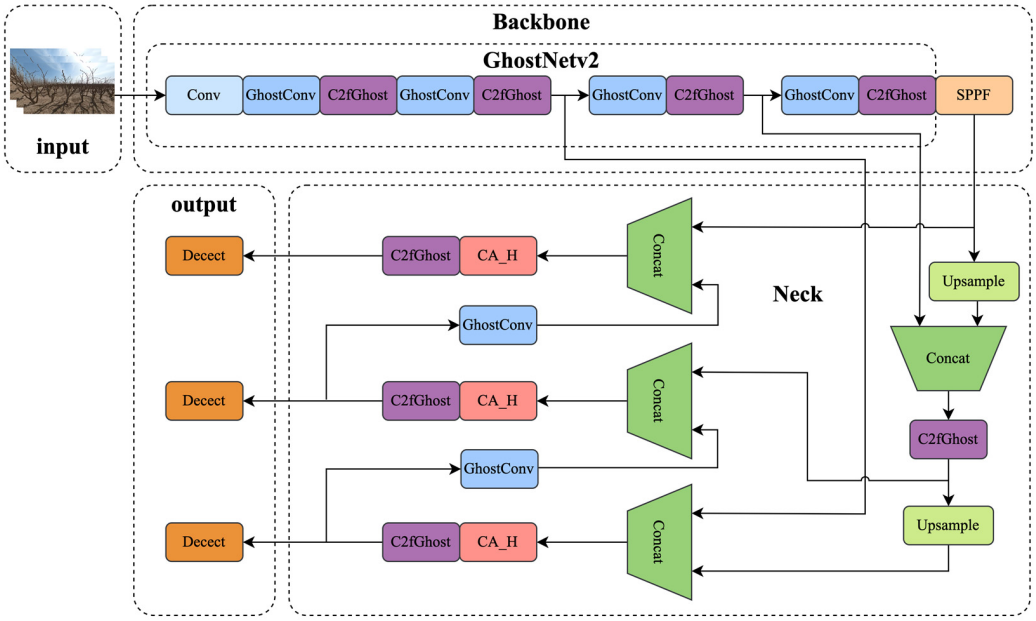

Figure 12. Overall network structure of trunk detection model.

4. Experiment Results with Relevant Analysis

4.1. Experimental Settings

In this paper, a series of algorithms formed by improvements are built under the PyTorch framework, and the hardware configuration used for the experiment is GPU NVIDIA RTX 2080Ti, and the environment configuration comprises CUDA10.2, Python 3.8, and PyTorch 1.8. Hyper-parameters are set to a batch size of eight, the training period (epochs) is 300, and the initial learning rate is 0.01.

The key frame data samples are obtained by batch screening using the key frame automatic extraction algorithm, and after the image data enhancement process, the effect of illumination is significantly eliminated to improve data quality, which in turn increases the model training efficiency and effectiveness. Since each image contains only a small number of objects due to the limitation of the viewing angle, in order to help the model learn and generalize better, it is necessary to ensure the quality and diversity of the samples, so it is necessary to extract as many key frames as possible. In this experiment, 16,000 front and side views of jujube trees were extracted, and the dataset was divided into a training set, validation set, and test set according to the ratio of 8:1:1, which were used for model training, validating, and adjusting the model hyper-parameters as well as testing the model performance, respectively.

4.2. Qualitative Evaluation

In order to quantitatively evaluate the effectiveness of the method in this paper, the evaluation metrics in this paper in the experiments are precision (P), recall (R), Intersection and Union Ratio (IoU), mean average precision (mAP), FPS, and model size.

4.3. Data Enhancement Comparison Test

In the actual application scenario, the data collection time of the jujube tree recognition test was from 10:00 to 13:00 on 20 March 2023, and the highest light intensity in the southern border region reached more than 18,000 lx during that time period. The strong light intensity led to a great difference in light between the sunny side and the shady side

of the jujube trees, which led to a sharp decrease in data quality and interfered with the training and detection of the model. Therefore, we selected the CLAHE method for data enhancement and processed the original dataset according to the actual characteristics, as shown in Figure 13. In order to verify the effect of data enhancement on the model, this paper designed an experiment to choose the YOLOv8s base model using the original dataset and the data-enhanced dataset for training, respectively, and the test results are shown in Table 1.

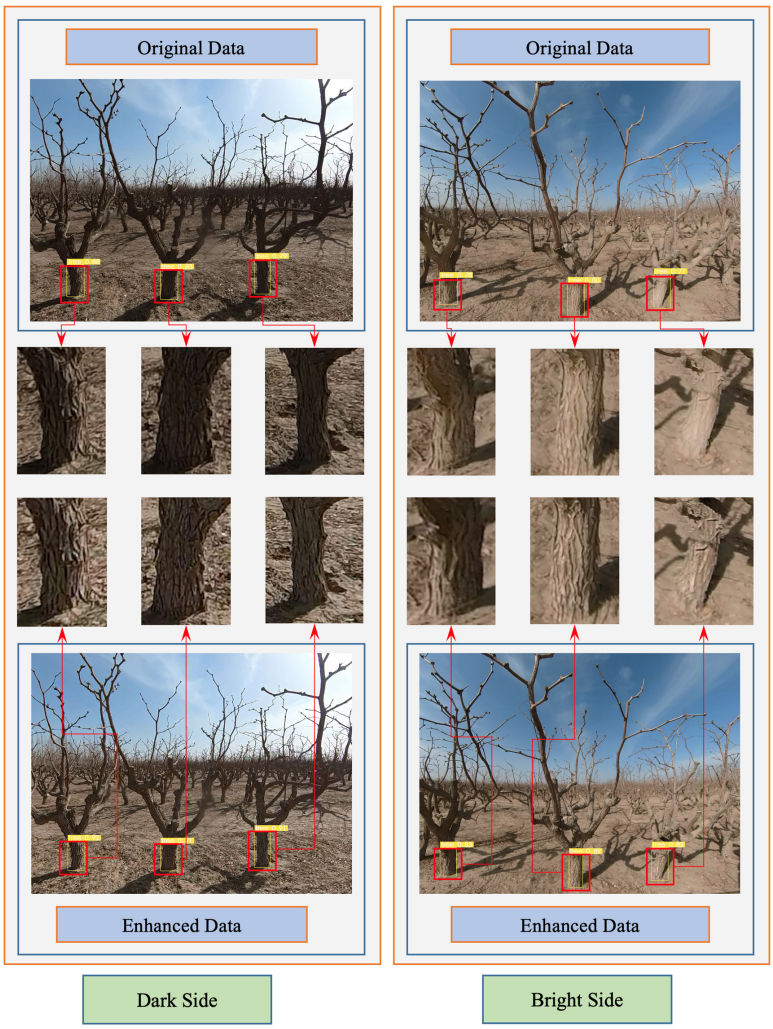

Figure 13. Image enhancement before and after comparison.

Table 1. Algorithm performance analysis for image enhancement.

Dataset	Precision (%)			mAP@0.5 (%)		
	Dark Side	Bright Side	Average	Dark Side	Bright Side	Average
Original	83.9	78.5	81.2	83.9	79.1	81.5
Enhanced	91.3	88.9	90.1	91.3	89.1	90.2

From the experimental results, it can be seen that the quality of the dataset is too low due to ambient light when using the original dataset, which is not conducive to the effective training of the model, resulting in the lower detection accuracy of the algorithmic model and a higher missing rate of detecting date palms, which is not in line with the needs of practical applications.

After data enhancement on dark side and bright side datasets, the base model performance was greatly improved in all aspects, according to which precision was improved by

8.8% and 13.2%, respectively; the average accuracy was improved by 10.9% from 81.2% to 90.1%; and the average mAP@0.5 was improved by 10.6%. From this, it can be concluded that the data enhancement method used in this paper has a greater impact on the object detection accuracy enhancement, and the degree of enhancement on the bright side is more obvious. Therefore, all subsequent experiments use the enhanced dataset.

Discussion

The main reasons for the significant improvement in data quality using the CLAHE method are as follows:

- Improving contrast: Due to the high light intensity in the Xinjiang jujube garden, there are cases of overexposure on the sunny side and underexposure on the shady side of the image, which lead to insufficient contrast of the image and have an impact on the accuracy of object detection. CLAHE is precisely the kind of enhancement method needed to address the situation, which is able to improve the contrast of the local area and make the image information more complete.
- Noise reduction: The CLAHE method is processed in chunks of the image, and histogram equalization can greatly reduce the phenomenon of excessive local contrast caused by noise or rapid changes in brightness. In addition, by limiting the contrast (contrast limiting), CLAHE amplifies the noise much less than the global histogram equalization method [39].
- Protection of details: The CLAHE method performs histogram equalization independently for each local region of the image and also ensures that the details are not over-amplified as the contrast limiting mechanism prevents too much concentration of a particular luminance value, thus protecting the details of the image as much as possible. This is especially important for object detection, such as the shape of tree trunks, texture, and other detailed features.
- Reducing the impact of brightness variation: Since the UAV is constantly traveling during the acquisition process, there are light variations. The CLAHE method for localized regions ensures that the contrast of local sub-regions is enhanced while not being affected by the brightness distribution of other regions, reducing the impact of uneven illumination on image characteristics. The accuracy and robustness of model detection is improved by the CLAHE method.

4.4. Lightweighting and Attention Mechanism Improvement Ablation Trial

The YOLOv8s-GhostNetv2-CA_H model proposed in this paper is trained on the homemade augmented jujube tree trunk detection dataset and its convergence performance is verified. The loss curves of the model in terms of class loss, confidence loss, and localization loss on both the training and test sets are shown in Figure 14. In addition, the precision, recall, and mAP@0.5 metrics of convergence are shown in Figure 15. As can be seen from Figure 14, the model's loss on the training and validation sets rapidly decreases to a lower level and then overall tends to flatten out and oscillate in a small range, indicating that the model has a good fitting ability for the task under study, with no overfitting or underfitting. As known from Figure 15, the model's performance is demonstrated by the fact that all of the model's metrics increase rapidly to higher values and then stabilize with a small amplitude.

In order to verify the effectiveness of the lightweighting and attention mechanism improvement scheme proposed in this paper, three sets of ablation experiments are designed according to each improvement module, namely, the YOLOv8s base model, the model for lightweight improvement using GhostNetv2, and the YOLOv8s-GhostNetv2-CA_H model with the addition of the CA_H attention mechanism, and the experiments are conducted using the same equipment and dataset for training and testing to ensure comparability. The experimental results are shown in Table 2.

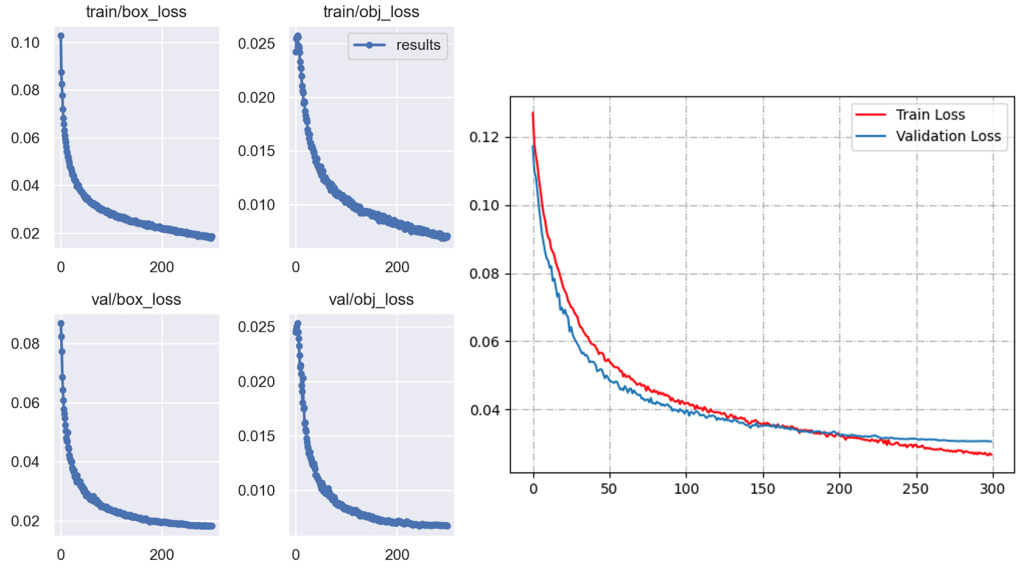

Figure 14. Loss variation curve.

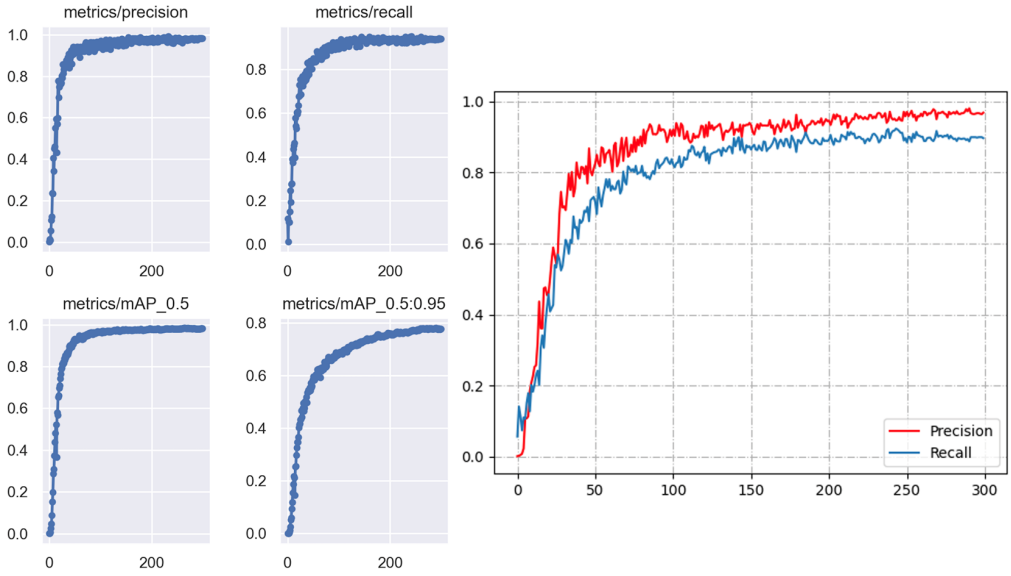

Figure 15. Metric variation curve with number of iterations.

Table 2. Ablation test results.

Model	P (%)	R (%)	FPS	mAP@0.5 (%)	Model Size (M)
YOLOv8s	90.1	88.7	153.5	90.2	21.5
YOLOv8s + GhostNetv2	87.6	85.8	186.3	87.9	16.9
YOLOv8s + GhostNetv2 + CA_H	92.3	89.9	179.8	91.8	17.3

Discussion

The main reasons for improving the model to enhance the detection performance are as follows:

- Lightweight improvement: In this paper, we use the Ghost bottleneck structure instead of the original structure on YOLOv8s to form the main part of the backbone, and we replace the standard convolution module with larger parameter counts in the neck and head with the Ghost bottleneck structure and depth-separable convolution, to generate the new lightweight network model YOLOv8s-GhostNetv2, which has

a 21.1% compression in the model size with a 3% reduction in recall and a 21.5% acceleration in computation. Precision and mAP@0.5 are only reduced by 2.8% and 2.5%, respectively, and recall is reduced by 3.3%, the model size is compressed by 21.3%, FPS is improved from 153.5 to 186.3, and computation is accelerated by 21.4%. It is proven that the improved network architecture based on the Ghost module proposed in this paper strongly helps to reduce the number of parameters and the complexity of the model, and the cost of losing precision and recall is exchanged for the speedup. In summary, YOLOv8s-GhostNetv2 maintains relatively efficient detection performance while having a lightweight framework.

- Introducing the attention mechanism: the lightweight improvement reduces the size of the algorithm model to a large extent, but in order to make up for the reduced detection performance caused by the lack of effective feature extraction brought about by the lightweight improvement, the CA_H channel attention is embedded in the Ghost bottleneck structure to generate the YOLOv8s-GhostNetv2-CA_H model, which is the same as the lightweight model. Compared with YOLOv8s-GhostNetv2, precision, recall, and mAP@0.5 are improved by 5.4%, 4.8%, and 4.4%, respectively, while the model size is increased by 2.4%, and the detection speed FPS is slowed down by 6.5 with a loss of 3.5%, which hardly brings too much computational consumption, which is attributed to the fact that the coordinate attention mechanism captures both the cross channel relationship and captures both orientation-aware and position-sensitive information. Compared with the YOLOv8s base model, the introduction of the CA_H attention mechanism mainly has the effect of improving the precision and recall, with the precision improved to 92.3%, recall improved to 89.9%, and mAP@0.5 improved to 91.8%. For the experiments in this paper, the addition of the attention mechanism can filter out the feature regions of important value from a large amount of irrelevant information, helping the model to process the information efficiently and thus obtaining performance gains without losing too much computational performance.

4.5. Comparative Experiments with Classical Algorithms

In order to objectively evaluate the detection effect of the improved model and verify the improvement of the improved algorithm in terms of accuracy and detection speed, the experiments compare the classical object detection algorithms Faster R-CNN, YOLOv5s, and YOLOv8s and the improved algorithms models YOLOv8s-GhostNetv2 and YOLOv8s-GhostNetv2-CA_H, and the results of the comparison tests are shown in Table 3.

Table 3. Comparative test results.

Model	P (%)	R (%)	FPS	mAP@0.5 (%)	Model Size (M)
Faster R-CNN	81.9	85.1	8	80.7	121.4
YOLOv5s	89.3	89.1	137.7	88.9	14.5
YOLOv8s	90.1	88.7	153.5	90.2	21.5
YOLOv8s-GhostNetv2	87.6	85.8	186.3	87.9	16.9
YOLOv8s-GhostNetv2-CA_H	92.3	89.9	179.8	91.8	17.3

Comparing the experimental results of different algorithm models in the table, the following can be seen:

- When the YOLOv8s base model is compared with the classical object detection algorithm Faster R-CNN, only recall is slightly lower than that of the YOLOv5s model by 0.4%, and the rest of the various aspects of the performance are achieved comprehensively beyond that.
- Compared with the YOLOv8s base model, the YOLOv8s-GhostNetv2-CA_H model proposed in this paper reduces the model size by 19.5%, improves the precision by 2.4% to 92.3%, the recall by 1.4%, mAP@0.5 by 1.8%, and FPS by 17.1%.

In addition, in order to better verify the feasibility of the improved model and evaluate it more intuitively, this paper selects the detection models YOLOv8s-GhostNetv2-CA_H and YOLOv8s before and after the improvement and conducts a comparison test for the special samples in the jujube tree trunk dataset, and the effect of the comparison test is shown in Figure 16.

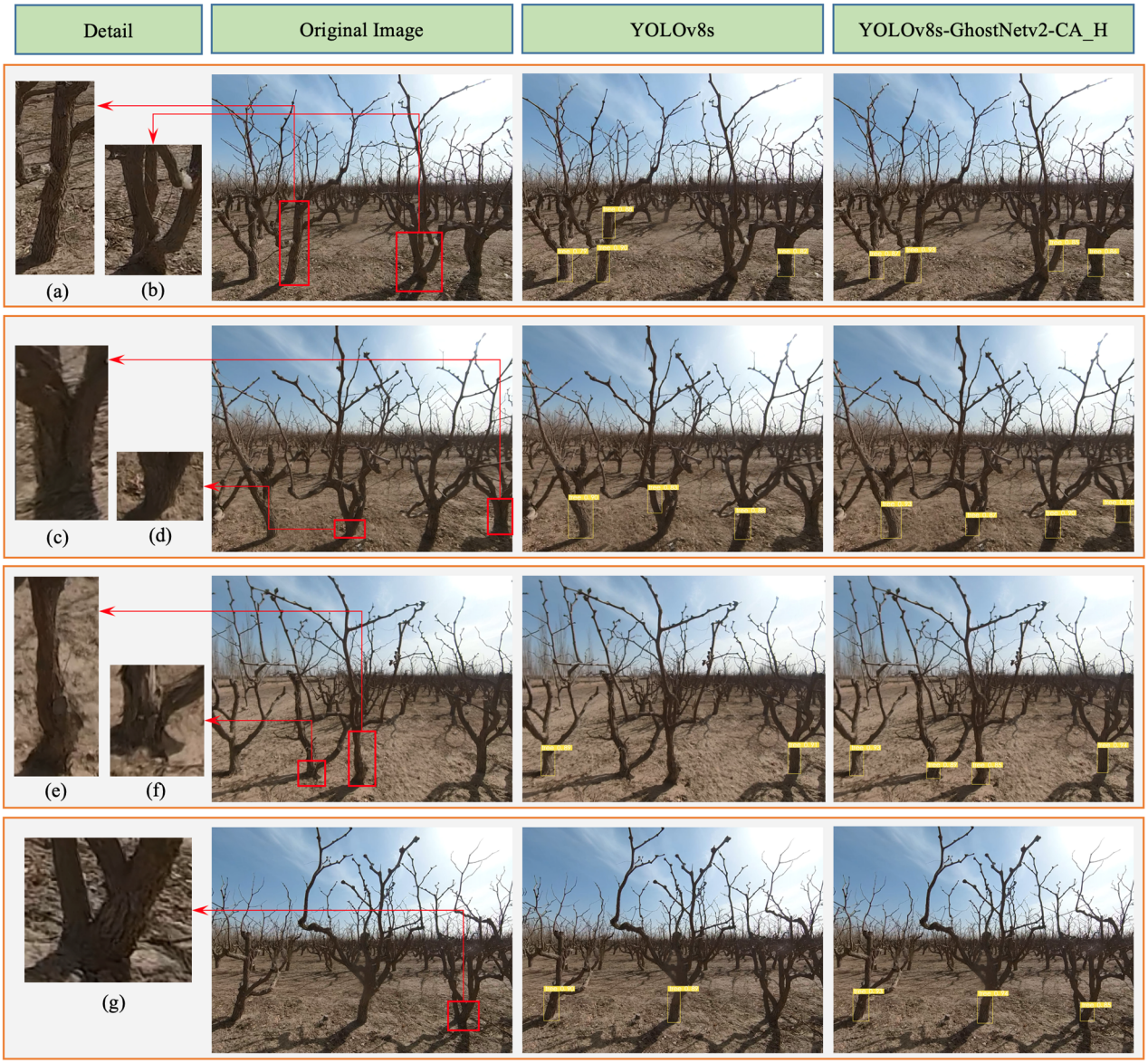

Figure 16. Comparison test effect ((**a**–**g**) for special sample cases).

Among the special sample cases are the following:

Figure 16a is a strong interference object with confusing features of the upper end branches of the trunk. The base model mistakenly detects the upper end branches of the trunk, and there is a problem of repeated detection, which can be correctly detected by the improved model.

Figure 16b is the detection of a dense occlusion object, where the trunk part has multiple trunks together, forming a partial occlusion between each other. The base model shows missed detection, while the improved model can detect the occluded object.

Figure 16c is a special tree type of jujube tree, and the trunk part is a single main root but has multiple trunk textures, which are highly interfering; the base model mistakenly detects the branch above, and the improved model with strong generalization can detect it.

Figure 16d,f are a small object detection where the trunk part is excessively buried in the soil resulting in a small object. The base model suffers from false and missed detections, while the improved model still detects and localizes accurately for small objects.

Figure 16e is a special tree type of jujube tree; the very fine texture of the trunk region is not clear enough and the lack of generalization of the base model leads to missed detection, while the improved model successfully detects it.

Figure 16g is a cross dense object with intersecting tree trunks. The base model has missed detections, while the improved model can detect the object but with average accuracy.

Discussion

Currently, the mainstream target recognition algorithms mainly contain Faster R-CNN, a representative of two-stage recognition, and the YOLO series, a representative of single-stage recognition, among which YOLOv5 is the most widely used. From the detection result comparison graph in Figure 16 and the detection accuracy shown in Table 3, it can be concluded that the improved model YOLOv8s-GhostNetv2-CA_H shows better performance compared with the mainstream target recognition model. The basic model performs poorly in complex and diverse real-world detection scenarios and misses and misdetects in the presence of strong interference, small targets, and cross dense targets, while the improved model has good robustness, generalization ability, and higher localization accuracy. The specific reasons are as follows:

- Compared to Faster R-CNN, which first uses a region proposal network (RPN) to generate candidate object regions and then performs classification and bounding box regression for each region, the YOLO series predicts the bounding box and category probabilities directly in a single neural network, and this one-step approach is more effective in real-world application scenarios with large amounts of jujube tree garden data because it reduces the steps in the inference process and the computational complexity.
- In addition, YOLO employs more advanced feature fusion mechanisms, such as cross-scale feature fusion, which can help the model better capture trunk targets of different sizes. In contrast, although Faster R-CNN can also handle multi-scale inputs, its feature fusion ability is weak, and its recognition effect is poor when facing the influence of tree branches with more disturbances.

5. Conclusions and Outlook

This study addresses two major difficulties encountered in the actual jujube garden detection process in a natural orchard environment: an excessive amount of collected redundant data and low data quality. Using data mining (key frame extraction, image processing), data enhancement, machine vision (lightweighting, attention mechanism improvement), and other techniques can propose a solution. A target recognition model of jujube tree trunks with good performance in a complex environment is constructed to realize the fast recognition of each tree in a large-scale jujube tree garden. And a combination with the key frame extraction algorithm to extract key frames containing the overall picture of jujube trees from redundant video data provides the basis for later jujube tree garden digitization to be carried out.

In our future work, we will focus on utilizing this improved object detection model in conjunction with key frame detection algorithms to realize the deployment of next-generation AI IoT systems in embedded devices. For example, this may be achieved by automatically extracting multi-view views of jujube trees and constructing spatial 3D models of jujube trees based on them [40], as well as pioneering applications for more efficient and diverse agricultural information data collection in the context of smart agriculture.

Author Contributions: Conceptualization, S.L. and N.W.; methodology, S.L. and N.W.; investigation, S.L. and L.D.; data curation, S.L.; writing—original draft preparation, S.L. and N.W.; writing—review and editing, S.L., J.L. and L.D.; funding acquisition, J.L. and L.D. 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 (52165037), and the Corps Regional Innovation Guidance Program (2021BB003).

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

Acknowledgments: We would also like to thank all reviewers for their valuable comments.

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

References

1. Liu, J.; Xiang, J.; Jin, Y.; Liu, R.; Yan, J.; Wang, L. Boost precision agriculture with unmanned aerial vehicle remote sensing and edge intelligence: A survey. *Remote Sens.* **2021**, *13*, 4387. [CrossRef]
2. Nie, J.; Jiang, J.; Li, Y.; Wang, H.; Ercisli, S.; Lv, L. Data and domain knowledge dual-driven artificial intelligence: Survey, applications, and challenges. *Expert Syst.* **2023**, e13425. [CrossRef]
3. Cheng, Z.; Cheng, Y.; Li, M.; Dong, X.; Gong, S.; Min, X. Detection of cherry tree crown based on improved LA-dpv3+ algorithm. *Forests* **2023**, *14*, 2404. [CrossRef]
4. Nie, J.; Wang, Y.; Li, Y.; Chao, X. Artificial intelligence and digital twins in sustainable agriculture and forestry: A survey. *Turk. J. Agric. For.* **2022**, *46*, 642–661. [CrossRef]
5. Donmez, C.; Villi, O.; Berberoglu, S.; Cilek, A. Computer vision-based citrus tree detection in a cultivated environment using UAV imagery. *Comput. Electron. Agric.* **2021**, *187*, 106273. [CrossRef]
6. Zhang, R.; Li, P.; Zhong, S.; Wei, H. An integrated accounting system of quantity, quality and value for assessing cultivated land resource assets: A case study in Xinjiang, China. *Glob. Ecol. Conserv.* **2022**, *36*, e02115. [CrossRef]
7. Li, Y.; Ercisli, S. Data-efficient crop pest detection based on KNN distance entropy. *Sustain. Comput. Inform. Syst.* **2023**, *38*, 100860.
8. Yang, Y.; Li, Y.; Yang, J.; Wen, J. Dissimilarity-based active learning for embedded weed identification. *Turk. J. Agric. For.* **2022**, *46*, 390–401. [CrossRef]
9. Ye, G.; Liu, M.; Wu, M. Double image encryption algorithm based on compressive sensing and elliptic curve. *Alex. Eng. J.* **2022**, *61*, 6785–6795. [CrossRef]
10. Li, Y.; Yang, J.; Zhang, Z.; Wen, J.; Kumar, P. Healthcare data quality assessment for cybersecurity intelligence. *IEEE Trans. Ind. Inform.* **2022**, *19*, 841–848. [CrossRef]
11. Xu, S.; Pan, B.; Zhang, J.; Zhang, X. Accurate and Serialized Dense Point Cloud Reconstruction for Aerial Video Sequences. *Remote Sens.* **2023**, *15*, 1625. [CrossRef]
12. Ahmed, M.; Ramzan, M.; Khan, H.U.; Iqbal, S.; Khan, M.A.; Choi, J.-I.; Nam, Y.; Kadry, S. *Real-Time Violent Action Recognition Using Key Frames Extraction and Deep Learning*; Tech Science Press: Henderson, NV, USA, 2021.
13. Wang, X.; Wang, A.; Yi, J.; Song, Y.; Chehri, A. Small Object Detection Based on Deep Learning for Remote Sensing: A Comprehensive Review. *Remote Sens.* **2023**, *15*, 3265. [CrossRef]
14. Ciarfuglia, A.T.; Motoi, M.I.; Saraceni, L.; Fawakherji, M.; Sanfeliu, A.; Nardi, D. Weakly and semi-supervised detection, segmentation and tracking of table grapes with limited and noisy data. *Comput. Electron. Agric.* **2023**, *205*, 107624. [CrossRef]
15. Ouhami, M.; Hafiane, A.; Es-Saady, Y.; El Hajji, M.; Canals, R. Computer vision, IoT and data fusion for crop disease detection using machine learning: A survey and ongoing research. *Remote Sens.* **2021**, *13*, 2486. [CrossRef]
16. Ling, S.; Wang, N.; Li, J.; Ding, L. Optimization of VAE-CGAN structure for missing time-series data complementation of UAV jujube garden aerial surveys. *Turk. J. Agric. For.* **2023**, *47*, 746–760. [CrossRef]
17. Chao, X.; Li, Y. Semisupervised few-shot remote sensing image classification based on KNN distance entropy. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8798–8805. [CrossRef]
18. Maity, M.; Banerjee, S.; Chaudhuri, S.S. Faster r-cnn and yolo based vehicle detection: A survey. In Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 8–10 April 2021; pp. 1442–1447.
19. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
20. Junos, M.H.; Khairuddin, A.S.M.; Dahari, M. Automated object detection on aerial images for limited capacity embedded device using a lightweight CNN model. *Alex. Eng. J.* **2022**, *61*, 6023–6041. [CrossRef]
21. Li, Y.; Chao, X.; Ercisli, S. Disturbed-entropy: A simple data quality assessment approach. *ICT Express* **2022**, *8*, 309–312. [CrossRef]
22. Osco, P.L.; de Arruda, S.D.M.; Gonçalves, N.D.; Dias, A.; Batistoti, J.; de Souza, M.; Gomes, F.D.G.; Ramos, A.P.M.; de Castro Jorge, L.A.; Liesenberg, W.; 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]

23. Li, Y.; Ercisli, S. Explainable human-in-the-loop healthcare image information quality assessment and selection. *CAAI Trans. Intell. Technol.* **2023**. [CrossRef]
24. Zhang, Y.; Yuan, B.; Zhang, J.; Li, Z.; Pang, C.; Dong, C. Lightweight PM-YOLO Network Model for Moving Object detection on the Distribution Network Side. In Proceedings of the 2022 2nd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS), Shenyang, China, 25–27 February 2022; pp. 508–516.
25. Li, Y.; Chao, X. Distance-entropy: An effective indicator for selecting informative data. *Front. Plant Sci.* **2022**, *12*, 818895. [CrossRef] [PubMed]
26. Yang, K.; Chang, S.; Tian, Z.; Gao, C.; Du, Y.; Zhang, X.; Liu, K.; Meng, J.; Xue, L. Automatic polyp detection and segmentation using shuffle efficient channel attention network. *Alex. Eng. J.* **2022**, *61*, 917–926. [CrossRef]
27. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Detection, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
28. Conroy, L.T.; Moore, B.J. Resolution invariant surfaces for panoramic vision systems. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 1, pp. 392–397.
29. Wan, S.; Ding, S.; Chen, C. Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles. *Pattern Detect.* **2022**, *121*, 108146. [CrossRef]
30. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-adaptive YOLO for object detection in adverse weather conditions. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 1792–1800.
31. Reza, M.A. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **2004**, *38*, 35–44. [CrossRef]
32. Ravikumar, M.; Rachana, G.P.; Shivaprasad, J.B.; Guru, S.D. Enhancement of mammogram images using CLAHE and bilateral filter approaches. In *Cybernetics, Cognition and Machine Learning Applications: Proceedings of ICCCMLA*; Springer: Singapore, 2021; pp. 261–271.
33. Terven, J.; Cordova-Esparza, D. A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond. *arXiv* **2023**, arXiv:2304.00501.
34. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Detection, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
35. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetv2: Enhance cheap operation with long-range attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9969–9982.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, N.A.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
37. Real, E.; Aggarwal, A.; Huang, Y.; Le, V.Q. Regularized evolution for image classifier architecture search. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4780–4789.
38. Gu, R.; Wang, G.; Song, T.; Huang, R.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; Zhang, S. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging* **2020**, *40*, 699–711. [CrossRef] [PubMed]
39. Zimmerman, B.J.; Pizer, M.S.; Staab, V.E.; Perry, R.J.; McCartney, W.; Brenton, C.B. An evaluation of the effectiveness of adaptive histogram equalization for contrast enhancement. *IEEE Trans. Med. Imaging* **1988**, *7*, 304–312. [CrossRef]
40. Ling, S.; Li, J.; Ding, L.; Wang, N. Multi-View Jujube Tree Trunks Stereo Reconstruction Based on UAV Remote Sensing Imaging Acquisition System. *Appl. Sci.* **2024**, *14*, 1364. [CrossRef]