agriculture

MDPI

Article

An Apple Detection and Localization Method for Automated Harvesting under Adverse Light Conditions

Guoyu Zhang [1], Ye Tian [1], Wenhan Yin [2] and Change Zheng [1,*]

[1] School of Technology, Beijing Forestry University, Beijing 100083, China; zgy914518160@bjfu.edu.cn (G.Z.); tytoemail@bjfu.edu.cn (Y.T.)
[2] School of Information Science and Technology, North China University of Technology, Beijing 100144, China; yinwenhan@mail.ncut.edu.cn
* Correspondence: zhengchange@bjfu.edu.cn

Abstract: The use of automation technology in agriculture has become particularly important as global agriculture is challenged by labor shortages and efficiency gains. The automated process for harvesting apples, an important agricultural product, relies on efficient and accurate detection and localization technology to ensure the quality and quantity of production. Adverse lighting conditions can significantly reduce the accuracy of fruit detection and localization in automated apple harvesting. Based on deep-learning techniques, this study aims to develop an accurate fruit detection and localization method under adverse light conditions. This paper explores the LE-YOLO model for accurate and robust apple detection and localization. The traditional YOLOv5 network was enhanced by adding an image enhancement module and an attention mechanism. Additionally, the loss function was improved to enhance detection performance. Secondly, the enhanced network was integrated with a binocular camera to achieve precise apple localization even under adverse lighting conditions. This was accomplished by calculating the 3D coordinates of feature points using the binocular localization principle. Finally, detection and localization experiments were conducted on the established dataset of apples under adverse lighting conditions. The experimental results indicate that LE-YOLO achieves higher accuracy in detection and localization compared to other target detection models. This demonstrates that LE-YOLO is more competitive in apple detection and localization under adverse light conditions. Compared to traditional manual and general automated harvesting, our method enables automated work under various adverse light conditions, significantly improving harvesting efficiency, reducing labor costs, and providing a feasible solution for automation in the field of apple harvesting.

Keywords: apple harvesting; adverse light; detection; localization

Citation: Zhang, G.; Tian, Y.; Yin, W.; Zheng, C. An Apple Detection and Localization Method for Automated Harvesting under Adverse Light Conditions. *Agriculture* **2024**, *14*, 485. https://doi.org/10.3390/agriculture14030485

Academic Editor: John M. Fielke

Received: 20 February 2024
Revised: 11 March 2024
Accepted: 12 March 2024
Published: 16 March 2024

1. Introduction

Agriculture is a vital part of the global economy and human life, and it faces unprecedented challenges due to population growth and the effects of climate change. Against this backdrop, improving the efficiency and sustainability of agricultural production has become particularly urgent. Automation technology, a tool to address these challenges, is a significant component of agricultural modernization [1–3].

As one of the world's principal fruits, apples are prominent agricultural products. However, traditional apple harvesting methods rely heavily on human labor and are often susceptible to seasonal labor shortages and rising labor costs [4]. To address these challenges, automated apple harvesting technology has emerged.

Apple harvesting robots have emerged as a potential solution to automate fruit harvesting, replacing manual labor [5–9]. These robots comprise various components, such as a vision detection system, mechanical structure, and control system. The vision detection system, in particular, is crucial for enabling the robot to perceive its surroundings and has

been extensively studied in recent years. The rapid and accurate detection of target fruits is a critical technology for apple harvesting robots, as it has significant practical implications for improving agricultural productivity and the efficiency of harvesting robots [10]. However, several challenges need to be addressed by harvesting robots in practice. One of the main challenges is the fast and accurate localization and detection of harvest objects in complex natural environments by vision detection systems. It is more challenging in an orchard with complex and variable lighting, dense distribution of fruit targets, and widespread mutual occlusion. Research on detection and localization algorithms with robust target detection capabilities in complex natural environments is necessary.

Deep-learning techniques have been widely used in visual detection [11–15]. In the field of target detection, mainstream target detection algorithms are usually categorized into single-stage target detection, e.g., the YOLO [16] (you only look once) series and SSD [17] (single shot multibox detector), and two-stage target detection, e.g., RCNN (Region CNN) [18], Fast RCNN [19], and Faster RCNN [20]. These algorithms show excellent results under conditions with proper illumination. However, their performance tends to degrade in the real world due to varying lighting conditions and noise.

In natural environments, lighting conditions are affected by the weather and time of day. Images captured under adverse light conditions, such as cloudy days and evening, often exhibit low contrast and low brightness [21]. These low-quality images can negatively impact the visual task of apple harvesting, leading to subpar performance in apple detection and localization, consequently affecting the success rate of the harvesting task. Figure 1 shows an example of apple detection under adverse light conditions.

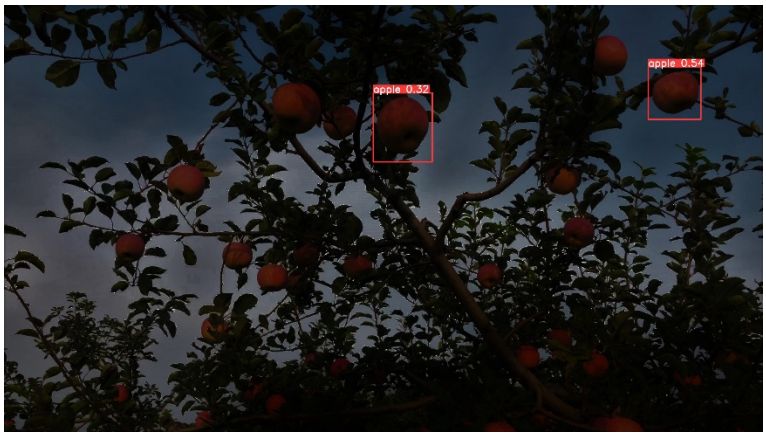

Figure 1. Apple detection in adverse light.

One way to address the challenges of enhancing brightness and contrast with low-light images is to utilize existing methods. For example, Retinex-Net, proposed by Wei et al., can decompose the input low-light image into reflectance and illuminance for brightness enhancement [22]. However, this method has limitations in terms of preserving color information. Another method proposed by Guo et al., called Zero-DCE, formulates light enhancement as an image-specific curve estimation task [23]. While it addresses some limitations, it may still struggle with noise reduction. Lv et al. proposed MBLLEN, a multi-branch low-light enhancement network that enhances images by merging different branches [24]. However, this method may not effectively preserve details and textures in the enhanced images. Although there are some methods to address these limitations, such as KinD (Kindling the Darkness), proposed by Zhang et al. [25], TSN-CA (Two-Stage Network with Channel Attention), proposed by Wei et al. [26], and RDGAN (Retinex Decomposition-Based Generative Adversarial Network), proposed by Wang et al. [27], they often suffer from a complex structure and may be time-consuming, limiting their practical application. Therefore, further research and development are necessary to overcome these

limitations and create more efficient and effective methods for enhancing brightness and contrast in low-light images.

We propose a fast, lightweight, and light-enhanced object detection method called LE-YOLO. Our method consists of a lightweight image enhancement module that adaptively adjusts the brightness and contrast of the input image while recovering color information. The module can be seamlessly embedded into the improved YOLOv5 network and combined with a binocular camera to achieve apple detection and localization under adverse light conditions.

The highlights of this study are as follows: (1) introducing a lightweight image enhancement module into the YOLOv5 network, which enables effective detection in adverse light conditions, (2) adding attention mechanisms to YOLOv5 to improve network performance, (3) and combining binocular camera ranging for the spatial localization of apples.

Our research provides significant innovations for automated apple harvesting by combining deep learning with target-detection techniques. Not only is it of practical value for improving the modernization of the apple industry, but it also provides valuable insights into the application of automation technology in agriculture. This paper describes our methodology, experimental design, and results. Furthermore, we discuss potential directions for future research. Through this research, we hope to contribute to the digitization and intelligence of agricultural production and to positively impact the global agricultural sector by addressing its challenges.

2. Materials and Methods

2.1. System Framework

The proposed system framework (Figure 2) for adverse-light object detection comprises three main modules: image enhancement, fruit detection, and the fruit localization module. The image enhancement module is based on the SARN network [28], which can enhance images. On the other hand, the fruit detection module utilizes YOLOv5, a state-of-the-art object detection algorithm, and implements an attention mechanism to improve its performance. Furthermore, the fruit localization module achieves the spatial localization of the fruit-picking point by adopting a binocular camera. This module utilizes the center of the bounding box generated by the fruit detection module to determine the fruit-picking point. By employing stereo-matching techniques, the module then calculates the parallax, which is subsequently used to determine the spatial location of the fruit-picking point by applying triangulation principles.

2.2. Image Enhancement Network

Most current image enhancement networks have high time costs due to their complex structure, which limits their practicality. A stacked attention residual network (SARN) [28] is a very lightweight and fast pipeline for image enhancement with simultaneous contrast improvement, noise removal, detail preservation, and color information recovery. A channel attention module (SE Module) is inserted into the residual block and its shortcuts to obtain an attention residual block (ARB). The ARBs are then stacked as the backbone of the SARN. Before the low-quality image is fed into the backbone, the network first extracts its shallow features, which contain the color information of the original image. The network is connected by global jumps, and the extracted shallow features are directly fused with the high-level output features after the brightness is enhanced by the backbone, thus effectively preserving the color information. The SARN network is lightweight, fast, and practical, making it suitable as a low-quality image enhancement module. The network architecture of SARN is shown in Figure 3.

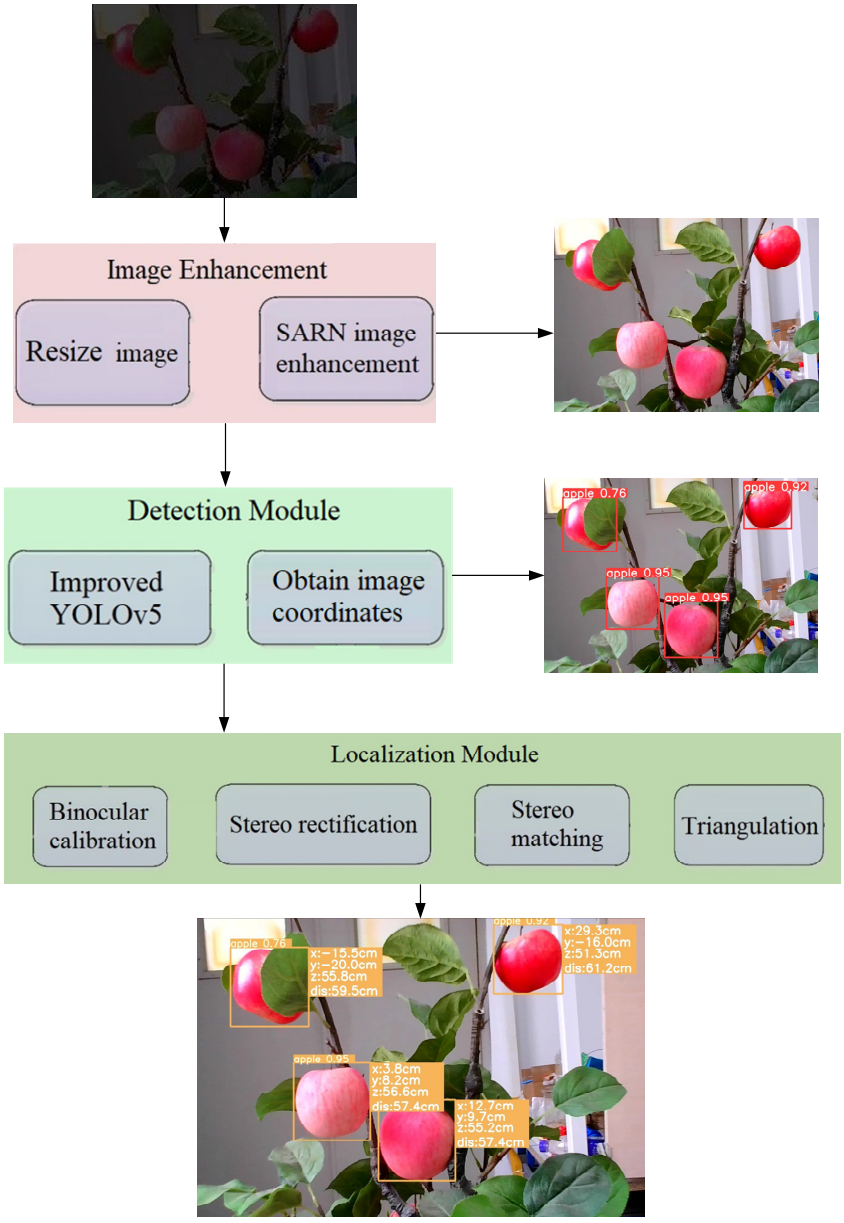

Figure 2. System framework.

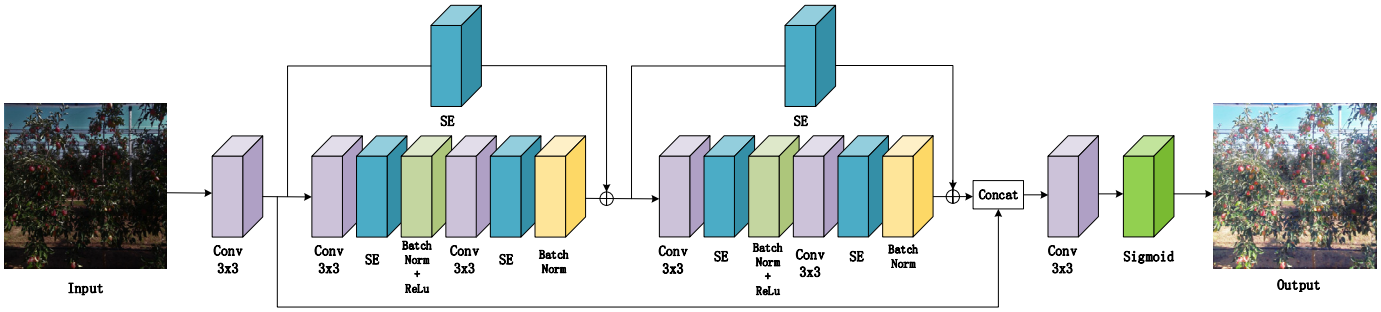

Figure 3. SARN network architecture.

2.3. YOLOv5 Model

The YOLO family, from YOLOv1 to YOLOv5 [29], has gained significant popularity in object detection due to its fast and efficient performance. Compared to the RCNN family, YOLO dramatically improves the model runtime speed while maintaining detection accuracy, making it suitable for real-time performance. YOLOv5 is a deep-learning algorithm for target detection that can quickly and accurately identify target objects in images or videos. The basic principle of this algorithm is to achieve target detection by segmenting the image into grids of different sizes and making predictions for each grid. The algorithmic flow of YOLOv5 can be divided into three main steps: input processing, feature extraction and target prediction. First, the input image is resized to the size required by the model and normalized. Next, after a series of convolution and pooling operations, the features in the image are extracted. Finally, the categories and bounding boxes of the target objects in each grid are predicted by classifying and regressing the feature maps. In the input processing stage, YOLOv5 splits the image into grids of different sizes, each of which is called an anchor point. Each anchor point is responsible for detecting one target object, while anchors of different sizes are responsible for detecting targets of different sizes. This multi-scale design allows YOLOv5 to detect target objects of different sizes, which improves the accuracy of detection. In the feature extraction phase, YOLOv5 uses a backbone network called CSPDarknet53 to extract features from the image. CSP Darknet53 is a lightweight network structure that improves the representation of features through the use of residual blocks and hopping connections. This network structure can efficiently extract semantic information from images and keep the computational complexity low. In the target prediction phase, the head structure of YOLOv5 is used to achieve the target prediction. The header consists of a series of convolutional and fully connected layers for the classification and regression of the feature map. Specifically, the classification layer is used to predict the class of the target object, while the regression layer is used to predict the bounding box of the target object. Since the size of the latest generation weights file is only 28 MB, and it is suitable for the initial model, YOLOv5 was chosen as the target of algorithm improvement in this study.

YOLOv5 consists of the input side, backbone, neck, and head [30]. The backbone mainly includes the focus and CSP modules as the feature extraction; the focus module not only realizes down sampling but also reduces the amount of computation, while the CSP module serves as the core to enhance feature extraction. A combination of the FPN and the PAN are used in the neck, and the head output structure includes the probability score of the target and the bounding box. The structure of the whole YOLOv5 is shown in Figure 4.

2.4. Improvements to the YOLOv5

2.4.1. Insertion Attention Mechanisms

By employing the attention mechanism, the model can identify and emphasize a small subset of crucial information while disregarding the less significant details from a large amount of information. To address the challenges posed by the intricate background in realistic orchard environments, the attention mechanism enables the model to effectively concentrate on the relevant information of the target region, even when there are numerous non-target regions within the image. YOLOV5 can be enhanced by incorporating an attention mechanism. Specifically, this paper utilizes squeeze-and-excitation networks (SE net) [31] to implement the attention mechanism. With its inherent learning capabilities, the SE net automatically determines the corresponding importance of each feature channel. The importance level obtained is then used to amplify relevant features while suppressing less relevant ones. The SE structure is shown in Figure 5.

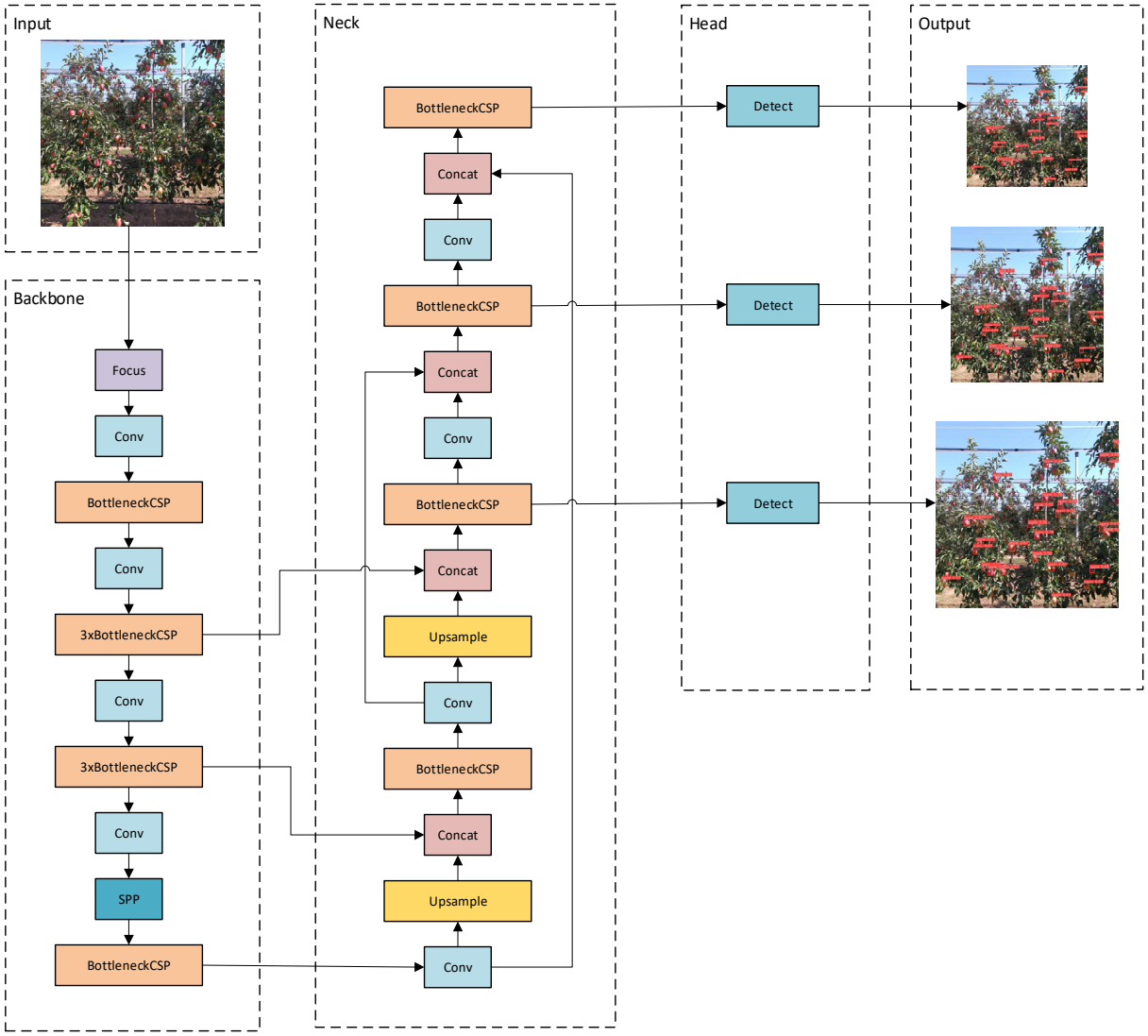

Figure 4. YOLOv5 network architecture.

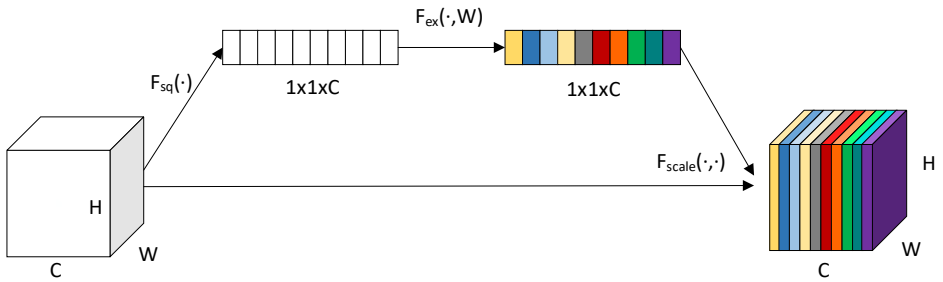

Figure 5. The SE structure.

The structure of the improved YOLOV5 is shown in Figure 6.

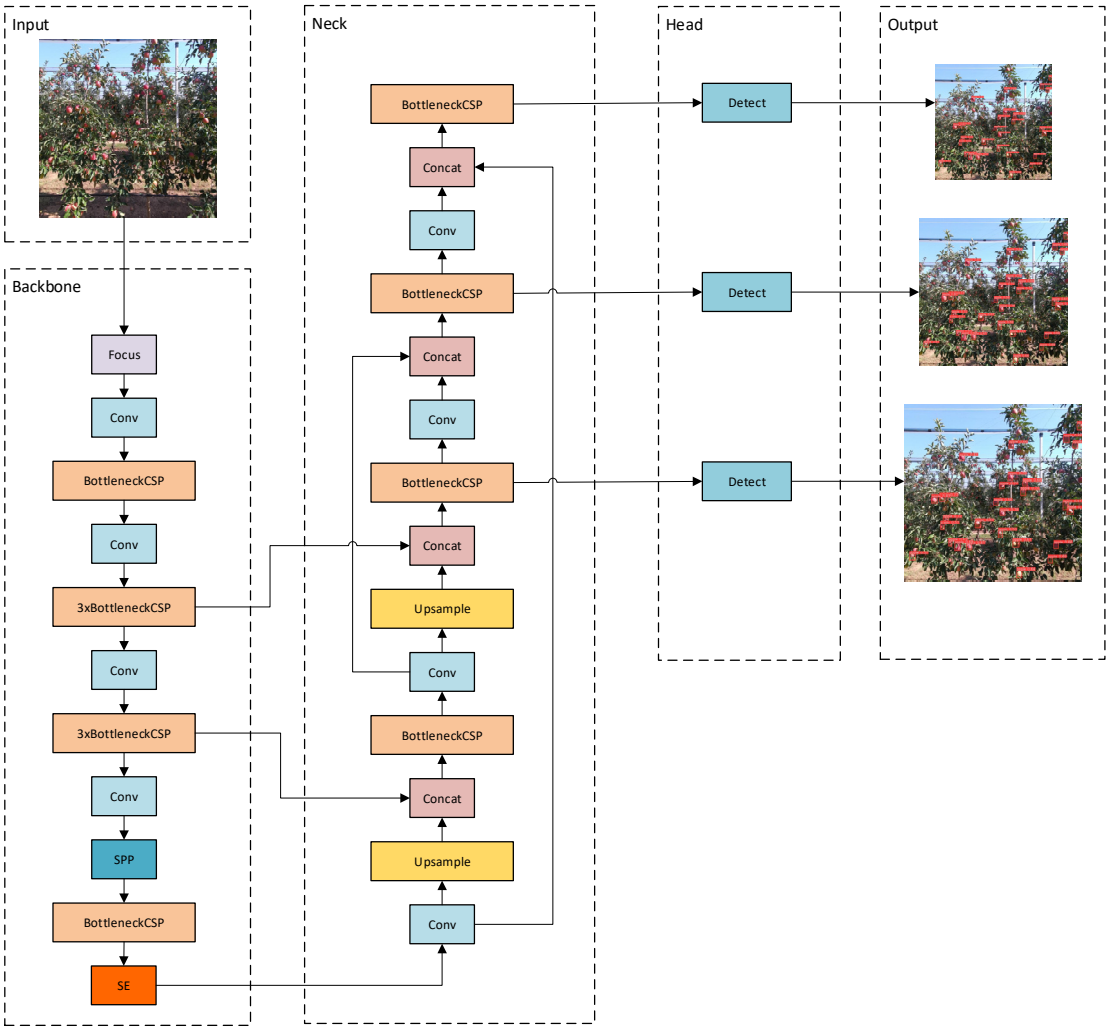

Figure 6. Improved YOLOv5 network architecture.

2.4.2. Improvement of the Loss Function

YOLOv5 incorporates the CIoU (complete intersection over union) loss function, which includes the overlap area, centroid distance, and aspect ratio of the bounding box regression. However, the aspect ratio difference represented by "v" in its formula does not accurately reflect the difference between the width and height of the box and its confidence levels. The "v" used by CIoU is the relative proportion of the width and height, not the value of the width and height. As long as the width and height of the predicted box correspond to a certain linear ratio with the width and height of the target box, then the added relative ratio penalty in CIoU will no longer work. Consequently, this sometimes prevents the model from effectively optimizing similarity. To address this issue, EIoU (enhanced intersection over union) loss is employed.

The penalty term of EIoU builds upon the penalty term of CIoU by dividing the influence factor of the aspect ratio to calculate the length and width of the target box and the predicted box, respectively. The loss function consists of three components: overlap loss, center distance loss, and width–height loss. The first two components follow the approach used in CIoU, while the width–height loss directly minimizes the difference in width and height between the target box and the predicted box. They facilitate faster convergence. The formula for the penalty term is as follows:

$$L_{EIOU} = L_{IOU} + L_{dis} + L_{asp} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} \quad (1)$$

where b and b^{gt} denote the central points of the predicted box and the target box respectively. $\rho(\cdot) = \left|\left|b - b^{gt}\right|\right|_2$ indicates the Euclidean distance. w and h are the width and height of the predicted box. w^{gt} and h^{gt} are the width and height of the target box. w^c and h^c are the width and height of the smallest enclosing box covering the two boxes.

2.5. Picking Point Localization Based on Binocular Stereo Vision

Stereo matching is used to obtain parallax information. The images used in this study were preprocessed by median filtering before stereo matching. The purpose of stereo matching is to recognize the same picking point of apples in the images taken by the left and right cameras. After stereo matching, the parallax is obtained. After calculating the parallax, a triangulation model can be constructed as shown in Figure 7, where P is the apple picking localization point with coordinates relative to the left camera coordinate system (X, Y, Z); pl is the projection point of point P on the left imaging plane, and pr is the projection point of point P on the right imaging plane; f is the focal length of the camera; and b is the baseline of the binocular vision system.

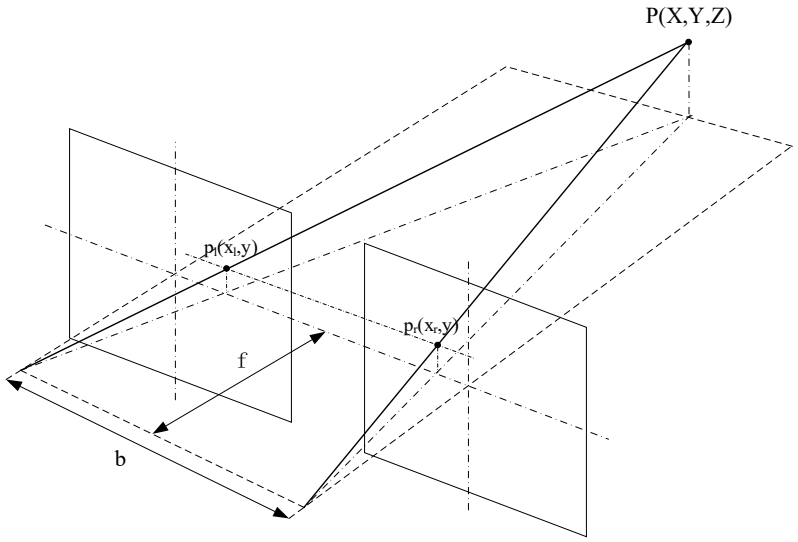

Figure 7. Triangulation model.

2.6. Implementation Details

The implementation details of the system include training the model on a desktop computer with a GeForce GTX 1080ti GPU and using the PyTorch (1.6.0) [32] deep-learning framework. Equipment and software systems used to capture images include the following: binocular camera (resolution: 1920 × 1080); laptop (i5-8300 CPU; 8G RAM; GTX1050Ti GPU); video cap.

2.7. Datasets

To obtain a rich and diverse dataset, we sampled data from an apple orchard maintained in the Changping District, Beijing, China. The orchard spans approximately 30 acres and consists of two varieties of apple trees: Fuji and Golden Crown. We used a digital camera to photograph various areas of the orchard. The data collection process took place from September to October 2022, allowing us to capture the seasonal changes and variations in the apple trees and their surroundings. This real-world data provided valuable insights into the features and transformations in an apple orchard environment to improve the model's performance and ensure its robustness in real-world applications. Additionally, annotating the image data collected in this authentic environment enhanced the reliability of the annotations and minimized the risk of overfitting and over-adaptation to specific scenes. Overall, realistic apple orchard images provided a valuable resource for our research. Images were collected under various weather conditions (e.g., sunny, cloudy, and

other weather conditions), at different time intervals (i.e., morning, noon, afternoon, and evening), from multiple angles (including light, sidelight, and backlight), and at different shooting distances, resulting in a dataset of 2043 images, as shown in Figure 8. This dataset represents the factors of weather, time of image collection, and angle of light uncertainty, resulting in images under different illumination conditions. We randomly selected 200 normal-light images and 200 adverse-light images as the test set and used the remaining 1643 images as the training set.

(**a**) (**b**)

Figure 8. (**a**) normal-light image, (**b**) severe-light image.

Given the insufficient training datasets, data enhancement allowed us to expand the dataset to increase the diversity of the data as well as improve the robustness of the model. The image enhancement methods for the training set included image brightness enhancement and reduction, horizontal mirroring, vertical mirroring, and multi-angle rotation. Data enhancement made it possible to extend the training set; the final training set included 9858 images.

2.8. Network Training

In this study, the improved YOLOv5s network was trained by stochastic gradient descent (SGD) in an end-to-end way. The batch size was set to 4, and each time, regularization was achieved using the BN layer to update the weight of model. The momentum factor (momentum) was set to 0.937, and the weight decay rate (decay) was set to 0.0005. The initial vector and IoU (intersection over union) threshold were all set to 0.01, and the enhancement coefficient of hue (H), saturation (S) and lightness (V) were set to 0.015, 0.7, and 0.4, respectively. The epochs were set to 100.

2.9. Evaluation Metrics

In this study, the precision (*P*), recall (*R*) and mean average precision (*mAP*) were used to assess the model's performance. These metrics were employed to evaluate the model's overall performance in a comprehensive manner.

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

Here, (*TP*) represents the model correctly predicting a sample as positive, i.e., the true class was indeed positive; (*FP*) represents the case in which a sample was incorrectly

predicted to be positive, but the true class was actually negative; (*FN*) means that the model incorrectly predicted a sample as negative, but the true class was actually positive.

$$AP = \int_0^1 P(r)dr \tag{4}$$

($P(r)$) represents the precision–recall curve, i.e., provides a graphical representation of the trade-off between precision (vertical axis) and recall (horizontal axis) for different thresholds in a binary classification model.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{5}$$

(*mAP*) refers to the average precision when the intersection over union (IoU) threshold is set to 0.5.

3. Results

3.1. Ablation Experiments

The ablation experiments were conducted to assess the impact of different improvement methods on the performance of the YOLOv5s model. The results (shown in Table 1) indicate that the SARN image enhancement module, SE attention mechanism module, and EIoU loss function can significantly improve model accuracy. The SARN image enhancement module significantly mitigated the influence of poor lighting and noise on input images, resulting in a 3% increase in precision (P), a 3.4% increase in recall rate (R), and a 3.2% increase in average precision (mAP) compared to the original YOLOv5s model. The SE attention mechanism enhanced the model's ability to capture contextual information surrounding pixels. This improvement led to a 0.9% increase in precision (P), a 1.5% increase in recall rate (R), and a 1.4% increase in average precision (mAP) compared to the original YOLOv5s model. Furthermore, the EIoU loss function was implemented for a more accurate measurement of the overlap between bounding boxes, thus improving the accuracy and robustness of bounding box regression. This enhancement resulted in a 0.6% increase in precision (P), a 1.1% increase in recall rate (R), and a 1% increase in average precision (mAP) compared to the original YOLOv5s model. By combining these three improvements, our modified model achieved significant performance gains. Compared to the original YOLOv5s model, our model demonstrated a 4.4% increase in precision (P), a 4.9% increase in recall rate (R), and a 5% increase in average precision (mAP). These improvements highlight the effectiveness of our approach for advancing the performance of the YOLOv5s model.

Table 1. Ablation experiments results.

SARN	SE	EIOU	P	R	mAP
×	×	×	0.852	0.802	0.853
√	×	×	0.882	0.836	0.885
×	√	×	0.861	0.817	0.867
×	×	√	0.858	0.813	0.863
√	√	√	0.897	0.851	0.903

3.2. Apple Detection Performance

Figure 9 shows the results of our method for recognizing apples under adverse lighting conditions. Figure 9a shows the input image and Figure 9b shows the detection as well as accuracy using YOLOv5s. Figure 9c shows the detection of apples by our method. By recognizing the output image, our method detected more apples and with better accuracy compared to the YOLOv5s model alone.

(**a**) (**b**) (**c**)

Figure 9. Detection results ((**a**) original image, (**b**) detection result of YOLOv5s, (**c**) detection result of our method).

As proof of the superiority of our method, we compared it with other models (YOLOv5s, RDGAN + YOLOv5s, and Zero-DCE + YOLOv5s) under different illumination conditions. Table 2 presents the results of these experiments. Our method consistently outperformed the other models in terms of mean average precision (mAP). Under normal lighting conditions, our method improved the mAP by 3.1%, 6.5%, and 7.5% compared to YOLOv5s, RDGAN + YOLOv5s, and Zero-DCE + YOLOv5s, respectively. Similarly, under overexposure and adverse lighting conditions, our method demonstrated superior performance with improvements of 6.8%, 3%, and 3.7% and 10.9%, 8.1%, and 10.4% in mAP, respectively. This suggests that our method is efficient for object detection under different lighting conditions. Furthermore, our method was faster for single-image inference compared to RDGAN and Zero-DCE, which highlights the efficiency of our approach in addition to its improved detection accuracy.

Table 2. Comparison of experimental results.

Model	Normal Light			Overexposure			Adverse Light			Inference Time (ms)
	P	R	mAP	P	R	mAP	P	R	mAP	
YOLOV5s	0.958	0.934	0.941	0.731	0.692	0.714	0.852	0.770	0.851	83
RDGAN + YOLOv5s	0.922	0.894	0.907	0.742	0.709	0.752	0.887	0.854	0.879	256
Zero-DCE + YOLOv5s	0.912	0.893	0.897	0.747	0.734	0.745	0.875	0.791	0.856	198
Ours	0.977	0.951	0.972	0.794	0.756	0.782	0.954	0.910	0.960	102

3.3. Apple Localization Performance

3.3.1. Binocular Camera Calibration

In order to effectively obtain internal and external parameter information from the left and right cameras, the stereo camera calibrator tool was utilized in MATLAB to calibrate the parameters of the binocular camera, and the size of the calibration checkerboard grid was chosen to have a side length of 24 mm and 8 × 6 corner points; the calibration effect is shown in Figure 10.

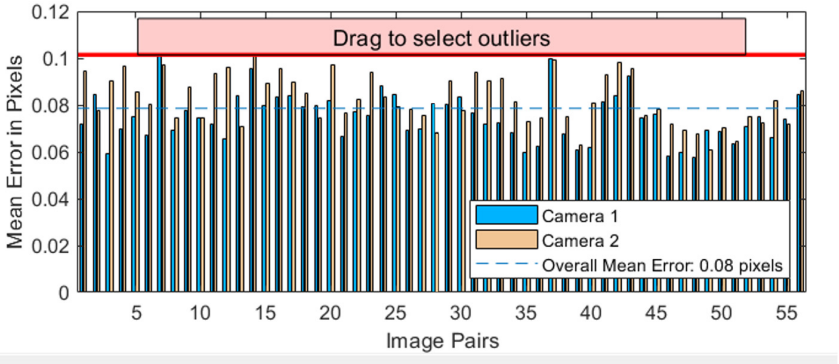

Figure 10. Calibration effect.

The obtained camera baseline d was 119.344 mm; the parameters of the left and right cameras are shown in Table 3.

Table 3. Left and right camera internal parameters.

Camera	k1	k2	k3	p1	p2	p3
Left	−0.06513102	0.30793057	0.00122804	−0.00045881	−0.41170792	−0.06513102
Right	0.032415219	0.128195917	0.00040042	0.00104828	0.06010414	−0.03241521

3.3.2. Localization Results

The localization accuracy of fruits determines the picking success rate of the picking robot. In order to assess the stability and robustness of the localization algorithm, the localization accuracy of the algorithm needed to be tested. Since it was not possible to directly measure the actual 3D coordinates of the localization point in the coordinate system of the left camera, it was not possible to calculate the errors in the X, Y and Z directions in the coordinates. Instead, the Euclidean distance error from the left camera to the localization point was used as a measure of the localization accuracy.

Table 4 shows the mean error of our method with respect to the original YOLOv5s localization under different lighting conditions. The experimental results show that our method improved the average accuracy of localization by 44% under normal lighting conditions and 73% under adverse lighting conditions relative to the localization method based on the original YOLOv5s. These results show that the localization stability of the algorithm is high, as is its robustness with respect to illumination variations, suggesting that the model is capable of meeting the application requirements of picking robots. Figure 11 shows the results of our method for the localization of apples under adverse light conditions.

Table 4. Distance measurement experiment results.

Model	Normal Light	Adverse Light
	Mean Error (cm)	Mean Error (cm)
YOLOV5s + localization	0.57	2.38
Ours	0.32	0.65

(**a**) (**b**)

Figure 11. Localization results ((**a**) an apple tree in adverse light, (**b**) apple localization result obtained using our method).

4. Discussion

Traditional manual apple picking involves significant labor costs and inefficiencies. To address this, the development of apple picking robots has become an area of interest. However, these robots have difficulty accurately detecting apples under varying light conditions, hindering their performance in natural environments. Deep-learning-based

fruit detection algorithms, such as the YOLOv5 model, offer faster and more accurate detection capabilities than traditional methods. We propose an enhanced version of the YOLOv5 model, called LE-YOLO, for apple detection and localization. The LE-YOLO model incorporates improvements in light-conditioned image enhancement using the SARN module, feature extraction using the SE attention mechanism, and robustness through the EIoU loss function. By integrating the improved YOLOv5 model with binocular camera depth ranging, the LE-YOLO model achieves accurate apple detection and localization under different lighting conditions. It is important to note that the proposed model does have limitations. Firstly, the image enhancement module may slightly reduce the inference speed of the model. In addition, the model's accuracy when detecting heavily occluded apples has not been fully validated. In future research, we plan to expand our dataset to include more images of apples in complex backgrounds, verifying the efficacy of the LE-YOLO algorithm in various environmental scenarios. We also aim to extend the application of LE-YOLO to other fruit datasets, optimizing the model for a broader range of fruit detection applications.

5. Conclusions

We propose a novel apple detection and localization method to improve apple detection under adverse light conditions, where each input image is light enhanced for better detection performance. We incorporated a lightweight image enhancement module and improved the YOLOv5s detection model by adding an attention mechanism and improving the loss function to further enhance the detection results. Meanwhile, we established a binocular camera-based apple localization method based on the binocular stereo matching algorithm and combined it with our improved detection algorithm to realize apple detection and localization under adverse light conditions. The experimental results show that our method improved detection mAP by 3.1%, 6.8%, and 10.9% under normal light, overexposure, and adverse light conditions, respectively, and the average accuracy in terms of localization by 44% and 73% under normal light and adverse light, respectively, compared to the original YOLOv5s algorithm. The experimental results demonstrate the superiority of our method.

This method provides technical support for apple picking robots to accurately detect multiple fruit targets in real time in an outdoor environment. Future work will include the detection of other apple tree species or other crop fruits.

Author Contributions: Methodology, G.Z. and Y.T.; resources, G.Z. and C.Z.; software, G.Z. and W.Y.; writing, G.Z.; format calibration, Y.T. and C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key R&D Program of China under Grant 2023YFC3006805, and in part by the National Natural Science Foundation of China under Grant 31971668.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to them also being necessary for future essay writing.

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

References

1. Fróna, D.; Szenderák, J.; Harangi-Rákos, M. The challenge of feeding the world. *Sustainability* **2019**, *11*, 5816. [CrossRef]
2. Rockström, J.; Williams, J.; Daily, G.; Noble, A.; Matthews, N.; Gordon, L.; Wetterstrand, H.; DeClerck, F.; Shah, M.; Steduto, P.; et al. Sustainable intensification of agriculture for human prosperity and global sustainability. *Ambio* **2017**, *46*, 4–17. [CrossRef]
3. Tzounis, A.; Katsoulas, N.; Bartzanas, T.; Kittas, C. Internet of Things in agriculture, recent advances and future challenges. *Biosyst. Eng.* **2017**, *164*, 31–48. [CrossRef]
4. Musacchi, S.; Serra, S. Apple fruit quality: Overview on pre-harvest factors. *Sci. Hortic.* **2018**, *234*, 409–430. [CrossRef]

5. Bogue, R. Fruit picking robots: Has their time come? *Ind. Robot Int. J. Robot. Res. Appl.* **2020**, *47*, 141–145. [CrossRef]
6. Legun, K.; Burch, K. Robot-ready: How apple producers are assembling in anticipation of new AI robotics. *J. Rural Stud.* **2021**, *82*, 380–390. [CrossRef]
7. Jia, W.; Zhang, Y.; Lian, J.; Zheng, Y.; Zhao, D.; Li, C. Apple harvesting robot under information technology: A review. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420925310. [CrossRef]
8. Shamshiri, R.R.; Weltzien, C.; Hameed, I.A.; Yule, I.J.; Grift, T.E.; Balasundram, S.K.; Pitonakova, L.; Ahmad, D.; Chowdhary, G. Research and development in agricultural robotics: A perspective of digital farming. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 1–14. [CrossRef]
9. Wan, S.; Goudos, S. Faster R-CNN for multi-class fruit detection using a robotic vision system. *Comput. Netw.* **2020**, *168*, 107036. [CrossRef]
10. Jiao, Y.; Luo, R.; Li, Q.; Deng, X.; Yin, X.; Jia, W. Detection and localization of overlapped fruits application in an apple harvesting robot. *Electronics* **2020**, *9*, 1023. [CrossRef]
11. Sigov, A.; Ratkin, L.; Ivanov, L.A.; Xu, L.D. Emerging enabling technologies for industry 4.0 and beyond. *Inf. Syst. Front.* **2022**, 1–11. [CrossRef]
12. Zennayi, Y.; Benaissa, S.; Derrouz, H.; Guennoun, Z. Unauthorized access detection system to the equipments in a room based on the persons identification by face recognition. *Eng. Appl. Artif. Intell.* **2023**, *124*, 106637. [CrossRef]
13. Zhu, H.; Han, T.; Alhajyaseen, W.K.M.; Iryo-Asano, M.; Nakamura, H. Can automated driving prevent crashes with distracted Pedestrians? An exploration of motion planning at unsignalized Mid-block crosswalks. *Accid. Anal. Prev.* **2022**, *173*, 106711. [CrossRef]
14. De Araujo PR, M.; Lins, R.G. Cloud-based approach for automatic CNC workpiece origin localization based on image analysis. *Robot. Comput. Integr. Manuf.* **2021**, *68*, 102090. [CrossRef]
15. Mao, M.; Zhao, H.; Tang, G.; Ren, J. In-Season Crop Type Detection by Combing Sentinel-1A and Sentinel-2 Imagery Based on the CNN Model. *Agronomy* **2023**, *13*, 1723. [CrossRef]
16. Diwan, T.; Anirudh, G.; Tembhurne, J.V. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimed. Tools Appl.* **2023**, *82*, 9243–9275. [CrossRef]
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
18. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
19. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef]
21. Liu, F.; Hua, Z.; Li, J.; Fan, L. Dual UNet low-light image enhancement network based on attention mechanism. *Multimed. Tools Appl.* **2023**, *82*, 24707–24742. [CrossRef]
22. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep retinex decomposition for low-light enhancement. *arXiv* **2018**. [CrossRef]
23. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1780–1789.
24. Lv, F.; Lu, F.; Wu, J.; Lim, C. MBLLEN: Low-Light Image/Video Enhancement Using CNNs. *BMVC* **2018**, *220*, 4.
25. Zhang, Y.; Zhang, J.; Guo, X. Kindling the darkness: A practical low-light image enhancer. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1632–1640. [CrossRef]
26. Wei, X.; Zhang, X.; Li, Y. Tsn-ca: A two-stage network with channel attention for low-light image enhancement. In Proceedings of the International Conference on Artificial Neural Networks, Bristol, UK, 6–9 September 2022; Springer Nature: Cham, Switzerland, 2022; pp. 286–298. [CrossRef]
27. Wang, J.; Tan, W.; Niu, X.; Yan, B. RDGAN: Retinex decomposition based adversarial learning for low-light enhancement. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1186–1191. [CrossRef]
28. Wei, X.; Zhang, X.; Li, Y. Sarn: A lightweight stacked attention residual network for low-light image enhancement. In Proceedings of the 2021 6th International Conference on Robotics and Automation Engineering (ICRAE), Guangzhou, China, 19–22 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 275–279.
29. Thuan, D. Evolution of Yolo Algorithm and Yolov5: The State-of-the-Art Object Detention Algorithm. Bachelor's Thesis, Oulu University of Applied Sciences, Oulu, Finland, 2021.
30. Deng, L.; Li, H.; Liu, H.; Gu, J. A lightweight YOLOv3 algorithm used for safety helmet detection. *Sci. Rep.* **2022**, *12*, 10981. [CrossRef] [PubMed]

31. Cheng, D.; Meng, G.; Cheng, G.; Pan, C. SeNet: Structured edge network for sea–land segmentation. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 247–251. [CrossRef]
32. Imambi, S.; Prakash, K.B.; Kanagachidambaresan, G.R. *PyTorch. Programming with TensorFlow: Solution for Edge Computing Applications*; Springer Nature: Cham, Switzerland, 2021; pp. 87–104.