agriculture

MDPI

Article

Prediction of Feed Quantity for Wheat Combine Harvester Based on Improved YOLOv5s and Weight of Single Wheat Plant without Stubble

Qian Zhang *, Qingshan Chen, Wenjie Xu, Lizhang Xu and En Lu

School of Agricultural Engineering, Jiangsu University, Zhenjiang 212013, China; 2212216076@stmail.ujs.edu.cn (Q.C.); x1049443576@163.com (W.X.); justxlz@ujs.edu.cn (L.X.); jsluen@163.com (E.L.)

* Correspondence: zhangq_jsu@ujs.edu.cn

Abstract: In complex field environments, wheat grows densely with overlapping organs and different plant weights. It is difficult to accurately predict feed quantity for wheat combine harvester using the existing YOLOv5s and uniform weight of a single wheat plant in a whole field. This paper proposes a feed quantity prediction method based on the improved YOLOv5s and weight of a single wheat plant without stubble. The improved YOLOv5s optimizes Backbone with compact bases to enhance wheat spike detection and reduce computational redundancy. The Neck incorporates a hierarchical residual module to enhance YOLOv5s' representation of multi-scale features. The Head enhances the detection accuracy of small, dense wheat spikes in a large field of view. In addition, the height of a single wheat plant without stubble is estimated by the depth distribution of the wheat spike region and stubble height. The relationship model between the height and weight of a single wheat plant without stubble is fitted by experiments. Then, feed quantity can be predicted using the weight of a single wheat plant without stubble estimated by the relationship model and the number of wheat plants detected by the improved YOLOv5s. The proposed method was verified through experiments with the 4LZ-6A combine harvester. Compared with the existing YOLOv5s, YOLOv7, SSD, Faster R-CNN, and other enhancements in this paper, the mAP_{50} of wheat spikes detection by the improved YOLOv5s increased by over 6.8%. It achieved an average relative error of 4.19% with a prediction time of 1.34 s. The proposed method can accurately and rapidly predict feed quantity for wheat combine harvesters and further realize closed-loop control of intelligent harvesting operations.

Keywords: feed quantity prediction; wheat combine harvester; neural network; vehicle vision; height estimation

Citation: Zhang, Q.; Chen, Q.; Xu, W.; Xu, L.; Lu, E. Prediction of Feed Quantity for Wheat Combine Harvester Based on Improved YOLOv5s and Weight of Single Wheat Plant without Stubble. *Agriculture* **2024**, *14*, 1251. https://doi.org/10.3390/agriculture14081251

Academic Editor: Lixia Hou

Received: 21 June 2024
Revised: 25 July 2024
Accepted: 26 July 2024
Published: 29 July 2024

1. Introduction

Wheat is a major crop in China, which ensures China's food security. With the development of wheat harvesting mechanization and intelligent detection technology, detecting the feed quantity for wheat combine harvesters accurately and rapidly has become an important research direction of intelligent harvesters. Excessive feed quantity increases the load on working components, such as the threshing drum [1], which can easily cause congestion. Insufficient feed quantity results in inadequate load on the threshing drum, which may reduce the work efficiency [2–4].

The current methods for detecting the feed quantity primarily use parameters such as the torque of the harvester's transmission shaft [5,6], the pressure [7], and the power of the header hydraulic cylinder [8,9] to detect the feed quantity indirectly. Although these methods can provide feedback on the feed status of harvesters, the information obtained is not predictive. And there is not enough time for subsequent adjustment of operational parameters for the harvesters. Therefore, predicting the feed quantity for wheat combine harvesters accurately and rapidly is crucial. This is based on the crop parameters within

the area to be harvested. Accurate predictions help reduce the congestion rate, ensure efficiency, and advance the development of intelligent harvesting in China.

The available technologies for predicting the feed quantity mainly include spectral [10], radar [11], and machine vision [12] technologies, which are mounted on unmanned aerial vehicles (UAV) [13–15] and harvesters. UAV-mounted spectral technologies and machine vision technologies have advantages in rapidly estimating the biomass and yield of wheat [16]. However, these methods have drawbacks such as low resolution, turbulence disturbance [17], difficulty in detecting sheltered crops, and working with combine harvesters. Harvester-mounted radar technologies are highly accurate but prone to environmental interferences, limited in sampling information, and costly. Compared to other technologies, harvester-mounted machine vision [18] offers higher resolution for local detection, faster sampling with more information, and lower costs. In this paper, harvester-mounted machine vision technologies are more suitable for predicting the feed quantity for wheat combine harvesters with the tilt-shot method.

The two main types of harvester-mounted tilt-shot machine vision technologies are those based on the pixel area [19,20] of wheat images and those based on the number of wheat spikes [21]. The pixel area of wheat images refers to the number of pixels occupied by wheat spikes in the image. Under normal growth conditions, each mature wheat plant can produce multiple spikes. In this paper, unless otherwise specified, the term "single wheat plant" refers to "a branch of a wheat spike", meaning each branch of a mature wheat plant that bears a spike. The number of spikes can approximate the number of mature wheat plant branches. Therefore, some scholars estimated the wheat biomass based on the pixel area of wheat images and the pixel–mass relationship. However, due to some factors, such as the tilt-shot method and perspective distortion (object appears larger when closer and smaller when farther), it is difficult to establish a pixel–mass relationship to estimate the feed quantity. Therefore, another group of scholars estimated the weight of a single wheat plant by averaging the wheat elevation in a whole field. They then estimated the wheat biomass using the number of wheat spikes and the weight of a single wheat plant, addressing the challenge of establishing a pixel–mass relationship.

However, due to variations in growth environment, soil nutrient distribution, and external disturbances, the height of wheat plants varies in different areas within the same field. Additionally, wheat tillering can also cause differences in spike height. It is difficult to improve the predicting accuracy of the feed quantity based on the number of wheat spikes and the weight of a single wheat plant in a whole field. Therefore, this paper proposes the concept of the weight of a single wheat plant without stubble, i.e., the weight of the remaining stalks and spike after cutting off the stubble part from a single wheat plant. By detecting the height of a single wheat plant without stubble, its weight can be estimated, and a relationship between height and weight can be established. Additionally, based on the weight of a single wheat plant without stubble and the number of wheat plants, it is possible to predict the feed quantity.

Currently, detecting wheat spikes primarily consists of two methods: traditional image processing and Convolutional Neural Networks (CNNs). The primary methods for spike detection based on traditional image processing include Super-pixel [22], skeleton extraction [23], morphology [24], watershed [25], and Support Vector Machine (SVM) [26], etc. These approaches are susceptible to variations in characteristics such as color, brightness, and texture. They also lack real-time detection capabilities. The watershed algorithm is extremely salient for the segmentation of objects with complex boundaries in images. Nonetheless, the outcome of the segmentation may encompass regions subjected to over-segmentation or under-segmentation, posing challenges for real-time detection in densely planted wheat fields.

CNNs exhibit exceptional generalization and robustness for detecting wheat spikes across varying lighting conditions. The leading CNN architectures for wheat spikes detection currently are: Single Shot Multi-Box Detector (SSD) [27], Faster Region-based Convolutional Neural Network (Faster R-CNN) [28], You Only Look Once version 5 small

(YOLOv5s) [29,30], and You Only Look Once version 7 (YOLOv7) [31]. YOLOv5s stands out among these methods for its exceptional balance of speed, accuracy, adaptability, and robustness, making it particularly suitable for real-time wheat spikes detection studies. Outdoor fields present challenges with dense wheat, overlapping organs, and various spike sizes. The existing YOLOv5s struggles to precisely detect small and dense wheat spikes in large field of view (FOV). This highlights the need for model improvements and optimizations.

Most commonly, methods used to estimate the height of a single wheat plant rely on binocular disparity to gather global point cloud data or average elevations of wheat in a whole field. The methods are amalgamated with algorithm inversion [32] and ground modeling [33] for the ascertainment of ground height, which is subsequently aimed at getting the height of wheat plants. These methods are advantageous for detecting wheat plant height across large areas. However, for single plant height detection, global point cloud data [34] computation is redundant and slow, and the accuracy of global elevation averages is relatively low. The dense growth and overlapping organs of wheat significantly increase congestion. Current methods like algorithm inversion and ground modeling use proximal ground data from the header and adjacent harvested areas to deduce ground height in upcoming detection areas. While accurate in regions with slow terrain changes, they struggle with areas of residue accumulation or significant height variability along harvest boundary ridges.

Based on the above issues, this paper enhances the existing YOLOv5s through three main modifications: introducing an attention optimization of the Backbone structure, implementing a multi-scale features extraction module with a tiered residual structure in the Neck, and amending a Head structure targeting the detection of small objects. This paper estimates the height of a single wheat plant without stubble based on the depth distribution of the wheat spike region and stubble height. It then establishes a relationship between the height and weight of a single wheat plant without stubble. Furthermore, analyzing the number of wheat spikes and the weight of a single wheat plant without stubble allows for predicting the feed quantity for wheat combine harvesters accurately and rapidly. The proposed method was verified through experiments with images acquired on the 4LZ-6A intelligent combine harvester, which facilitates the prediction of feed quantity for wheat combine harvesters. The proposed method contributes to furthering the advancement toward closed-loop control of intelligent harvesting operations.

2. Wheat Combine Harvester Feed Quantity Prediction System and Dataset Construction

2.1. Feed Quantity Prediction Definition and Acquisition System Construction

There is no definitive definition for the feed quantity for wheat combine harvesters within China. The "Guidelines for the Promotion and Certification of Agricultural Machinery" defines the feed quantity for ratoon rice combine harvesters. It describes this quantity as the total mass of grain, stalks, and cleaning residuals received by the combine harvester per second. This is measured in kilograms per second (kg/s). Referencing the feed quantity definition for ratoon rice combine harvesters, this study employs the total mass of grain, stalks, and cleaning residuals that the wheat combine harvesters receives per second as the feed quantity for wheat combine harvesters. To predict the feed quantity, the calculation is based on the harvester's operating speed V, the direction of operation, the cutter height H_{lc}, and the maximum cutting width L_g. The total biomass from the cutter to the top of wheat plants for a unit of time (area = L_g m $\times$ (V m/s $\times$ 1 s)) is taken as the predictive value of feed quantity. In this study, unless specified otherwise, the predictive value of feed quantity for wheat combine harvester is collectively referred to as feed quantity.

As illustrated in Figure 1, due to some factors, such as the vibration during harvesting operations and the load-bearing capacity of the combine harvester's outer wall, a visual data acquisition system with the tilt-shot method was mounted on the 4LZ-6A multi-functional intelligent crawler-type combine harvester. This harvester was developed by our team in Zhenjiang, China. The camera used was the STEREOLABS ZED 2i, manufactured

in San Francisco, CA, USA. It has a 110° (H) × 70° (V) × 120° (D) field of view, 2.1 mm focal length, 0.3–20 m depth range, and 5.07% TV distortion. It was mounted on top of the harvester's cabin, tilted forwards at a 35° angle relative to the horizontal. The image processing unit used an Advantech MIC-7700 industrial computer, manufactured in Taipei, Taiwan. It was equipped with a 10th generation Intel motherboard and an Intel Core i7-6700 processor, manufactured in Santa Clara, California, USA. The unit also featured an NVIDIA GTX 1650 graphics card, manufactured in Santa Clara, California, USA, 32 GB of memory, and was capable of operating in temperatures from 0 °C to 60 °C. The main control unit communicates with the image processing unit via the CAN bus.

Figure 1. Visual data acquisition system for feed quantity.

2.2. Feed Quantity Prediction Coordinate Model and Distortion Correction

Figure 2 depicts the construction of a feed quantity prediction coordinate model. $O_{c1} - X_{c1}Y_{c1}Z_{c1}$ and $O_{c2} - X_{c2}Y_{c2}Z_{c2}$ represent the camera coordinate systems of the left and right cameras, respectively, with $O_{c1} - X_{c1}Y_{c1}Z_{c1}$ being the base camera coordinate system. The angle θ signifies the inclination of the Z_{c1} axis of this coordinate system with respect to the horizontal plane. Image coordinate systems are denoted by $O_{i1} - X_{i1}Y_{i1}$ and $O_{i2} - X_{i2}Y_{i2}$, while pixel coordinate systems are signified by $O_{o1} - U_1V_1$ and $O_{o2} - U_2V_2$. The world coordinate system is represented by $O_w - X_wY_wZ_w$, where the X_w and Y_w axes are parallel to the horizontal plane, and the Z_w axis extends vertically upwards. The origin of the world coordinate system, O_w, and the base camera coordinate system origin, O_{c1}, share the same vertical axis perpendicular to the horizontal plane, at a distance H apart.

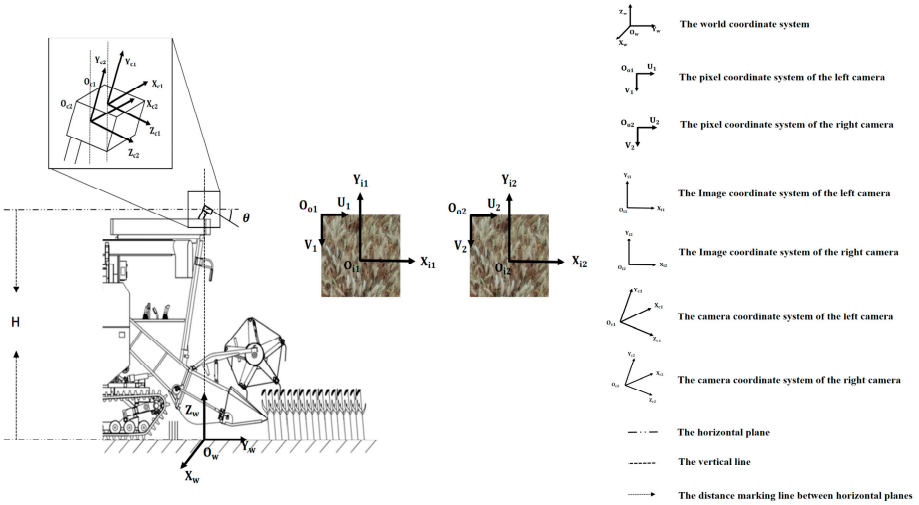


Figure 2. Statistics of feed quantity prediction coordinate model.

Within the pixel coordinate system $O_{o1} - U_1V_1$, the image coordinate system $O_{i1} - X_{i1}Y_{i1}$ has its origin coordinate at (u_{o1}, v_{o1}), and within the pixel coordinate system $O_{o2} - U_2V_2$, the

image coordinate system $O_{i2} - X_{i2}Y_{i2}$ has its origin coordinate at (u_{o2}, v_{o2}). Both left and right cameras have identical pixel densities with lengths and widths denoted by d_x and d_y. Point coordinates in the pixel coordinate system are (u_1, v_1) and (u_2, v_2), while in the image coordinate system, they are (x_{i1}, y_{i1}) and (x_{i2}, y_{i2}). The transformation relations from pixel coordinates to image coordinates for both cameras are provided in Equations (1) and (2):

$$\begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_{o1} \\ 0 & \frac{1}{d_y} & v_{o1} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{i1} \\ y_{i1} \\ 1 \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_{o2} \\ 0 & \frac{1}{d_y} & v_{o2} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{i2} \\ y_{i2} \\ 1 \end{bmatrix} \tag{2}$$

Optical properties and design constraints of camera lenses can lead to lens distortion, which degrades the image quality [35]. Radial distortion, as opposed to tangential distortion, more significantly affects the geometric fidelity of wheat images. Thus, our study prioritizes radial distortion correction. Calibration images are gathered during the camera calibration phase to determine and adjust radial distortion coefficients k_1, k_2, k_3 by minimizing the discrepancies between the actual image coordinates and the ideal coordinates. In the image coordinate system, the points of the left and right cameras are (x_{ic1}, y_{ic1}) and (x_{ic2}, y_{ic2}), respectively, while (x_{i1}, y_{i1}) and (x_{i2}, y_{i2}) correspond to the coordinates on the distorted image. Equations (3) and (4) are employed for the radial distortion correction of the images:

$$\begin{cases} x_{i1} = x_{id1} + (x_{id1} - x_{ic1}) \cdot \left(k_1 r_1^2 + k_2 r_1^4 + k_3 r_1^6\right) \\ y_{i1} = y_{id1} + (y_{id1} - y_{ic1}) \cdot \left(k_1 r_1^2 + k_2 r_1^4 + k_3 r_1^6\right) \end{cases} \tag{3}$$

$$\begin{cases} x_{i2} = x_{id2} + (x_{id2} - x_{ic2}) \cdot \left(k_1 r_2^2 + k_2 r_2^4 + k_3 r_2^6\right) \\ y_{i2} = y_{id2} + (y_{id1} - y_{ic2}) \cdot \left(k_1 r_2^2 + k_2 r_2^4 + k_3 r_2^6\right) \end{cases} \tag{4}$$

where (x_{id1}, y_{id1}) and (x_{id2}, y_{id2}) are the image coordinates after distortion correction, r_1 is the distance from (x_{id1}, y_{id1}) to (x_{ic1}, y_{ic1}), and r_2 is the distance from (x_{id2}, y_{id2}) to (x_{ic2}, y_{ic2}).

The focal lengths of both the left and right cameras are f, and the transformation relationship between the image coordinate systems after radial distortion correction, and the camera coordinate systems are given by Equations (5) and (6):

$$\begin{bmatrix} x_{i1} \\ y_{i1} \\ 1 \end{bmatrix} = \frac{1}{z_{c1}} \cdot \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_{c1} \\ y_{c1} \\ z_{c1} \\ 1 \end{bmatrix} \tag{5}$$

$$\begin{bmatrix} x_{i2} \\ y_{i2} \\ 1 \end{bmatrix} = \frac{1}{z_{c2}} \cdot \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_{c2} \\ y_{c2} \\ z_{c2} \\ 1 \end{bmatrix} \tag{6}$$

In the base coordinate system of the camera $O_{c1} - X_{c1}Y_{c1}Z_{c1}$, the depth value of the point (x_{c1}, y_{c1}, z_{c1}) is denoted by its Euclidean distance from the origin O_{c1}. As shown in Figure 3a,b, the RGB image in the image coordinate system can obtain the depth image in the camera coordinate system through Equations (5) and (6). By color coding, the image depth values are mapped into the RGB color space. After aligning and superimposing the calibrated coordinate systems of the RGB image and the color-coded depth image, the fused image shown in Figure 3c is obtained, which can represent the image's depth information more intuitively.

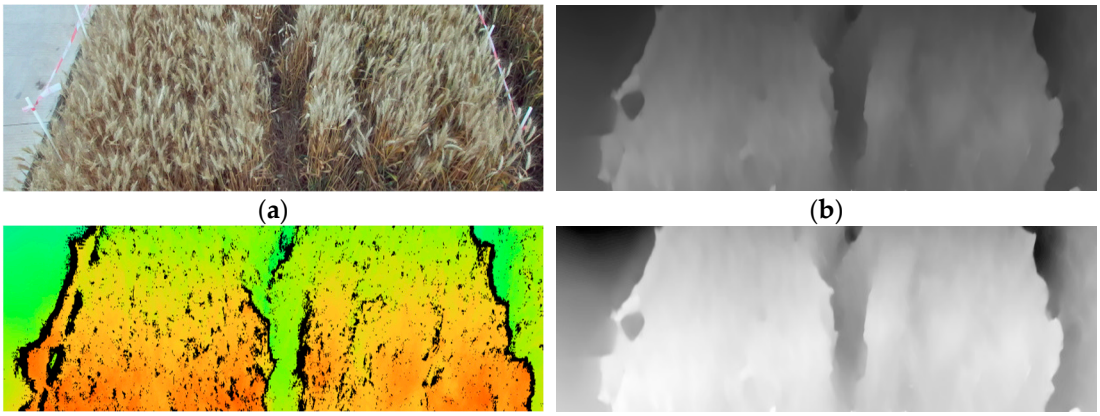


Figure 3. Schematic of coordinate conversion. (**a**) Left camera RGB image. (**b**) Depth image. (**c**) Fusion of depth image and RGB image. (**d**) Elevation image.

The acquisition of calibration images via the stereo camera facilitates the transformation relationship between the camera coordinate systems and the world coordinate systems through Equations (7) and (8).

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = T \cdot \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{7}$$

$$T = \begin{bmatrix} R & t_v \\ 0 & 1 \end{bmatrix}, \; t = \begin{pmatrix} 0 & 0 & \mathrm{H} \end{pmatrix}^{-1} \tag{8}$$

where R denotes the rotation matrix, illustrating the rotation transformation that occurs from the camera coordinate system to the world coordinate system, as deduced via Equation (9). concurrently, t_v defines the translation vector, portraying the shift transformation from the camera coordinate system to the world coordinate system.

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(90+\theta) & -\sin(90+\theta) \\ 0 & \sin(90+\theta) & \cos(90+\theta) \end{bmatrix} \tag{9}$$

The elevation value of point (x_w, y_w, z_w) in the world coordinate system $\mathrm{O_w} - \mathrm{X_w Y_w Z_w}$ is determined by its perpendicular distance to the $\mathrm{X_w} - \mathrm{Y_w}$ plane, laying the groundwork for subsequent estimations of wheat plant height. As illustrated in Figure 3d, the elevation image in the world coordinate system can be obtained from the depth image in the camera coordinate system through Equations (7) and (8) [36].

2.3. Wheat Spikes Dataset Construction

In May 2023, a data collection initiative was undertaken in the wheat fields of Shiye Satellite Village, Zhenjiang, Jiangsu Province, focusing on mature phases of three wheat varieties: "Zhenmai15", "Zhenmai18", and "Zhenmai12". ZED 2I camera was used to collect image data of each experimental area during the 09:00–17:00 period under both sunny and cloudy weather conditions. The camera had a shooting angle of 35° and a shooting height of 2.8 m, capturing a total of 1720 images of mature wheat spikes of the three varieties across different weather conditions and time periods. Figure 4 shows the wheat spikes image data collected in a field environment.

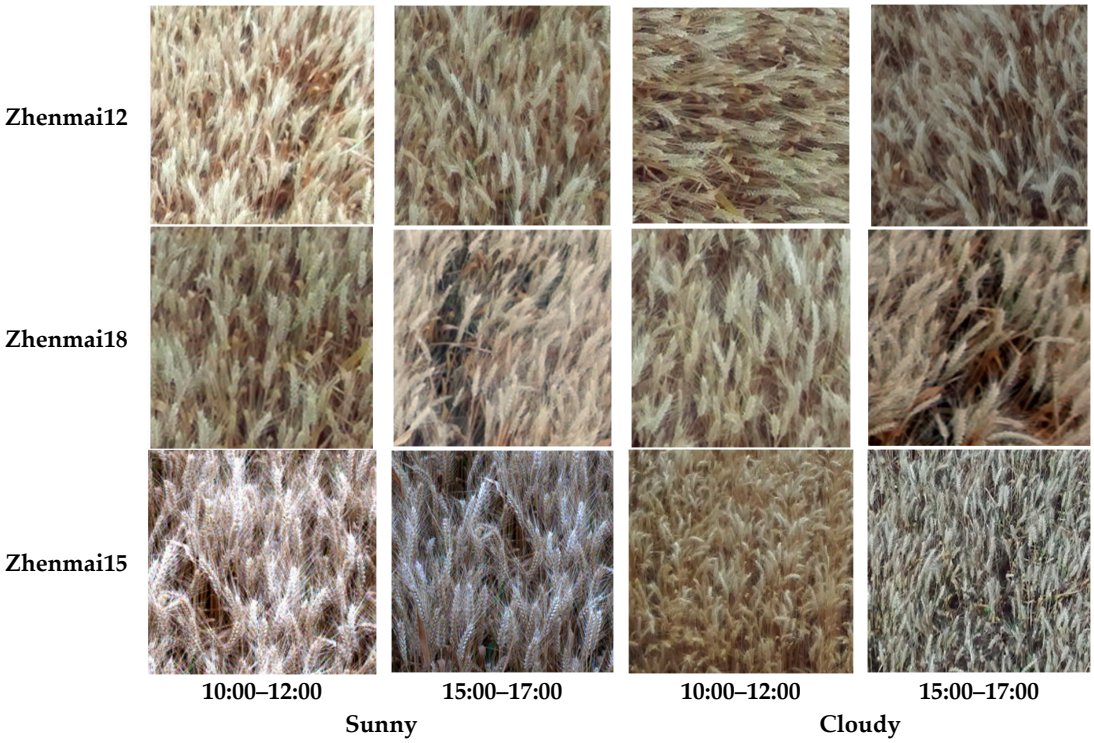


Figure 4. Wheat spikes image data collected in a field environment.

As depicted in Figure 5, the collected images were processed to extract the region where wheat was located to construct a wheat spikes dataset: This was achieved first through a series of image preprocessing techniques such as denoising, filtering, and sharpening to augment the contrast of the images. Subsequently, the areas designated for wheat detection were extracted based on color space transformation and thresholding segmentation, followed by alternating morphological opening and closing operations to enhance edge information in the images, thereafter creating a binary mask for Region of Interest (ROI) to fill the non-wheat detection areas in the images with a grayscale value set to 0. Finally, the images are divided into image tiles, with each image patch sized at 640 × 640 pixels.

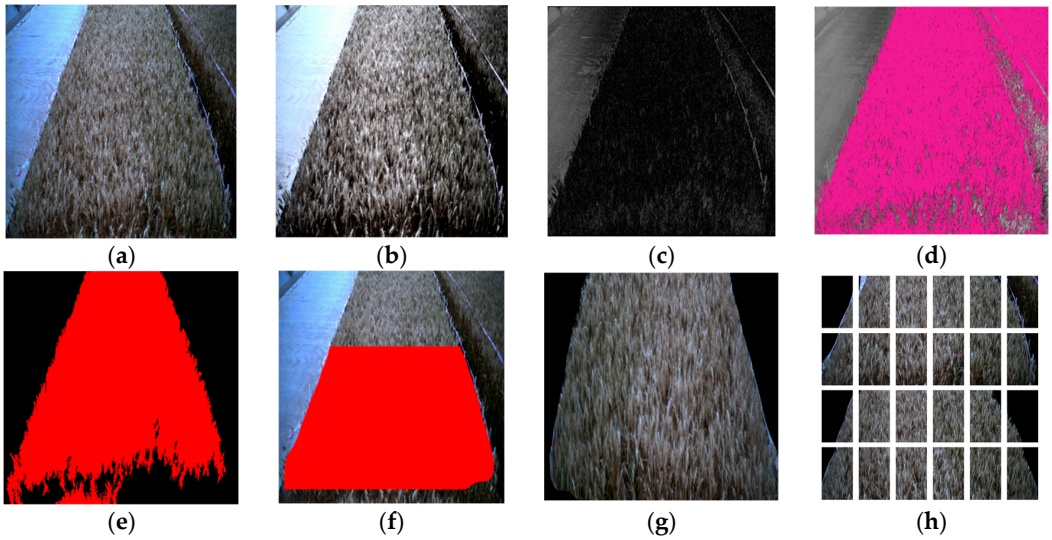


Figure 5. Image preprocessing. The purple areas indicate wheat regions after global thresholding, while the red areas indicate wheat regions after morphological processing and ROI segmentation. (**a**) Original image. (**b**) Multi-scale enhancement. (**c**) Color space conversion. (**d**) Global thresholding. (**e**) Morphological processing. (**f**) ROI segmentation. (**g**) Image filling. (**h**) Image tiling.

After performing the image processing operations shown in Figure 5, the initial wheat spikes dataset was obtained, with each image having a pixel size of 640 × 640. The initial wheat spikes dataset obtains 4800 sample images (400 images of each of the three different wheat varieties across variable lighting conditions, specifically during morning and afternoon periods under sunny and cloudy weather conditions). This paper used data augmentation to expand each image in the initial dataset with the following steps: flipping images vertically or horizontally with 50% probability, adjusting brightness between 0.8 and 1.5 times, applying random Gaussian blur, rotating images between −15° and 15°, and resizing images to 0.8 to 0.95 times their original size. These steps increased data quantity and diversity, helping the model handle image deformations, lighting variations, and noise. Figure 6 shows some wheat spike images after data augmentation.

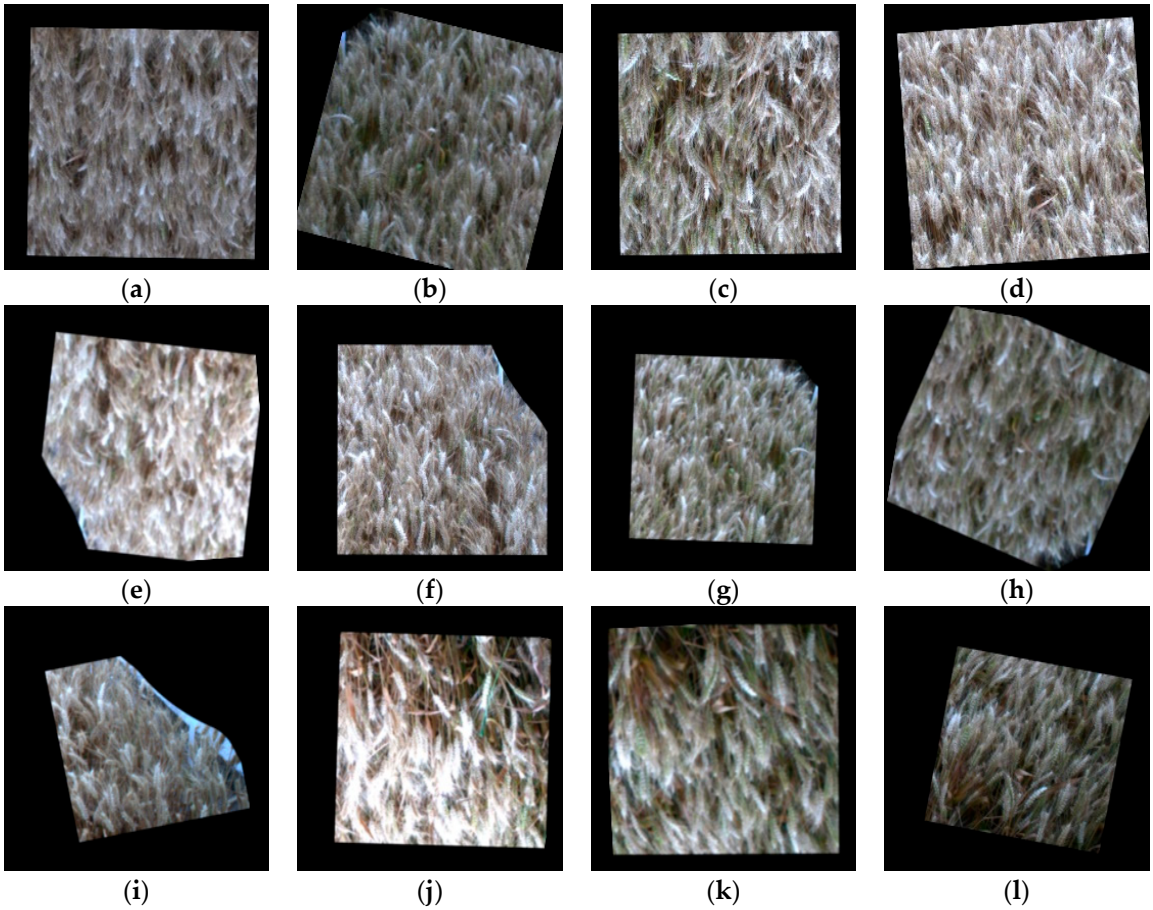


Figure 6. Schematic of partial wheat spikes images after data augmentation. (**a**) Vertical flip and resizing. (**b**) Rotation and resizing. (**c**) Brightness adjustment and resizing. (**d**) Brightness adjustment and rotation. (**e**) Gaussian blur and rotation. (**f**) Brightness adjustment and resizing. (**g**) Gaussian blur and resizing. (**h**) Vertical flip and rotation. (**i**) Rotation and resizing. (**j**) Vertical flip and brightness adjustment. (**k**) Vertical flip and Gaussian blur. (**l**) Rotation and resizing.

The augmented dataset was annotated using Labelme, and the annotated dataset was named VOC_Wheatear. The VOC_Wheatear dataset encompasses 12,914 images, comprising one category of 'Wheatear' labels with a total of 9.81×10^4 annotation boxes. Figure 7 presents the statistical information of this dataset. The VOC_Wheatear dataset was randomly divided into training, validation, and test sets in a ratio of 7:1.5:1.5, with the training set containing 9040 images. Both the validation and test sets contain 1937 images each.

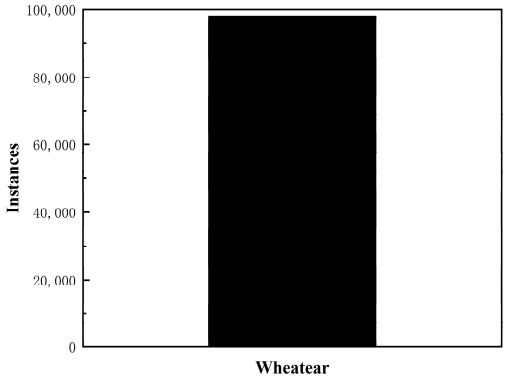


Figure 7. The names and quantity of samples in the VOC_Wheatear dataset.

3. Improvements of the Existing YOLOv5s Based on Multi-Scale Features of Small Objects and Attention Optimization

3.1. The Existing YOLOv5s

As shown in Figure 8, the existing YOLOv5s mainly consists of three parts: Backbone, Neck, and Head [37]. Each part is primarily composed of three basic blocks: CBS (Convolution, Batch Normalization, and SiLU), C3 (three CBS units and one Bottleneck unit), and SPPF (Spatial Pyramid Pooling Fusion). The Backbone harnesses residual frameworks and feature reuse stratagem to distill image features, comprising CBS, C3, and SPFF modules, with C3 acting as the pivotal residual feature learning module, containing three standard convolutional layers, and the SPPF module uses three 5 × 5 max-pooling operations. The Neck employs PAN (Path Aggregation Network) and FPN (Feature Pyramid Network), where the PAN combines bottom-up spatial information with top-down semantic information, and the FPN fuses feature maps from different levels through up-sampling and down-sampling to generate a multi-scale feature pyramid. The Head transforms features of varying scales and generates detection results.

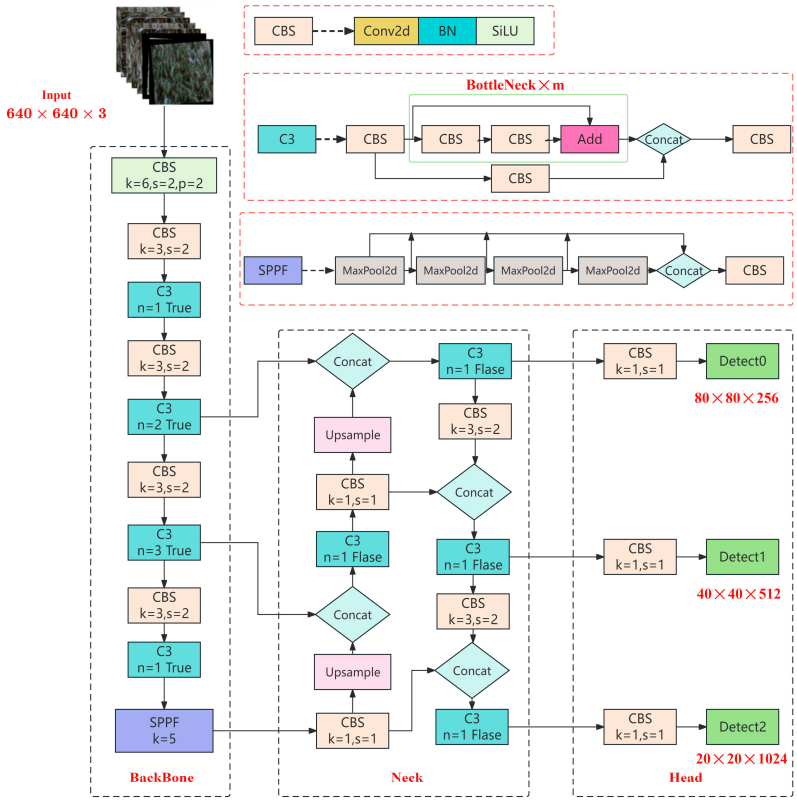


Figure 8. Schematic of the existing YOLOv5s structure.

3.2. The Improved YOLOv5s

The existing YOLOv5s is compromised by an absence of attention optimization mechanisms, leading to low attention to wheat spikes in the feature space during feature extraction. This shortfall makes it difficult to effectively weaken the interference of background information such as leaves, stalks, and ground, resulting in low recognition accuracy of wheat spikes.

Existing global attention mechanisms [38,39] are typically utilized to process each pixel or sector in an image so that the model can focus on the entire image simultaneously. However, their redundant computational load does not suit the real-time prediction needs of the feed quantity discussed in this paper. Additionally, as the wheat combine harvester commences its continuous operation, the same wheat spike exhibits scale differences. The C3 module entrenched within the existing Neck architecture is imbued with a rather homogeneous scale of feature extraction, making it challenging to meet the multi-scale feature extraction requirements of wheat spikes.

The harvester-mounted tilt-shot camera yields a large FOV, and it is challenging to detect small and dense wheat spikes. The detection layers within the existing Head architecture struggle to accurately detect these small targets, engendering a proclivity for missed detections. Consequently, this paper advances the existing YOLOv5s by integrating multi-scale feature extraction capabilities in large FOV. The main improvements are as follows:

(1) To amplify the network's attentiveness towards small wheat spikes, mitigate background interference, and diminish computational superfluity, an attention optimization based on a set of compact bases is proposed for the Backbone structure. The attention optimization does not follow a full-image process. The existing Backbone structure, which lacks attention optimization, is strategically enhanced through the integration of an Expectation–Maximization Attention (EMA) mechanism [40] alongside Dropout layers.

(2) To enhance the multi-scale feature extraction process within the network, a C3Res2NetBlock module featuring a hierarchical residual structure is incorporated into the Neck structure. This module improves the resolution in extracting multi-scale features of wheat spikes while also reducing the network's parameters and computational costs.

(3) Aiming to improve the detecting accuracy of small wheat spikes, an improved Head architecture focused on small targets is delineated. This Head framework employs larger-scale feature maps to replace the original feature maps. The improved YOLOv5s model structure is shown in Figure 9.

3.2.1. An Attention-Optimized Backbone Structure Based on a Set of Compact Bases

EMA, using the Expectation–Maximization (EM) algorithm, iteratively computes the maximum likelihood solution of the variable model. Based on this maximum likelihood, running the attention mechanism on this basis can reduce the complexity of the attention process. Within the dataset where the variable x_i is situated, a corresponding latent variable z_i is discerned, emblematic of the class S_i to which x_i is ascribed, and the probability p_i of its occurrence. The EM algorithm aims to maximize likelihood through the E-step and the M-step process. The theoretical Equation for the EM algorithm is:

$$\begin{cases} Q(\theta_e, \theta^t) = \sum\limits_{z} \sum\limits_{i=1}^{n} [\ln(p(x_i, z_i|\theta_e))] \cdot p(z_i|x_i, \theta^t) \\ \theta^{t+1} = argmax\{Q(\theta_e, \theta^t)\} \end{cases} \tag{10}$$

where θ_e is the set of all parameters of the model. In the E-step, θ^t is used to denote the posterior distribution of z, and then this posterior distribution is used to calculate the expected value of the likelihood function $\ln(p(x_i, z_i|\theta_e))$. In the M-step, maximize function to determine the new posterior distribution θ^{t+1}; alternate execution of the E-step and M-step until the convergence criterion is met.

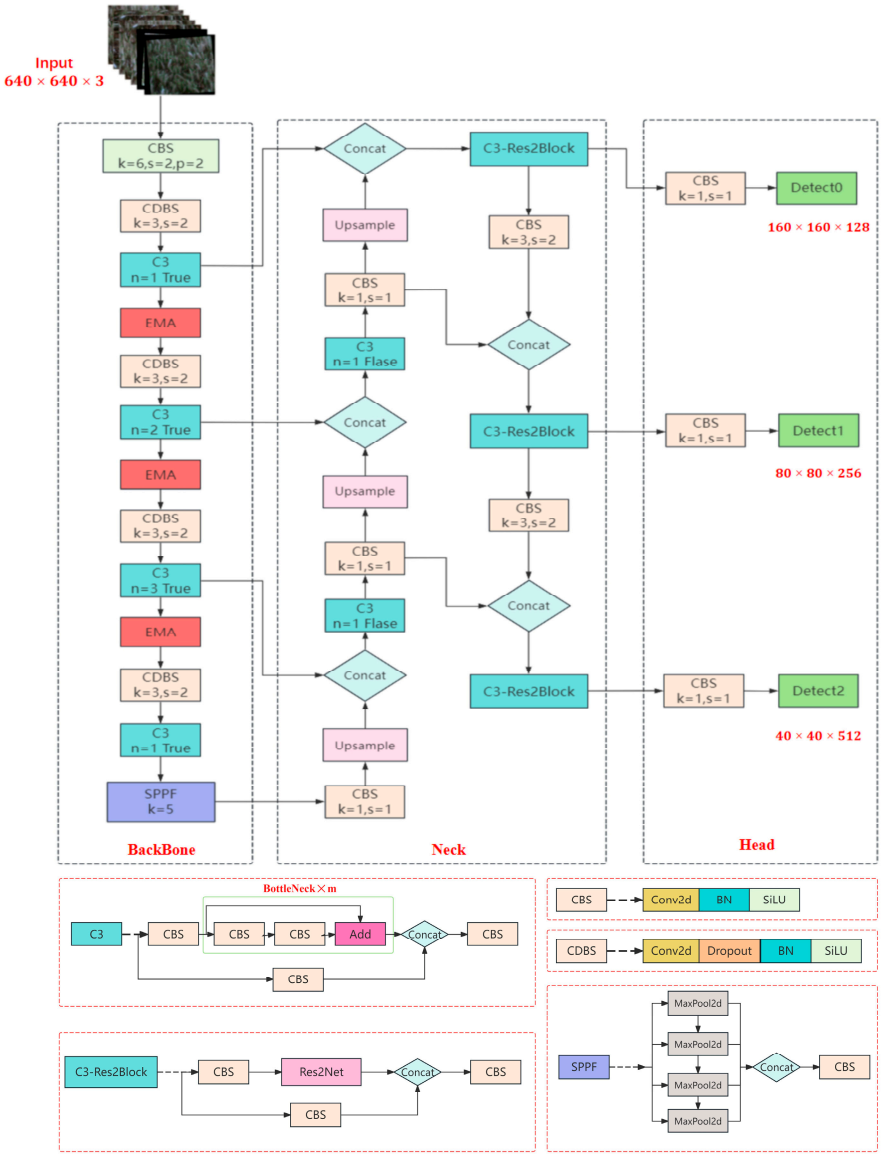


Figure 9. Schematic of the improved YOLOv5s structure.

The structure of EMA is depicted in Figure 10, encompassing three distinct steps: Attribution Estimation (AE), Likelihood Maximization (AM), and Data Re-estimation (AR). Commencing with the input x_i and a foundational base μ, the AE phase is dedicated to ascertaining the latent variable z_i, effectively operating as the E-step within the EM algorithmic sequence.

The AM-step proceeds to refine the base μ through the utilization of likelihood estimates, akin to the M-step in EM methodology. Subsequently executing the AE-step and AM-step in an alternating sequence for a predetermined number of cycles, the AR-step employs the stably converged base μ alongside z_i to reconstruct x_i into y_i, completing the process with the output.

Presented with an input $X = [x_1, x_2, \cdots x_{ec}] \in R^{N \times C}$, a foundational base value $\mu \in R^{K \times C}$, a latent variable $Z \in R^{N \times K}$. Within the confines of the AE-step, the duty attributed to the kth base relative to the nth pixel can be ascertained:

$$p(X_n|\mu_k) = K(X_n, \mu_k) \tag{11}$$

$$z_{nk} = \frac{K(X_n, \mu_k)}{\sum_{j=1}^{K} K(X_n, \mu_j)} \tag{12}$$

where K represents the general kernel function.

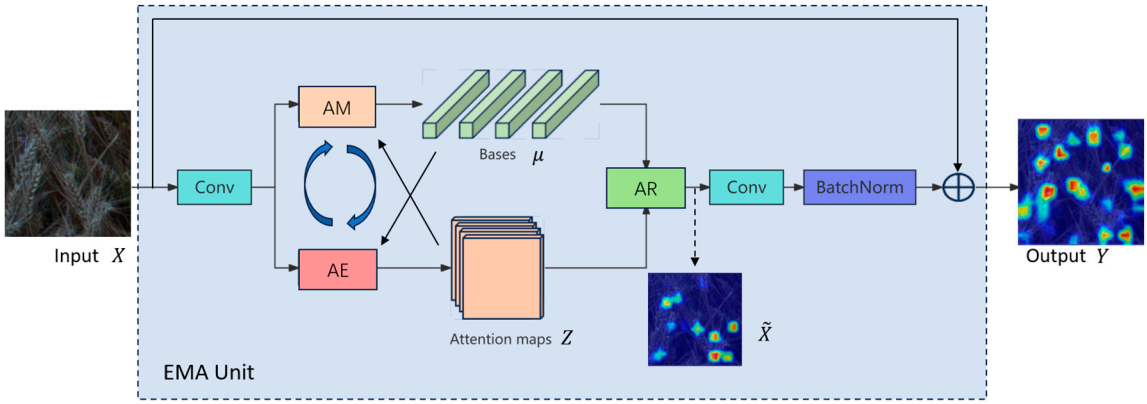


Figure 10. Schematic of the EMA structure.

Progressing to the AM-step, there is an adaptation in the base values where μ is ascertained as the weighted summation of X, the kth base thus defined:

$$\mu_k = \frac{z_{nk}^{t_e} X_n}{\sum_{m=1}^{N} z_{mk}^{t_e}} \tag{13}$$

The AE-step and AM-step are iteratively executed for t_e steps, therefore allowing for the re-estimation of X using the approximate convergence of μ and z.

The main purpose of utilizing EMA in this paper is to enhance the network's attentiveness towards wheat spikes within the feature space, reduce interference from irrelevant background information, and eliminate the process of computing attention over the entire image. By iterating a set of compact bases through the EM algorithm and running the attention mechanism on these bases, it is possible to significantly reduce the network's computational redundancy.

3.2.2. Multi-Scale Feature Extraction Module with Hierarchical Residual Structure

Aiming to improve the network's multi-scale feature extraction capability, this paper incorporates a C3Res2NetBlock module with a hierarchical residual structure into the Neck. The Res2Net [41] enhances the network's expressivity by introducing a multi-branch structure and incrementally increased resolution. As depicted in Figure 11, the structure of the C3Res2NetBlock involves channeling input feature maps through twin convolutional layers with identical kernel dimensions and halved channel outputs. These maps then undergo BatchNorm and SiLU processes before bifurcating into two branches. These branches serve the dual purposes of reducing dimensions and extracting salient features. One branch, having passed through the Res2Net module, merges with the other along the channel dimension. Subsequently, it undergoes convolution, BatchNorm, and SiLU processes to generate an output feature map containing multi-scale feature information.

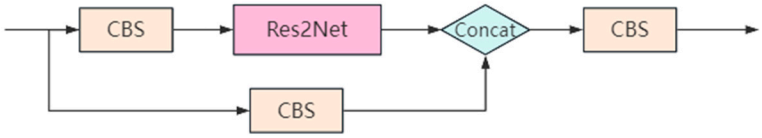


Figure 11. Schematic of the C3Res2NetBlock structure.

As illustrated in Figure 12, within the confines of a singular residual block, Res2Net employs grouped convolution to subdivide the input feature map's channels equitably; the resulting subdivisions are tackled with an array of smaller convolutional kernels to

imbue the model with a lightweight architecture; and a stairway-like concatenation is used to augment the count of scales that the output feature map may represent.

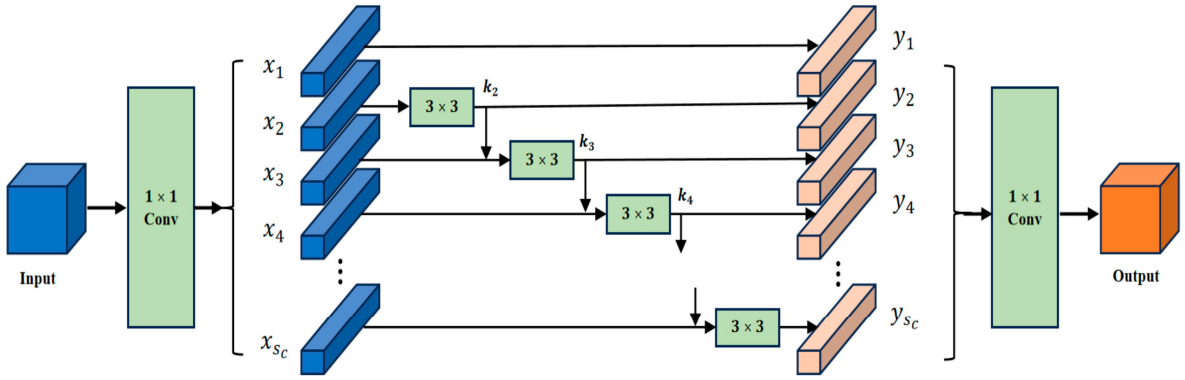


Figure 12. Schematic of the Res2Net structure.

The Equation of Res2Net is shown below:

$$y_m = \begin{cases} x_m & m = 1 \\ K_m(x_m) & m = 2 \\ K_m(x_m + y_{m-1}) & 2 < m \leq s_c \end{cases} \quad (14)$$

Res2Net adopts a 1 × 1 convolution that modifies the output channel count of the feature map to n_m. Subsequently, a split operation is leveraged to evenly segregate the input feature map along the channel dimension into s_c subsets, denoted as x_m, where $m = \{1, 2, \ldots, s_c\}$. Each subset x_m retains the same scale as the original feature map, with the channel quantity reduced to n_m/s_c. Besides the initial convolution set x_1, each x_m undergoes a subsequent 3 × 3 convolution, represented as K_m, and the output post-convolution is indicated as y_m. The current x_m, summed with the prior output y_{m-1}, forms the input for K_m. This hierarchical input amalgamation of each K_m ensures that every y_m is enriched with more comprehensive multi-scale features atop the basis established by y_{m-1}.

3.2.3. Enhanced Head Architecture for Small Targets Detection

The existing YOLOv5s model processes images through down-sampling at ratios of 8×, 16×, and 32×, producing feature maps of dimensions 80 × 80, 40 × 40, and 20 × 20. The VOC_Wheatear dataset developed for this paper comprises predominantly small wheat spikes, with the majority being less than 32 × 32 pixels in size. After down-sampling, the feature maps provide sparser details of small wheat spikes. This makes detection layers, designed for the original scale, inadequate for spotting these small targets. Consequently, this results in missed detections.

Therefore, this paper focuses on the scale characteristics of wheat spikes and network structure improvements. The Head structure was enhanced by introducing a 160 × 160 feature map size layer in the detection layer, replacing the original 20 × 20 size. This new layer offers rich positional information and detailed features for small targets. Consequently, it significantly improves the detection accuracy of small and dense wheat spikes within a large FOV. In this paper, unless specified otherwise, the detection layer with a feature map size of 20 × 20 is denoted as the P5 layer, and the one with a size of 160 × 160 is denoted as the P2 layer. The term P2–P5 refers to the substitutive enhancement where the P2 detection layer replaces the P5 detection layer.

4. Estimation of Wheat Height without Stubble Based on Depth Distribution of Spikes and Stubble Height

The height of a single wheat plant significantly affects its weight, which is crucial for predicting feed quantity for combine harvesters. This paper estimates the height of a single

wheat plant without stubble based on the stubble height, which is used for feed quantity predictions. This height refers to the remaining stalk and spike after removing the stubble.

Images from the harvester's tilt-shot camera include various disruptive elements, making stubble height estimation challenging. To address this, the paper introduces a method for detecting the depth distribution of wheat spikes and estimating stubble height using diverse ground information, therefore improving the estimation of the height of a single wheat plant without stubble.

4.1. Wheat Spikes Detection and Counting Based on the Improved YOLOv5s

Using transfer learning [42], this paper accelerates training, enhances model generalization, and reduces overfitting. The improved YOLOv5s was pretrained on the COCO dataset [43], which includes a wide range of common objects and annotations. We then fine-tuned it on the VOC_Wheatear dataset to focus on wheat spike features, with training parameters set as follows: learning rate = 0.001, weight decay = 1×10^{-4}, and momentum = 0.9. Training, conducted with a batch size of 8, was monitored for loss reduction and validated every 10 epochs until the loss stabilized. The enhanced YOLOv5s was then used to detect and count wheat spikes in the harvest area.

As shown in Figure 13, factoring in the harvester's operating speed V(V = 0.6 m/s) and the maximum cutting width L_g = 2 m. The area to be harvested (A1–A2–A3–A4) is defined after correcting for perspective using Inverse Perspective Mapping (IPM). The reference area is A1–A4–A5–A6, and the ROI within the FOV is A2–A3–A5–A6. After preprocessing the image (denoising, filtering, and sharpening), the improved YOLOv5s detects and counts wheat spikes. The bounding box coordinates are then used to extract spike regions for calculating depth distribution and elevation values.

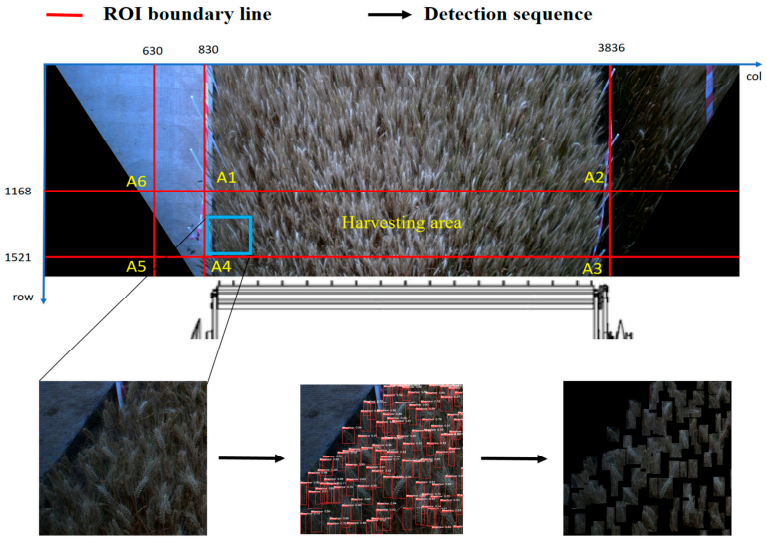


Figure 13. Schematic of wheat spikes detection and counting.

4.2. Calculation of Depth Distribution and Elevation Values of Spike Region of Single Wheat Plant

The depth distribution and the elevation values of the spike area are computed using the coordinates within the image coordinate system based on the transformation relations from the image coordinate system to the camera coordinate system through Equations (5) and (6).

By iterating through and tallying the depth distribution of each spike area in the depth image, one-dimensional grayscale histograms for the wheat spike areas are constructed, as shown in Figure 14. The histogram predominantly exhibits a unimodal distribution. The peak portion represents the wheat spike within the area, while the remaining parts indicate the non-spike areas within the detection area. The most frequently occurring grayscale value at the peak is selected as the depth value of the wheat spike. Subsequently, through the conversion equations demarcated by Equations (7) and (8), the elevation value h_i of a

single wheat plant within the world coordinate system is calculated from the spike's depth value within the camera coordinate system.

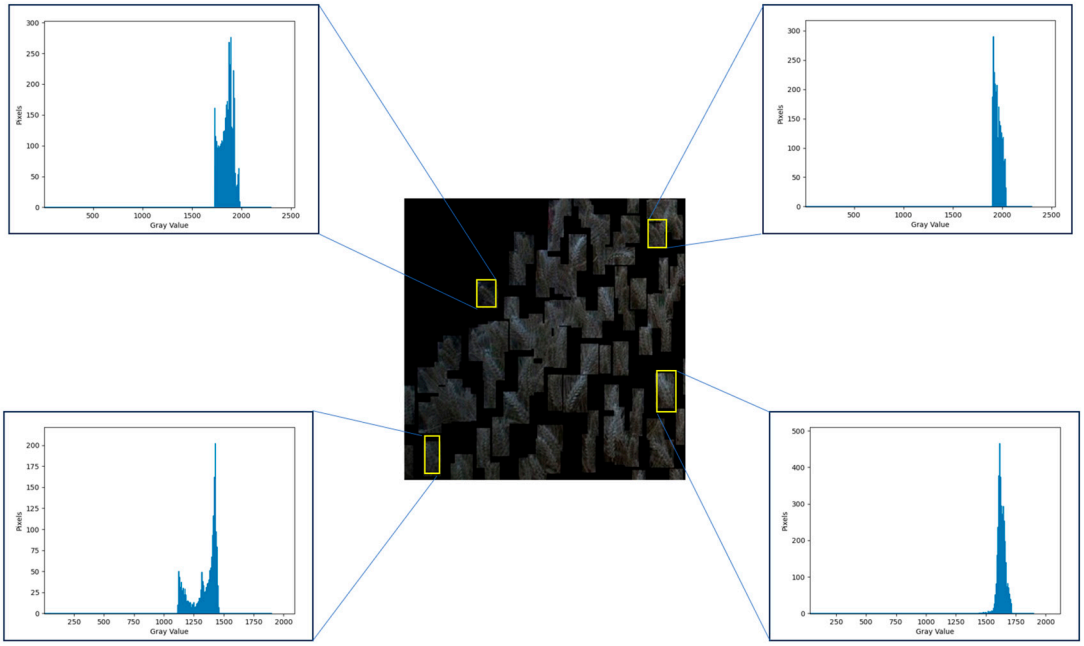


Figure 14. Depth histograms of the spike area from wheat plants.

4.3. Estimating the Height of a Single Wheat Plant without Stubble Based on Multiple Types of Ground

Accurate ground information assessment is vital for determining the height of a single wheat plant without stubble in the harvesting area. The dense growth and overlapping parts of wheat make it hard to calculate the ground height using harvester-mounted vision systems. Current methods include algorithm inversion and ground modeling. They use data from the header and adjacent harvested areas to calculate ground height in upcoming detection areas. These methods fail in places with thick residue or significant height variation along the harvest boundary. Thus, this paper examines various ground types. It determines the height of different grounds to calculate the height of a single wheat plant without stubble, considering known stubble height and plant elevation.

As illustrated in Figure 15, the types of adjacent areas are diverse, as routinely encountered in the fieldwork of the wheat combine harvester. The terrains can be broadly bifurcated into two categories: a. the harvest boundary area (which includes field ridges and cement pavements); b. the already harvested area (which encompasses scenarios of no residue accumulation, minor residue accumulation, and severe residue accumulation). This paper focuses on analyzing the ground conditions of the adjacent harvested areas and estimating the height of a single wheat plant without stubble within the harvesting area.

4.3.1. The Harvest Boundary Area

Figure 16 illustrates the scenario where the wheat combine harvester operates adjacent to cement pavements. Images captured by the camera are processed through Equations (5)–(8) to obtain the elevation images of the A2–A3–A5–A6 area. Within the elevation image, an elevation detection line A–B is established to perform an elevation profile analysis on the A2–A3–A5–A6 area. There is a distinct elevation difference between the ground area A1–A4–A5–A6 and the harvesting area A1–A2–A3–A4, indicating a significant shift in elevation values, as shown in Figure 16b. By instituting an elevation threshold, a demarcation is achieved between the ground area and the harvesting area, with results presented in Figure 16a.

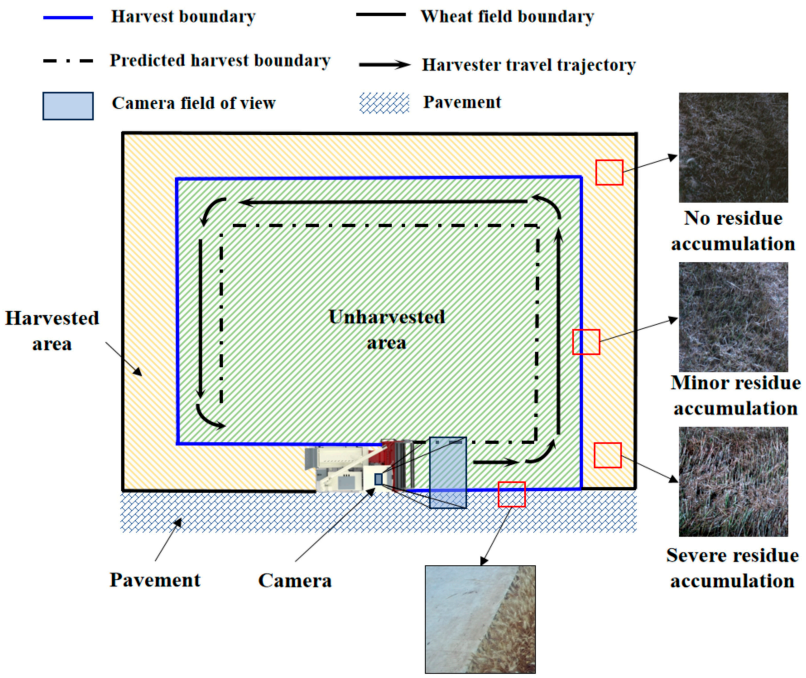
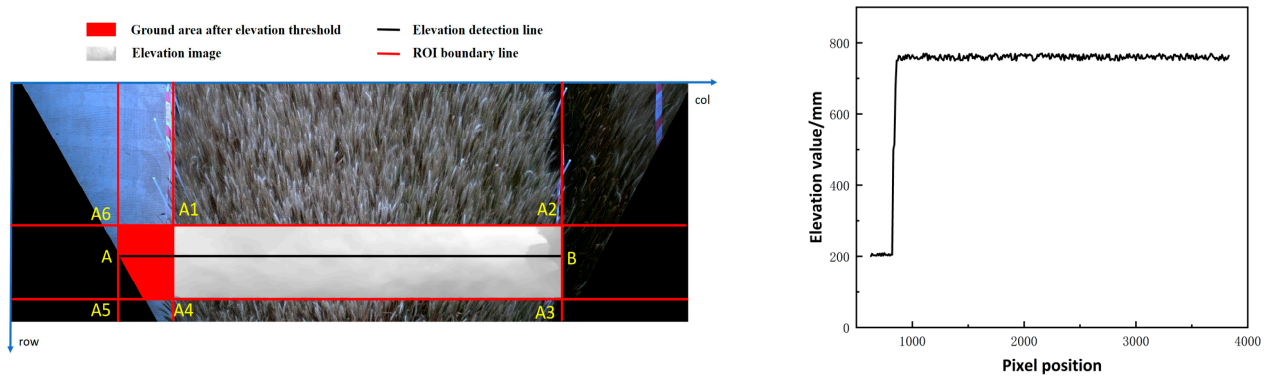


Figure 15. Schematic of wheat combine harvester operation.



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

Figure 16. Schematic of the operating area adjacent to the cement pavement. (**a**) Schematic of the operating area adjacent to the cement pavement. (**b**) Elevation profile analysis of the A2–A3–A5–A6 area.

For the operation of the wheat combine harvester operates proximal to the harvest boundary area (including field ridges and cement pavements), the boundary height h_b between the area A1–A2–A3–A4 and the adjacent ground area A1–A4–A5–A6 is detected in advance. At time t, the elevation value h_{i-t} of a single wheat plant, ground elevation h_{d-t} adjacent to the harvesting area, boundary height h_b, and wheat stubble height h_{lc} are calculated. Based on Equation (15), the height of a single wheat plant without stubble H_{i-t} within the harvesting area at time t is calculated.

$$H_{i-t} = h_{i-t} - (h_{d-t} - h_b) - h_{lc} \tag{15}$$

4.3.2. The Already Harvested Area

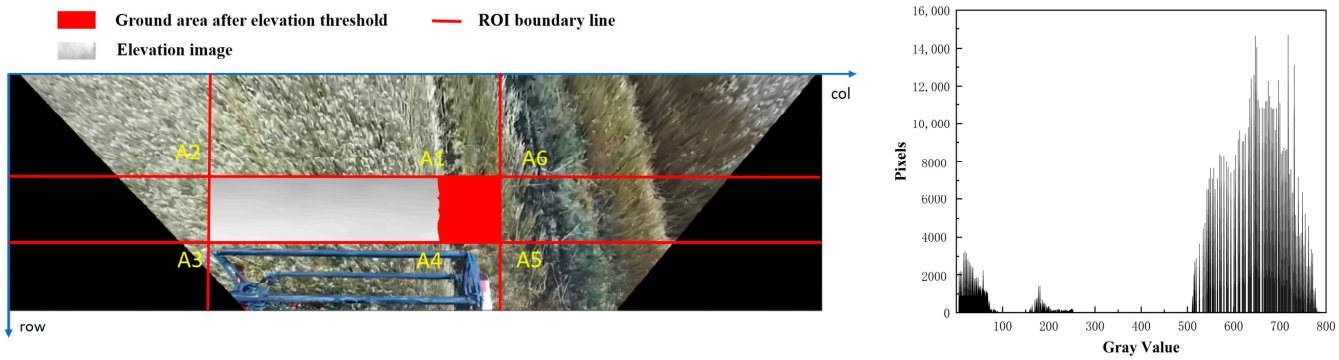
When the wheat combine harvester operates near an already harvested area, the adjacent areas mainly include three types of ground conditions: no residue accumulation, minor residue accumulation, and severe residue accumulation. The images collected by the camera are processed using Equations (5)–(8) to obtain the elevation images of A2–A3–A5–A6 area. As shown in Figure 17b, the elevation histogram of A2–A3–A5–A6 area roughly presents a three-peak distribution. The peaks correspondingly represent the

ground area, the residue accumulation area, and the harvesting area. By setting an elevation threshold, the adjacent area is distinguished from the harvesting area. The result is shown in Figure 17a. The peak values p_d and p_z of the ground area and the residue accumulation area in the elevation histogram, respectively, are used to determine the ground conditions as no residue accumulation, minor residue accumulation, and severe residue accumulation through Equations (16)–(18). The ground conditions of the adjacent area and the results of its elevation histogram are shown in Figure 18.

$$p_d > 2p_z \tag{16}$$

$$\frac{1}{2}p_z \leq p_d \leq 2p_z \tag{17}$$

$$p_d < \frac{1}{2}p_z \tag{18}$$



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

Figure 17. Schematic of the operating area adjacent to the harvested area. (**a**) Schematic of the operating area adjacent to the harvested area. (**b**) The elevation histogram of A2–A3–A5–A6 area.

In cases of the no residue accumulation type, the ground area A1–A4–A5–A6 within the FOV is visible. Images captured by the camera are processed through Equations (5)–(8) to obtain the elevation images. By analyzing the elevation histogram, the mean ground elevation value h_{d-t} at time t can be determined. By iterating through and tallying the elevation value h_{i-t} of single wheat plant within the harvesting area, the mean ground elevation value h_{d-t}, and the wheat stubble cutting height h_{lc}, the stubble-removed height H_{i-t} of a single wheat plant within the harvesting area at time t is calculated using Equation (19).

$$H_{i-t} = h_{i-t} - h_{d-t} - h_{lc} \tag{19}$$

Scenarios involving severe residue accumulation within the ground area adjacent to the already harvested area are less common in practice, and the ground information within the A1–A4–A5–A6 area at time t is difficult to detect. This paper considers using the mean ground elevation value h_{d-t-n} at time $(t-n)$, when the ground is visible, to represent the ground elevation value in the area adjacent to the already harvested area at time t in regions with slow terrain changes. n denotes the time interval between t and the nearest previous time point where the ground elevation could be accurately detected. By iterating through and tallying the elevation value h_{i-t} of single wheat plant within the harvesting area, the mean ground elevation h_{d-t-n}, and the wheat stubble height h_{lc}, the height of a single wheat plant without stubble H_{i-t} at time t is calculated using Equation (20).

$$H_{i-t} = h_{i-t} - h_{d-t-n} - h_{lc}\,, n = 1, 2, 3\ldots \tag{20}$$

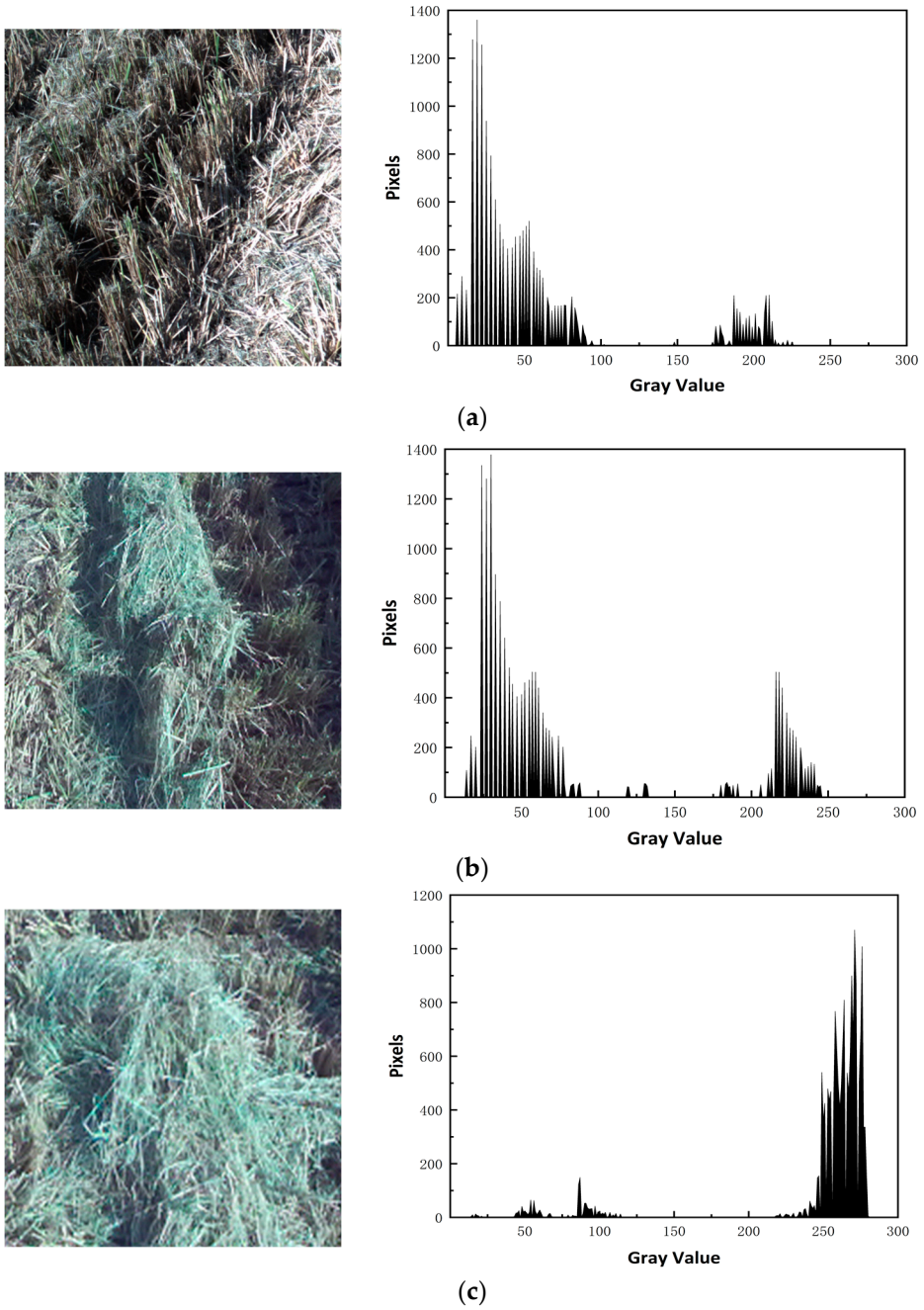


Figure 18. Analysis of the accumulation of detritus and elevation histograms. (**a**) No residue accumulation. (**b**) Minor residue accumulation. (**c**) Severe residue accumulation.

5. Prediction of Feed Quantity for Wheat Combine Harvester Based on the Weight of a Single Wheat Plant without Stubble

5.1. Height–Weight Relationship Model of Wheat Plant without Stubble

This paper defines the height from the cutter to the wheat plants' top, excluding the stubble, as the weight of a single wheat plant without stubble. The height plays a significant role in determining the weight of a single wheat plant without stubble. According to the offline experimental data, this paper constructs a height–weight relationship model of wheat plants without stubble.

As illustrated in Figure 19, concerning various mature wheat varieties and their growth conditions, five groups of experimental areas sized 2 m × 0.6 m were designated, and each experimental area was divided into six sections of 0.3 m × 0.6 m. Considering the stubble height $h_{lc} = 0.2$ m, wheat plants were manually stripped at a height of 0.2 m in

different sections. The height of a single wheat plant without stubble was measured with a tape measure. Through the cumulative analysis of data, as portrayed in Figure 20, the height–weight relationship model was formulated in accordance with Equation (21).

$$m_i = 0.0178 \times (H_i - h_{lc}) - 4.974 \quad (21)$$

where H_i represents the height of a single wheat plant without stubble, measured in meters (m); h_{lc} denotes the stubble height, also in meters (m).

Figure 19. Schematic of the experimental area.

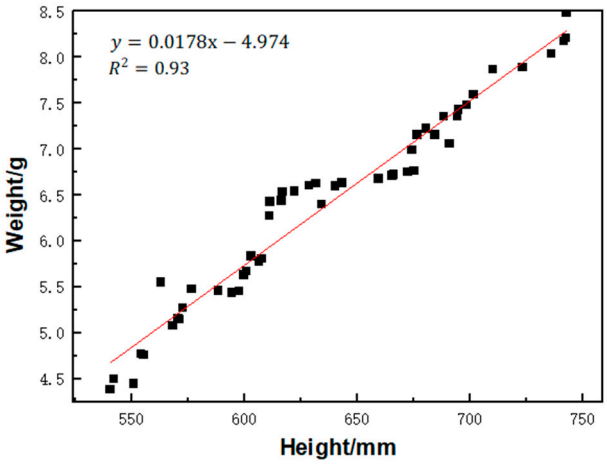


Figure 20. The height–weight relationship model.

The model exhibits a coefficient of determination (R^2) valued at 0.93, offering a quantitative foundation for predictive algorithms to estimate the weight by predicting the height of a single wheat plant without stubble.

5.2. Prediction of Feed Quantity for Wheat Combine Harvester

Based on the height H_{i-t} of a single wheat plant without stubble in the harvesting area derived from Equations (18)–(20), the height–weight relationship model from Equation (21), and the statistical number s of wheat spikes within the harvesting area, the prediction model for the feed quantity at time t was obtained, with the following Equation:

$$Q_t = \sum_{i=1}^{s} m_{i-t} = 1.667 \times \sum_{i=1}^{s} V \times [0.0178 \times (H_{i-t} - h_{lc}) - 4.974] \quad (22)$$

where V denotes the harvester's operating speed, measured in meters per second (m/s).

6. Experimental Methods and Evaluation Metrics

This paper conducts a comparative experiment on wheat spike detection using the improved YOLOv5s and other networks based on field images. It also confirms the efficiency and benefits of using the improved YOLOv5s and a feed quantity prediction model for wheat combine harvester feed quantity prediction in field trials.

6.1. Methods and Evaluation Metrics for the Comparative Experiment of the Improved YOLOv5s

Based on the VOC_Wheatear dataset, this paper conducts wheat spikes detection experiments with the existing YOLOv5s, YOLOv7, SSD, Faster R-CNN, the improved YOLOv5s, and other improved network models.

The mAP_{50}, Precision, and Recall are selected as evaluation metrics for wheat spikes detection according to Equations (23)–(25). mAP_{50} stands for mean Precision at 50% Intersection over Union (IoU) threshold, reflecting the algorithm's comprehensive classification capacity. Precision measures the proportion of true positives within the predicted positive samples, while Recall represents the proportion of true positives correctly identified by the model among all positive samples. True Positives (*TP*) are the number of correctly detected positive samples, False Positives (*FP*) are the number of negative samples incorrectly detected as positive, and False Negatives (*FN*) are the number of positive samples incorrectly classified as negative, and $p(r)$ is the Precision–Recall curve.

$$mAP_{50} = \int_0^1 p(r)dr \tag{23}$$

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

6.2. Experiment Methods and Evaluation Metrics for Wheat Combine Harvester Feed Quantity Prediction

In actual agronomic operations, experiments for predicting the feed quantity were conducted with the harvester's operating speed V = 0.6 m/s and the maximum cutting width L_g = 2 m, and the stubble height h_{lc} = 0.2 m, with a 2-s prediction rhythm. In the experimental crop fields of three wheat varieties, 35 experimental areas were randomly designated near field ridges, cement pavements, and areas with minor and severe residue accumulation, resulting in 35 sets of experimental data. The ZED 2I camera was mounted on top of the harvester's cabin with a shooting angle of 35°and a shooting height of 2.8 m, ensuring that the ground areas adjacent to the harvesting area were captured within the image frame. First, the boundaries of the 35 experimental areas were clearly marked with caution lines and markers to ensure each area measured precisely 0.6 m × 2 m. Second, the visual acquisition system described in this paper was used to collect data from each area. The wheat feed quantity for each zone was then predicted using the method outlined in this paper. Third, the harvester was operated unloaded for half a minute to clear any residual wheat grains from the grain bunker, ensuring that the data collected during the experiments was not influenced by previous contaminants. During the actual experiment, the harvester processed the wheat, and a dedicated collection device gathered the cleaning residues produced. After the experiment, the harvester was again operated unloaded for half a minute to collect any remaining residues and wheat grains from the grain bunker. Finally, the collected cleaning residues and wheat grains were manually weighed to provide the true feed quantity for each experimental set. This measurement was then used to validate the predictions made by the visual prediction system and assess the method's accuracy for this prediction rhythm.

As delineated by Equation (26), this paper defines Q_ε as the relative error in wheat feed quantity prediction, Q_{yc} as the predicted feed quantity and Q_{cl} as the true feed quantity. $\overline{Q}_\varepsilon$ denotes the mean value of relative error. N_Q denotes the number of relative error samples.

$Q_{\varepsilon i}$ denotes the ith value of relative error. σ denotes the standard deviation. This paper uses Q_ε and σ to measure the prediction accuracy of the algorithm for the wheat combine harvester feed quantity.

$$Q_\varepsilon = \frac{|Q_{yc} - Q_{cl}|}{Q_{cl}} \times 100\% \tag{26}$$

$$\overline{Q}_\varepsilon = \frac{1}{N_Q}\sum_{i=1}^{N_Q} Q_{\varepsilon i} \tag{27}$$

$$\sigma = \sqrt{\frac{1}{N_Q - 1}\sum_{i=1}^{N_Q}\left(Q_{\varepsilon i} - \overline{Q}_\varepsilon\right)^2} \tag{28}$$

7. Discussion

7.1. Comparison of the Improved YOLOv5s with Other Networks for Wheat Spikes Detection

7.1.1. Comparison of the Improved YOLOv5s with the Existing Neural Networks

Samples collected from three wheat varieties, "Zhenmai12", "Zhenmai15", and "Zhenmai18", during different periods of sunny and cloudy weather were used to construct the VOC_Wheatear dataset. Comparative experiments for wheat spikes detection were conducted using the VOC_Wheatear dataset with the improved YOLOv5s versus the existing YOLOv5s, YOLOv7, SSD, and Faster R-CNN.

Depicted in Figure 21 are the outcomes of the assay, where the original test images and the outputs derived from the existing YOLOv5s, YOLOv7, SSD, Faster R-CNN, and the improved YOLOv5s were, respectively, enumerated as numbers 0 through 5. The visual exhibit shows that, compared to other neural networks, the improved YOLOv5s has fewer false detections of wheat spikes in ground areas and harvesting boundary regions under cloudy and low-light conditions. In sunny conditions, the improved YOLOv5s pays more attention to wheat spikes at the image edges and small dense spikes, with fewer omissions.

By applying Equations (23)–(25), the detection results for the existing YOLOv5s, YOLOv7, SSD, and Faster R-CNN on the VOC_Wheatear test set were compiled. These results were then compared with those from the improved YOLOv5s for wheat spikes detection, as presented in Table 1.

Table 1. Comparison of wheat spikes detection results between the improved YOLOv5s and the existing neural networks.

No.	Detection Method	mAP_{50}/%	Pre/%	Recall/%	Time/ms
1	Existing YOLOv5s	71.3	80.3	68.2	94.5
2	YOLOv7	65.3	72.6	60.4	109.4
3	SSD	68.5	75.6	67.5	98.6
4	Faster R-CNN	70.5	80.6	65.2	116.3
5	Improved YOLOv5s	78.1	85.2	70.9	101.7

The following data have been obtained from the table: Compared to the existing YOLOv5s, YOLOv7, SSD, and Faster R-CNN, the improved YOLOv5s achieved a 6.8%, 12.8%, 9.6%, and 7.6% increase in mAP_{50} for wheat spikes detection on the VOC_Wheatear test set, respectively. The Precision of wheat spikes detection was improved by 4.9%, 12.6%, 9.6%, and 4.6%. The Recall increased by 2.7%, 10.5%, 3.4%, and 5.7%, respectively. Although the average processing time of the improved YOLOv5s for wheat spikes detection increased by 7.2 ms and 3.1 ms compared to the existing YOLOv5s and SSD, the increase was marginal and still within the real-time detection and control requirements of the wheat combine harvester's closed-loop control time rhythm.

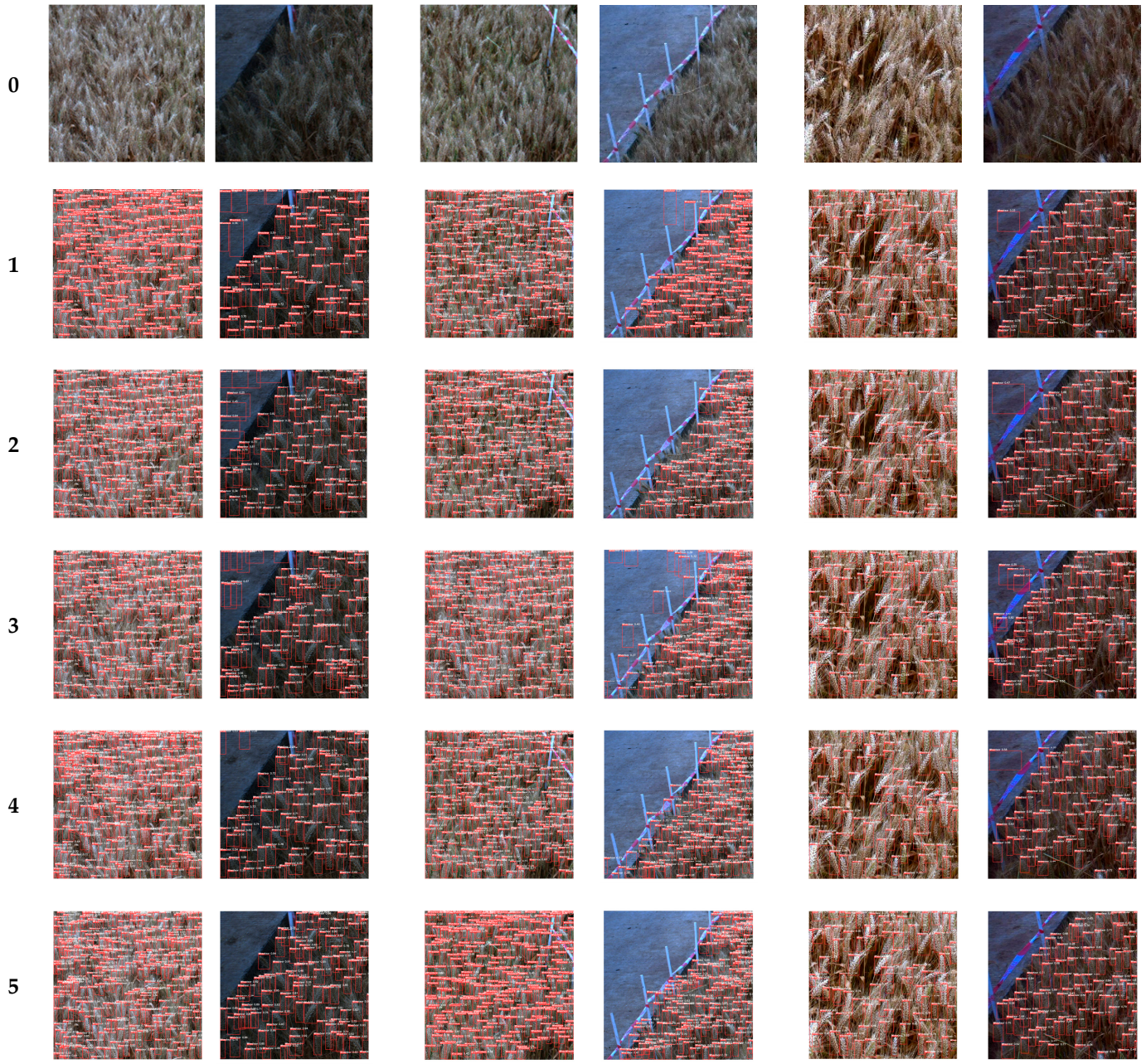


Figure 21. Comparison of wheat spikes detection results between the improved YOLOv5s and the existing neural networks.

7.1.2. Comparison of YOLOv5s Improvements

After conducting comparative experiments on various existing network models, this paper chose the existing YOLOv5s as the baseline. The paper also trained and tested some improved models generated during the process of network structure improvement, comparing the existing YOLOv5s with these improved methods and the improved YOLOv5s. The improvements made to the existing YOLOv5s focus on the Backbone, Neck, and Head structures. The enhancements include the C3 module, attention mechanism module, and feature map scales. The methodologies involve a comparative analysis of the accuracy and efficiency of wheat spikes detection for different improvements. Within this exegesis, "Existing YOLOv5s" refers to the current model; "YOLOv5s + C3Res2NetBlock" denotes the replacement of the

C3Res2NetBlock in the existing YOLOv5s Neck structure; "YOLOv5s + EMA" indicates the addition of EMA in the Backbone structure of the existing YOLOv5s; "YOLOv5s + P2–P5" refers to the substitution of the 160 × 160 P2 detection layer in place of the 20 × 20 P5 detection layer in the Head structure of the existing YOLOv5s; "YOLOv5s + P2–P5 + C3Res2NetBlock" signifies the replacement of the C3Res2NetBlock in the Neck structure and the use of the P2 detection layer instead of the P5 detection layer in the Head structure; "Improved YOLOv5s" points to the comprehensive improvements including the replacement of C3Res2NetBlock in the Neck structure, addition of EMA in the Backbone, and the use of the P2 detection layer in the Head structure. Designations A through F correspondingly represent "Existing YOLOv5s", "YOLOv5s + C3Res2NetBlock", "YOLOv5s + EMA", "YOLOv5s + P2–P5", "YOLOv5s + P2–P5 + C3Res2NetBlock", and "Improved YOLOv5s".

Employing a method predicated on transfer learning, this paper aims to expedite the training regimen, bolster model generalization, and curtail the propensity for overfitting. The improved YOLOv5s underwent pretraining upon the COCO dataset, acquiring pretrained weights in the interim. Then, we used the pretrained weights for further training on the VOC_Wheatear dataset to learn the object features of wheat spikes. The comparative results of the training are depicted in Figure 22, with "epoch" denoting the number of training cycles. The illustration reveals that, in comparison with other improvements and the existing YOLOv5s, the improved YOLOv5s yields the fastest convergence speed and the highest mAP_{50} during dataset training. The other modifications demonstrate quicker convergence and relatively greater mAP_{50} than the existing YOLOv5s.

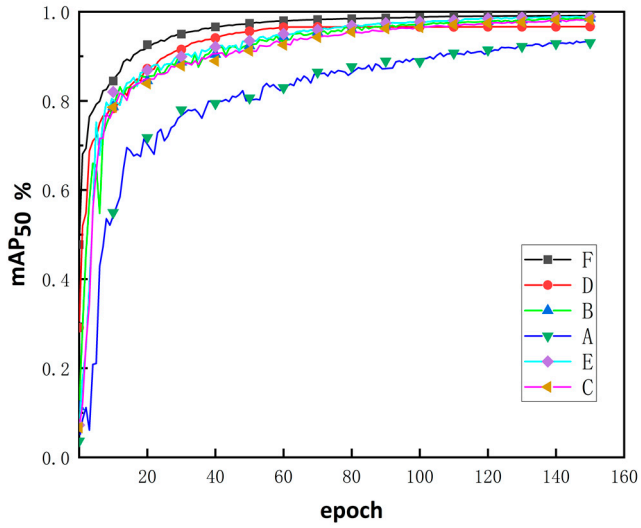


Figure 22. Training comparison of the improved YOLOv5s.

To further provide an intuitive analysis of the comparison in detection performance of YOLOv5s structural modifications, visualized Gradient-weighted Class Activation Mapping (Grad-CAM) [44] heatmaps of some training images are displayed in Figure 23. From the figure, it is evident that in comparison with other improvement methods and the existing YOLOv5s, the improved YOLOv5s exhibits the highest focus on areas containing wheat spikes in the training images, possesses the best generalization capabilities, and affords the most precise identification and localization of small wheat spikes. Relative to the existing YOLOv5s, the alternative improvements also display a considerably higher focus on areas with wheat spikes in the training images and exhibit more accurate identification and positioning of the small wheat spikes.

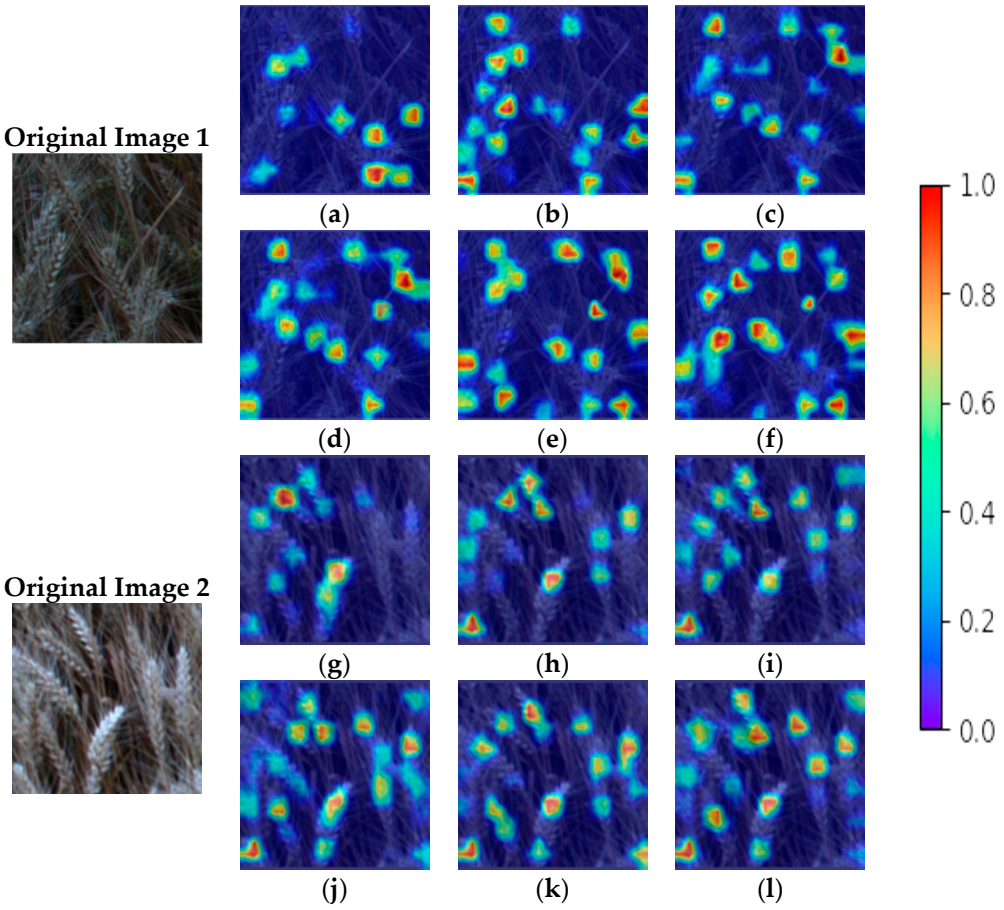


Figure 23. Comparison of heat maps for YOLOv5s structural modification. (**a**) Heat map of Original Image 1 from "Existing YOLOv5s". (**b**) Heat map of Original Image 1 from "YOLOv5s + C3Res2NetBlock". (**c**) Heat map of Original Image 1 from "YOLOv5s + EMA". (**d**) Heat map of Original Image 1 from "YOLOv5s + P2–P5". (**e**) Heat map of Original Image 1 from "YOLOv5s + P2–P5 + C3Res2NetBlock". (**f**) Heat map of Original Image 1 from "Improved YOLOv5s". (**g**) Heat map of Original Image 2 from "Existing YOLOv5s". (**h**) Heat map of Original Image 2 from "YOLOv5s + C3Res2NetBlock". (**i**) Heat map of Original Image 2 from "YOLOv5s + EMA". (**j**) Heat map of Original Image 2 from "YOLOv5s + P2–P5". (**k**) Heat map of Original Image 2 from "YOLOv5s + P2–P5 + C3Res2NetBlock". (**l**) Heat map of Original Image 2 from "Improved YOLOv5s".

By applying Equations (23)–(25), the detection results for the existing YOLOv5s, "YOLOv5s + C3Res2NetBlock", "YOLOv5s + EMA", "YOLOv5s + P2–P5", "YOLOv5s + P2-P5 + C3Res2NetBlock" on the VOC_Wheatear test set were compiled. These results were then compared with those from the improved YOLOv5s for wheat spikes detection, as presented in Table 2.

From the table, it is apparent that models with structural modifications based on the existing YOLOv5s demonstrate improvements in mAP_{50}, Precision, and Recall for the wheat spikes test set over the existing YOLOv5s. Compared to the existing YOLOv5s, "YOLOv5s + C3Res2NetBlock", "YOLOv5s + EMA", "YOLOv5s + P2–P5", and "YOLOv5s + P2–P5 + C3Res2NetBlock", the improved YOLOv5s exhibited increments in mAP_{50} of 6.8%, 10%, 9%, 7%, and 7%, respectively; and in Precision of 4.9%, 0.1%, −0.2%, 0.2%, and 0.6%. Relative to other improvement methods and the existing YOLOv5s, the average processing time of the improved YOLOv5s for wheat spikes detection increased slightly but remained within the operational time rhythm of feed quantity prediction, satisfying the real-time detection and control requirements of wheat combine harvester.

Table 2. Comparison of test results of YOLOv5s model structure modification.

No.	Detection Method	mAP_{50}/%	Pre/%	Recall/%	Time/$_{ms}$
A *	Existing YOLOv5s	71.3	80.3	68.2	94.5
B *	YOLOv5s + C3Res2NetBlock	77.1	85.1	71.9	87.5
C *	YOLOv5s + EMA	77.2	85.4	71.8	96.1
D *	YOLOv5s + P2–P5	77.4	85.0	72.1	97.5
E *	YOLOv5s + P2–P5 + C3Res2NetBlock	77.4	84.6	70.3	98.3
F *	Improved YOLOv5s	78.1	85.2	70.9	101.7

* Designations A through F correspondingly represent "Existing YOLOv5s", "YOLOv5s + C3Res2NetBlock", "YOLOv5s + EMA", "YOLOv5s + P2–P5", "YOLOv5s + P2–P5 + C3Res2NetBlock", and "Improved YOLOv5s".

7.2. Wheat Combine Harvester Feed Quantity Prediction

Utilizing Equations (15)–(22), each experimental group used the improved YOLOv5s to calculate the number of wheat plants, the height–weight relationship model, the height of a single wheat plant without stubble in the harvesting area, and the wheat combine harvester feed quantity prediction model. Comparison of predicted value and true value for the combine harvester's feed quantity were made for 35 datasets, as shown in Figure 24. The prediction time includes image processing, wheat spikes detection, estimation of the height of a single wheat plant without stubble, and feed quantity prediction.

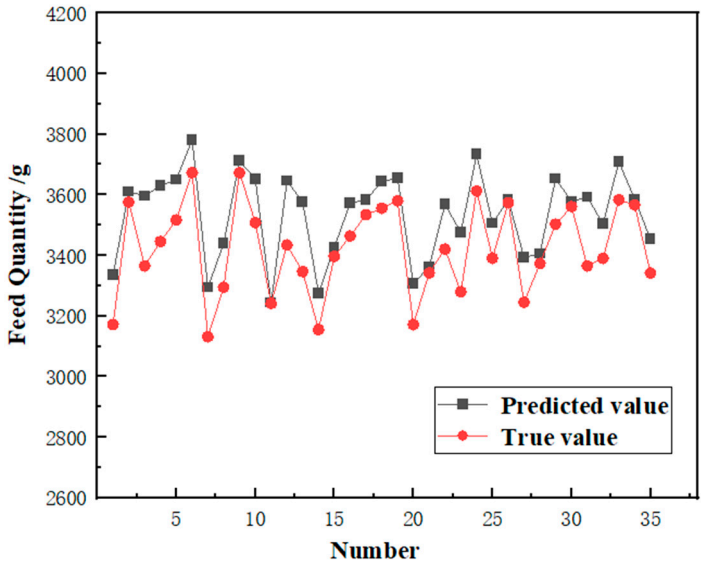


Figure 24. Comparison of predicted value and true value of feed quantity.

In accordance with Equations (26)–(28), the relative errors for feed quantity predictions in Figure 24 were computed, and the outcomes are displayed in Figure 25. The relative error in predicting the feed quantity ranged between 1.08% and 7.42%, with an average of 4.19% and a standard deviation of 1.904%. The average prediction time was 1.34 s, which conformed to the closed-loop control threshold of the harvester. The outcomes of these experiments verified the effectiveness and advantages of the method for predicting the feed quantity derived from the improved YOLOv5s and the weight of a single wheat plant without stubble.

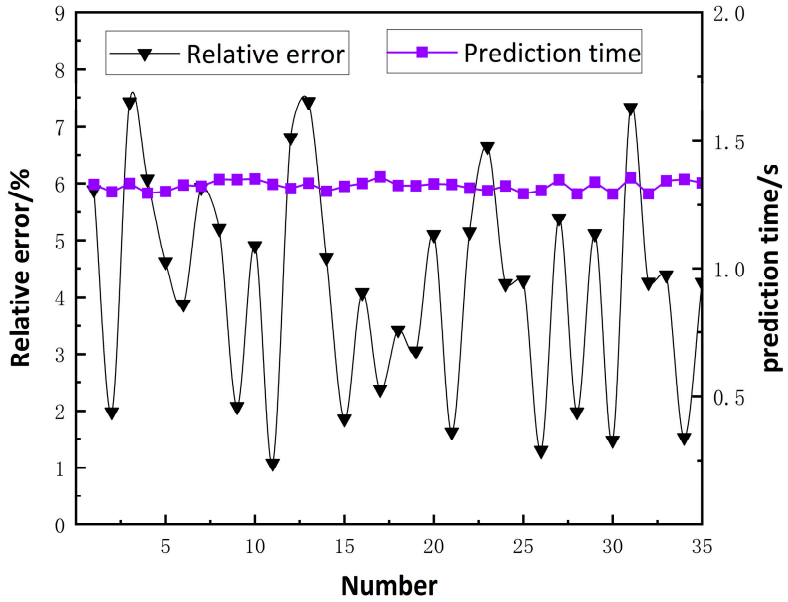


Figure 25. Prediction time and relative errors of wheat combine harvester feed quantity prediction.

8. Conclusions

Wheat grows densely with overlapping organs and different weights of a single plant in a complex field environment. It is difficult to predict the feed quantity accurately for a wheat combine harvester based on the existing YOLOv5s and the uniform weight of a single wheat plant for a whole field. This paper improves the existing YOLOv5s based on the multi-scale features of small objects and attention optimization. In addition, we proposed a wheat combine harvester feed quantity prediction method based on the number of wheat plants and the weight of a single wheat plant without stubble. The main conclusions are as follows:

(1) An optimization of the attention mechanism based on a set of compact bases was proposed for the Backbone structure. The existing YOLOv5s Backbone structure, which lacks attention optimization, was strategically enhanced through the integration of an EMA mechanism alongside Dropout layers. This enhancement boosted the attentiveness towards the distinctive features and reduced computational redundancy. A multi-scale feature extraction C3Res2NetBlock module with a hierarchical residual structure was integrated into the existing YOLOv5s Neck structure. This module facilitated an enhanced resolution in the extraction of multi-scale features of wheat spikes while reducing the network's parameter framework and computational expenditure. An improved Head architecture focused on small targets was delineated. This remodeled Head employed larger-scale detection layers to replace the original detection layers. It improved the recognition accuracy of small dense wheat spikes in large FOV and reduced the missed detection.

(2) Based on the wheat spikes detection results from the improved YOLOv5s, the depth distribution and elevation value of a single wheat plant were calculated. This paper examined various ground types in the harvesting area. It determined the height of these grounds, allowing for the estimation of the height of a single wheat plant without stubble. This estimation used known stubble height and wheat plant elevation. Combining the statistical count of wheat plants and the height of a single wheat plant without stubble from the improved YOLOv5s, a feed quantity prediction model was established. In addition, we proposed a wheat combine harvester feed quantity prediction method based on the number of wheat plants and the weight of a single wheat plant without stubble.

(3) The proposed method was verified through experiments with images acquired on the 4LZ-6A intelligent combine harvester. Compared with the existing YOLOv5s,

YOLOv7, SSD, and Faster R-CNN, the mAP_{50} of wheat spikes detection by the improved YOLOv5s increased by over 6.8%. Compared with the improved YOLOv5s in other ways in this paper, the mAP_{50} of wheat spikes detection by the improved YOLOv5s increased by over 6.8%. The average relative error of feed quantity prediction based on the proposed method was 4.19%. The average time of prediction using the proposed method was 1.34 s. The proposed method can accurately and rapidly predict the feed quantity of wheat combine harvester and further realize closed-loop control of intelligent harvesting operations.

Author Contributions: Conceptualization, Q.Z. and Q.C.; Data curation, Q.C. and W.X.; Formal analysis, Q.Z. and W.X.; Funding acquisition, Q.Z. and L.X.; Investigation, Q.C., E.L. and W.X.; Methodology, Q.Z. and Q.C.; Project administration, Q.Z. and L.X.; Resources, Q.Z.; Software, Q.Z., Q.C. and W.X.; Supervision, L.X.; Validation, Q.C. and E.L.; Visualization, Q.Z. and Q.C.; Writing—original draft, Q.Z. and Q.C.; Writing—review and editing, Q.Z. and Q.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 52302495; High-tech Key Laboratory of Agricultural Equipment and Intelligence of Jiangsu Province, grant number MAET202329; Jiangsu Province Higher Education Basic Science (Natural Science) Research Project, grant number 23KJB210006; Zhenjiang Key R&D Plan (Industry Foresight and Common Key Technology) Project, grant number GY2023001; Jiangsu Agricultural Science and Technology Innovation Fund, grant number CX(22)1005.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Acknowledgments: Thanks to the authors cited in this article and the referees for their helpful comments and suggestions.

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

References

1. Wang, F.; Liu, Y.; Li, Y.; Ji, K. Research and Experiment on Variable-Diameter Threshing Drum with Movable Radial Plates for Combine Harvester. *Agriculture* **2023**, *13*, 1487. [CrossRef]
2. Shi, J.; Jiang, M.; Zhao, Y.; Liao, N.; Wang, Z. Research on the Fault-Diagnosing Method in the Operation of the Threshing Cylinder of the Combine Harvester. In Proceedings of the 2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA), Chengdu, China, 1–4 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1279–1284.
3. Hao, S.; Tang, Z.; Guo, S.; Ding, Z.; Su, Z. Model and Method of Fault Signal Diagnosis for Blockage and Slippage of Rice Threshing Drum. *Agriculture* **2022**, *12*, 1968. [CrossRef]
4. Liang, Z.; Wada, M.E. Development of cleaning systems for combine harvesters: A review. *Biosyst. Eng.* **2023**, *236*, 79–102. [CrossRef]
5. Yu, W.; Xin, W.; Jiangjiang, Z.; Dong, W.; Shumao, W. Wireless feeding rate real-time monitoring system of combine harvester. In Proceedings of the 2017 Electronics, Palanga, Lithuania, 19–21 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
6. Zhang, Y.; Chen, D.; Yin, Y.; Wang, X.; Wang, S. Experimental study of feed rate related factors of combine harvester based on grey correlation. *IFAC-PapersOnLine* **2018**, *51*, 402–407. [CrossRef]
7. Chen, X.; He, X.; Wang, W.; Qu, Z.; Liu, Y. Study on the Technologies of Loss Reduction in Wheat Mechanization Harvesting: A Review. *Agriculture* **2022**, *12*, 1935. [CrossRef]
8. Liang, Z.; Qin, Y.; Su, Z. Establishment of a Feeding Rate Prediction Model for Combine Harvesters. *Agriculture* **2024**, *14*, 589. [CrossRef]
9. Chen, M.; Jin, C.; Ni, Y.; Yang, T.; Zhang, G. Online field performance evaluation system of a grain combine harvester. *Comput. Electron. Agric.* **2022**, *198*, 107047. [CrossRef]
10. Kanning, M.; Kühling, I.; Trautz, D.; Jarmer, T. High-resolution UAV-based hyperspectral imagery for LAI and chlorophyll estimations from wheat for yield prediction. *Remote Sens.* **2018**, *10*, 2000. [CrossRef]
11. Kim, Y.; Jackson, T.; Bindlish, R.; Hong, S.; Jung, G.; Lee, K. Retrieval of wheat growth parameters with radar vegetation indices. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 808–812.
12. Chen, J.; Fu, S.; Wang, Z.; Zhu, L.; Xia, H. Research on the method of predicting feeding volume of rice combine harvester base on machine vision. In Proceedings of the International Conference on Image Processing and Intelligent Control (IPIC 2021), Lanzhou, China, 30 July–1 August 2021; SPIE: Bellingham, WA, USA, 2021; pp. 28–32.

13. Olson, D.; Anderson, J. Review on unmanned aerial vehicles, remote sensors, imagery processing, and their applications in agriculture. *Agron. J.* **2021**, *113*, 971–992. [CrossRef]
14. Zhu, W.; Feng, Z.; Dai, S.; Zhang, P.; Wei, X. Using UAV multispectral remote sensing with appropriate spatial resolution and machine learning to monitor wheat scab. *Agriculture* **2022**, *12*, 1785. [CrossRef]
15. Xu, S.; Xu, X.; Zhu, Q.; Meng, Y.; Yang, G.; Feng, H.; Yang, M.; Zhu, Q.; Xue, H.; Wang, B. Monitoring leaf nitrogen content in rice based on information fusion of multi-sensor imagery from UAV. *Precis. Agric.* **2023**, *24*, 2327–2349. [CrossRef]
16. Wei, L.; Yang, H.; Niu, Y.; Zhang, Y.; Xu, L.; Chai, X. Wheat biomass, yield, and straw-grain ratio estimation from multi-temporal UAV-based RGB and multispectral images. *Biosyst. Eng.* **2023**, *234*, 187–205. [CrossRef]
17. Shi, Q.; Liu, D.; Mao, H.; Shen, B.; Li, M. Wind-induced response of rice under the action of the downwash flow field of a multi-rotor UAV. *Biosyst. Eng.* **2021**, *203*, 60–69. [CrossRef]
18. Chen, J.; Lian, Y.; Zou, R.; Zhang, S.; Ning, X.; Han, M. Real-time grain breakage sensing for rice combine harvesters using machine vision technology. *Int. J. Agric. Biol. Eng.* **2020**, *13*, 194–199. [CrossRef]
19. Zhang, Q.; Chen, Q.; Xu, L.; Xu, X.; Liang, Z. Wheat Lodging Direction Detection for Combine Harvesters Based on Improved K-Means and Bag of Visual Words. *Agronomy* **2023**, *13*, 2227. [CrossRef]
20. Wen, J.; Yin, Y.; Zhang, Y.; Pan, Z.; Fan, Y. Detection of wheat lodging by binocular cameras during harvesting operation. *Agriculture* **2022**, *13*, 120. [CrossRef]
21. Maji, A.K.; Marwaha, S.; Kumar, S.; Arora, A.; Chinnusamy, V.; Islam, S. SlypNet: Spikelet-based yield prediction of wheat using advanced plant phenotyping and computer vision techniques. *Front. Plant Sci.* **2022**, *13*, 889853. [CrossRef] [PubMed]
22. Wang, M.; Liu, X.; Gao, Y.; Ma, X.; Soomro, N.Q. Superpixel segmentation: A benchmark. *Signal Process. Image Commun.* **2017**, *56*, 28–39. [CrossRef]
23. Au, O.K.-C.; Tai, C.-L.; Chu, H.-K.; Cohen-Or, D.; Lee, T.-Y. Skeleton extraction by mesh contraction. *ACM Trans. Graph. (TOG)* **2008**, *27*, 1–10. [CrossRef]
24. Fabricius, A.M.; Diegeler, A.; Doll, N.; Weidenbach, H.; Mohr, F.W. Minimally invasive saphenous vein harvesting techniques: Morphology and postoperative outcome. *Ann. Thorac. Surg.* **2000**, *70*, 473–478. [CrossRef] [PubMed]
25. Kornilov, A.S.; Safonov, I.V. An overview of watershed algorithm implementations in open source libraries. *J. Imaging* **2018**, *4*, 123. [CrossRef]
26. Chaganti, S.Y.; Nanda, I.; Pandi, K.R.; Prudhvith, T.G.; Kumar, N. Image Classification using SVM and CNN. In Proceedings of the 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 13–14 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5.
27. 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: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
28. Zhang, Q.; Gao, G. Prioritizing robotic grasping of stacked fruit clusters based on stalk location in RGB-D images. *Comput. Electron. Agric.* **2020**, *172*, 105359. [CrossRef]
29. Ji, W.; Wang, J.; Xu, B.; Zhang, T. Apple Grading Based on Multi-Dimensional View Processing and Deep Learning. *Foods* **2023**, *12*, 2117. [CrossRef] [PubMed]
30. Wang, D.; He, D. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. *Biosyst. Eng.* **2021**, *210*, 271–281. [CrossRef]
31. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
32. Chirarattananon, P. A direct optic flow-based strategy for inverse flight altitude estimation with monocular vision and IMU measurements. *Bioinspir. Biomim.* **2018**, *13*, 036004. [CrossRef]
33. Zhao, L.; Jiao, S.; Wang, C.; Zhang, J. Research on terrain sensing method and model prediction for height adjustment of sugarcane harvester base cutter. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 7344498. [CrossRef]
34. Sun, Y.; Luo, Y.; Zhang, Q.; Xu, L.; Wang, L.; Zhang, P. Estimation of crop height distribution for mature rice based on a moving surface and 3D point cloud elevation. *Agronomy* **2022**, *12*, 836. [CrossRef]
35. Zhang, Q.; Gao, G.-Q. Hand–eye calibration and grasping pose calculation with motion error compensation and vertical-component correction for 4-R (2-SS) parallel robot. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420909012. [CrossRef]
36. Luo, Y.; Wei, L.; Xu, L.; Zhang, Q.; Liu, J.; Cai, Q.; Zhang, W. Stereo-vision-based multi-crop harvesting edge detection for precise automatic steering of combine harvester. *Biosyst. Eng.* **2022**, *215*, 115–128. [CrossRef]
37. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; Fang, J.; Wong, C.; Yifu, Z.; Montes, D. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo* **2022**. [CrossRef]
38. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
39. Zhuang, X.; Li, Y. Segmentation and Angle Calculation of Rice Lodging during Harvesting by a Combine Harvester. *Agriculture* **2023**, *13*, 1425. [CrossRef]

40. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9167–9176.
41. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]
42. Torrey, L.; Shavlik, J. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; IGI Global: Hershey, PA, USA, 2010; pp. 242–264.
43. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13. Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
44. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.