agronomy

MDPI

Article

Cross-Platform Wheat Ear Counting Model Using Deep Learning for UAV and Ground Systems

Baohua Yang *, Ming Pan, Zhiwei Gao, Hongbo Zhi and Xiangxuan Zhang

School of Information and Computer, Anhui Agricultural University, Hefei 230036, China
* Correspondence: ybh@ahau.edu.cn

Abstract: Wheat is one of the widely cultivated crops. Accurate and efficient high-throughput ear counting is important for wheat production, yield evaluation, and seed breeding. The traditional wheat ear counting method is inefficient due to the small scope of investigation. Especially in the wheat field scene, the images obtained from different platforms, including ground systems and unmanned aerial vehicles (UAVs), have differences in density, scale, and wheat ear distribution, which makes the wheat ear counting task still face some challenges. To this end, a density map counting network (LWDNet) model was constructed for cross-platform wheat ear statistics. Firstly, CA-MobileNetV3 was constructed by introducing a collaborative attention mechanism (CA) to optimize the lightweight neural network MobileNetV3, which was used as the front end of the feature extraction network, aiming to solve the problem of occlusion and adhesion of wheat ears in the field. Secondly, to enhance the model's ability to learn the detailed features of wheat ears, the CARAFE upsampling module was introduced in the feature fusion layer to better restore the characteristics of wheat ears and improve the counting accuracy of the model for wheat ears. Finally, density map regression was used to achieve high-density, small-target ear counting, and the model was tested on datasets from different platforms. The results showed that our method can efficiently count wheat ears of different spatial scales, achieving good accuracy while maintaining a competitive number of parameters (2.38 million with a size of 9.24 MB), which will benefit wheat breeding and screening analysis to provide technical support.

Keywords: UAVs; wheat ear counting; density map; lightweight

Citation: Yang, B.; Pan, M.; Gao, Z.; Zhi, H.; Zhang, X. Cross-Platform Wheat Ear Counting Model Using Deep Learning for UAV and Ground Systems. *Agronomy* **2023**, *13*, 1792. https://doi.org/10.3390/agronomy13071792

Academic Editors: Ahmed Rady and Ahmed Kayad

Received: 11 May 2023
Revised: 7 June 2023
Accepted: 13 June 2023
Published: 4 July 2023

1. Introduction

Wheat has the largest sown area, the largest yield, and is the most widely distributed food crop in the world. The number of ears of wheat is one of the most critical parameters in wheat phenotyping research, which is the key evaluation parameter for wheat planting density and yield estimation. How to count wheat ears accurately, efficiently, and non-destructively is of great significance to phenotype analysis, field management, and food security [1]. However, the wheat ears are not only similar in color to the leaves but also densely distributed. Only relying on the traditional manual counting method has disadvantages, such as strong subjectivity, low efficiency, large error, long counting cycle, and lack of unified statistical standards [2]. Therefore, the counting method of wheat ears is still an important issue in current research.

In fact, with the development of high-throughput phenotype acquisition platforms, RGB images can be acquired not only by hand-held cameras and ground devices but also by convenient drone platforms. Among them, images acquired by ground-based phenotyping platforms have a high spatial resolution, which has been applied to crop plant counting by some scholars. The UAV remote sensing platform has the advantages of miniaturization, practicability, and high resolution, which has also provided a new way for large-scale field crop counting. In the past ten years, detection and counting methods based on deep

learning, including single-stage target detection network models, such as FCN [3], Faster-RCNN [4], etc., and two-stage target detection network models, such as EfficientDet-D0 [5], YOLO v3 [6], YOLO v4 [7], YOLO v5 [8,9], etc., were used to generate multiple location boxes by detecting the target area of wheat ears, thus realizing the counting of wheat ears in the field. The above studies show that the identification and counting of wheat ears in the field environment based on the convolutional neural network model can not only make up for the subjectivity, experience, and unsustainability of manual detection but also provide higher detection accuracy. However, there are more practical problems to be overcome when counting wheat ears in high-density wheat areas in crowded field scenes. In particular, wheat in the field grows densely and has smaller ears. The method based on small target detection still needs to be further improved. On the one hand, the detector needs to be trained to capture more information, resulting in a complex structure of the training detector and a large amount of calculation. On the other hand, after multiple downsampling in the deep CNN architecture, the deep feature maps will lose more spatial information, resulting in a decrease in counting accuracy [10]. Therefore, the difficulty of the above target detection limited the further improvement of the accuracy and efficiency of counting the number of ears of wheat in large-scale and intensive fields.

Currently, methods based on density maps, which skip the recognition and classification tasks and directly generate density maps, have been favored in the study of dense object counting [11]. Specifically, the method is to learn the mapping relationship between the local key features of the wheat image and the density map, thereby obtaining the count of the target object based on the integral of the density map [12]. The research showed that the method of constructing density map regression improved the accuracy of wheat ear counting in the real scene of wheat field. For example, Ma et al. proposed an EarDensityNet model based on a fully convolutional neural network to generate a wheat canopy density map to achieve satisfactory ear counting results, which solved the problem of mutual occlusion between wheat ears [13]. Therefore, counting wheat ears using a density map model has a positive effect on reducing labor intensity and time consumption. Although the above-mentioned method successfully implements the count, the design of the counting model needs to consider not only the accuracy but also the complexity and size of the model. On the one hand, to facilitate the producers and breeders to better realize the high-throughput phenotype analysis of wheat, it is necessary to reduce the number of network parameters as much as possible without reducing the accuracy. For example, Khaki et al. (2022) designed a lightweight neural network to successfully implement the counting of wheat ears based on public datasets [14]. On the other hand, in addition to ground-based field crop counts, images obtained from UAV platforms were often used for field high-throughput monitoring due to their convenience, high speed, and low cost. Studies have shown that when the vertical heights from the wheat canopy are 2.5 m [4], 2.9 m [15], 3.5 m [16], 5 m [17], and 20–30 m [18], the spikes were successfully detected. In addition, wheat ear counting can also be achieved through UAV platforms based on different heights, such as 0.5–1.2 m [19] and 25 m [20]. Although high-throughput phenotypic data collection has become a reality, most of the previous studies were based on building models for phenotypic acquisition on a single platform, and few studies involved counting wheat ears across different platforms [21]. Therefore, it is necessary to develop a high-precision model to meet the difficulty of counting small wheat ears for different spatial scales, which is of significance for wheat field monitoring and wheat yield estimation in application scenarios of different spatial scales.

Therefore, a cross-platform wheat ear counting model based on deep learning of UAV and ground system was constructed in this study. The main contributions were to (1) improve the lightweight neural network using the collaborative attention mechanism to solve the problems of dense distribution, overlapping, severe occlusion, and adhesion in the images of wheat ears obtained at different spatial scales, aiming to improve the robustness of cross-platform ear counting models, (2) optimize the feature fusion layer with the CARAFE upsampling module, aiming to improve the feature extraction ability

of small target wheat ear details on different platforms, (3) build a lightweight wheat ear density map counting network, which reduces the number of learning parameters of the model, making the model easy to be deployed in field applications on ground systems or UAV platforms.

2. Materials and Methods

2.1. Data Collection

2.1.1. Data Sources

The data collection area is located at the National Agricultural Science and Technology Innovation and Integration Demonstration (31°25′~31°42′ N, 117°09′~117°16′ E). The area belongs to the subtropical monsoon climate, and the agricultural climate resources are relatively superior. Winter wheat is the main food crop in the area, and the main camera (48 megapixels) of a nova5pro mobile phone (Huawei Technologies Co., Ltd., Shenzhen, China) was used for data acquisition on 7 May and 17 May 2021. The camera lens was kept at a distance of about 50 cm from the wheat ears, and a total of 500 images of winter wheat ears were collected, with an image pixel resolution of 3024 × 3024 pixels. These images contain many samples of different light intensities, different densities, and different periods, including 313 for filling and 187 for maturity. Our team members manually counted the number of ears of wheat while taking pictures and recorded it as a wheat ear counting (WEC) data set.

In addition, to avoid the problem of overfitting the training model due to the single feature of wheat images, it is necessary to obtain diverse and representative wheat images. We added another 560 wheat images from GWHD with a resolution of 1024 × 1024 pixels. To reflect the diversity of the data, 140 wheat images from four continents in the GWHD dataset were selected, including Asia, Europe, North America, and Oceania, as shown in Figure 1, to verify the transferability and universality of the model, whether it can effectively count wheat ears in a denser wheat scene. Wheat images were obtained on 8 May and 26 May 2022 at a height of about 3–5 m from the top of the wheat ears using UAV (unmanned aerial vehicle) 3P (SZ DJI Technology Co., Shenzhen, China) at the High-tech Agricultural Park in Hefei, China, as the UAV-WE (UAV-based wheat ears counting) dataset, which was used to test the models individually, as shown in Figure 1c.

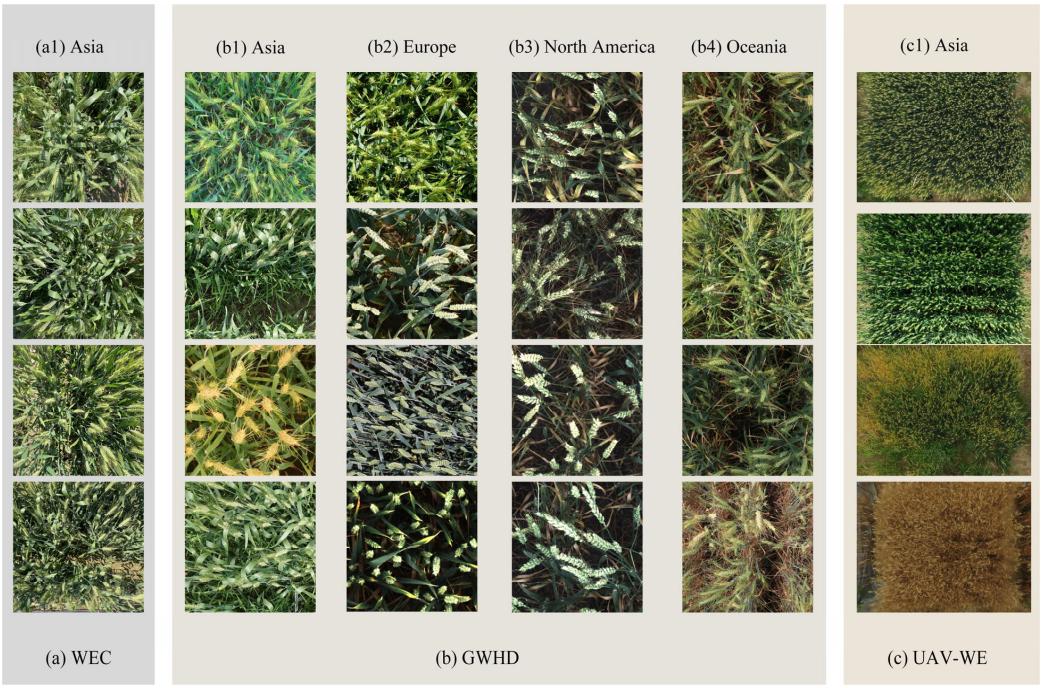

Figure 1. Examples of wheat ears images from different datasets.

2.1.2. Data Annotation

The wheat images in the dataset are marked with ear points, that is, a point is marked at the position of each wheat ear and saved in the UCF-QNRF dataset format. This labeling method not only does not affect the resolution of the image but also can better preserve spatial information of the wheat ears. In order to save computation time and labeling workload, the wheat images in the WEC dataset are scaled to 1024 × 1024 pixels before labeling. An example of wheat ear labeling and an example of a saved label in ".mat" format, as shown in Figure 2.

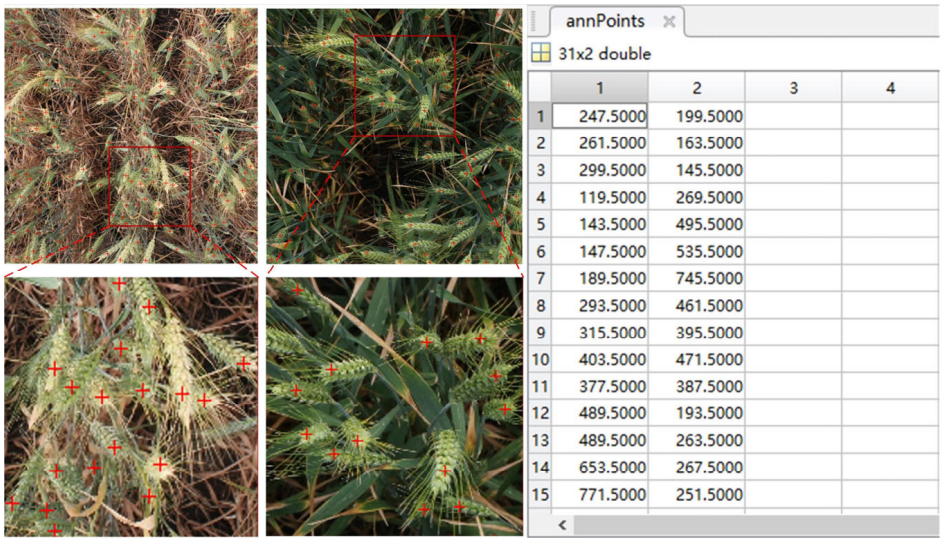

Figure 2. An example of wheat image annotation.

2.1.3. Dataset Construction

To enhance the superiority of the counting model, the wheat images in the training set, validation set, and test set are all in the same distribution and independent of each other. The WEC and GWHD datasets include 560 wheat images from different regions in different periods. In addition, the data set UAV-WE specifically includes two parts. One part is the original image of 24 wheat-filling stage images taken at a distance of 5 m from the ground, and its resolution is 4000 × 3000 pixels. The other part is to use a sliding window with a resolution of 512 × 512 pixels to intercept the above-mentioned image of 4000 × 3000 pixels to obtain a UAV image with a resolution of 512 × 512 pixels, from which 24 relatively dense images from UAV were selected. In this study, the data obtained by the ground system and UAV platform were randomly divided into training set, validation set, and test set according to the ratio of about 70%, 20%, and 10%, and 770, 224, and 114 pictures were obtained. The details are shown in Table 1.

Table 1. Construction of wheat ear image dataset.

Data Sets	Continent	Platform	Number	Training Set	Validation Set	Test Set	Size	Wheat Ear Num
WEC (filling)	Asian	G	313	219	63	31	1024 × 1024	8295
WEC (maturity)	Asian	G	187	131	37	19	1024 × 1024	4426
GWHD	Asian	G	140	98	28	14	1024 × 1024	6100
GWHD	Europe	G	140	98	28	14	1024 × 1024	6947
GWHD	North America	G	140	98	28	14	1024 × 1024	4742
GWHD	Oceania	G	140	98	28	14	1024 × 1024	7423
UAV-WE	Asian	U	24	14	6	4	4000 × 3000	80,274
	Asian	U	24	14	6	4	512 × 512	7682
Total			1108	770	224	114		125,889

Note: G represents the data obtained from the ground system, and U represents the data obtained from the UAV platform.

2.2. CA-MobileNetV3

MobileNetV3-Small is a lightweight neural network in the MobileNetV3 series, which combines the depthwise separable convolution (DSC) and the inverse residual structure of the linear bottleneck, the former of which comes from MobileNetV1 and the latter from MobileNetV2. In addition, MobileNetV3-Small also includes attention architecture, named squeeze and excite (SE), which is lightweight [22]. The DSC is used in the block instead of the standard convolution, which can reduce the calculation parameters of the network, aiming to make the network lightweight, and the linear bottleneck can effectively prevent the ReLU activation function from destroying information in low-dimensional space. In addition, the reverse residual structure can avoid the failure of the convolution kernel during the depth convolution process and the disappearance of the gradient when the network depth is too deep. However, the SE-block module has a large amount of calculation, and it only pays attention to channel information.

To improve the feature extraction ability of occluded wheat ears, MobileNetV3 is improved by introducing a coordinated attention mechanism (CA), which enables MobileNetv3 not only to capture the long-distance dependence in the spatial direction but also to obtain more accurate wheat ears location information, thereby improving the accuracy of wheat ear counting. The specific structure is shown in Figure 3.

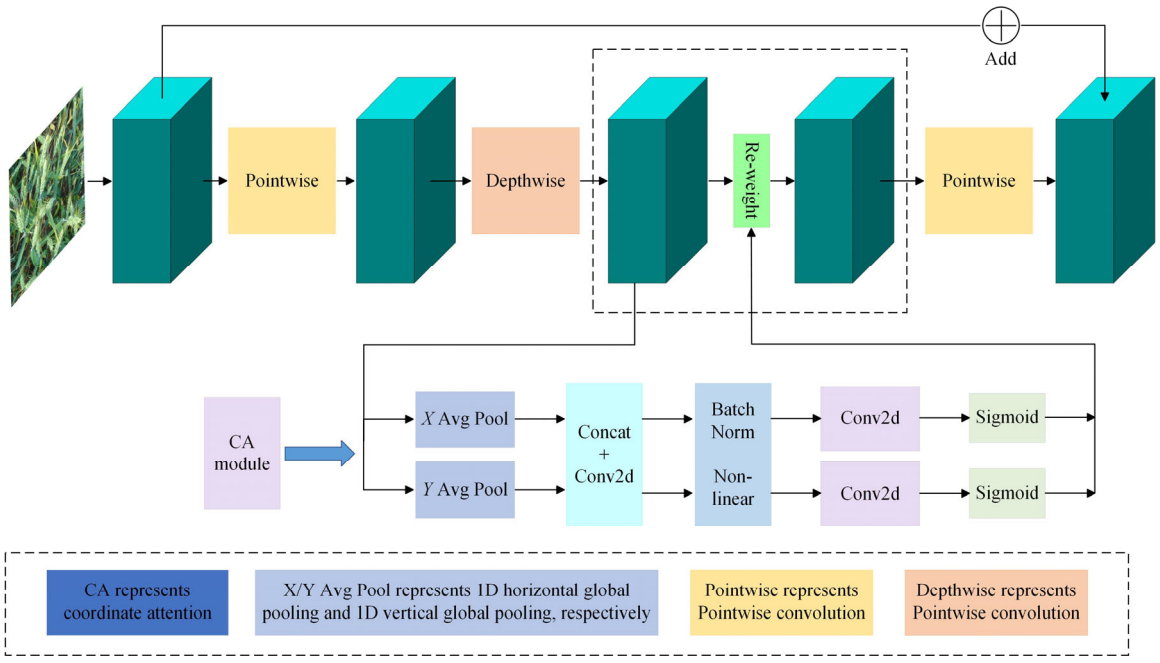

Figure 3. Structure diagram of CA-MobileNetV3 block.

It can be seen from Figure 3 that the input wheat images are trained in different directions and jointly learned to obtain mixed attention weights to activate input features. The channel outputs of height and width are, respectively:

$$T_c^h(h) = \frac{1}{W} \sum_{0 \le i \le W} x_{c(h,i)} \tag{1}$$

$$T_c^w(w) = \frac{1}{H} \sum_{0 \le j \le H} x_{c(j,w)} \tag{2}$$

where H and W are the height and width, respectively, x_c is the channel feature map, T_c^h represents the cth channel with height h, T_c^w represents the cth channel with height of w.

2.3. CARAFE Upsampling

CARAFE consists of two parts, Kernel Prediction Module and Content-aware Reassembly Module, and its structure is shown in Figure 4. In the Kernel Prediction Module, firstly, the channel number of the input feature map of $H \times W \times C$ is compressed to C_x, the multiple of upsampling is T, and the size after upsampling is $S_k \times S_k$, and the number of channels is changed from C_x becomes ${T^2} \times {S_k}^2$ to realize content coding, and then expands the channel features in the spatial dimension, and then performs Softmax normalization on the obtained upsampling kernel, so that the weight sum of the convolution kernel is 1. In the Content-aware Reassembly Module, each position in the output feature map is mapped to the input feature map, and the original feature map and the upsampled map are subjected to a dot product operation to obtain a feature map with a size of $TH \times TW \times C$.

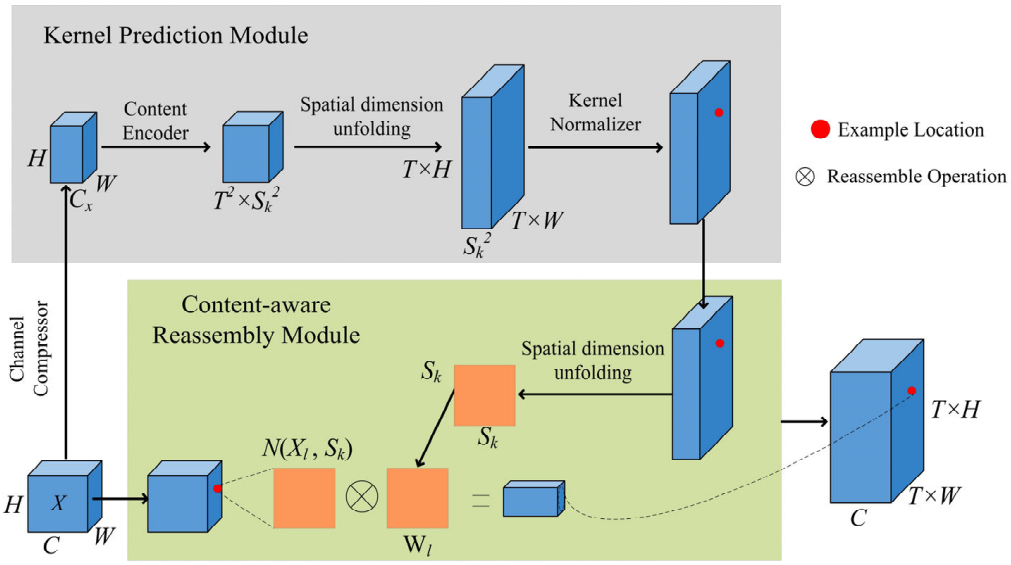

Figure 4. Structure diagram of CARAFE upsampling.

The CARAFE upsampling module has the characteristics of small redundancy, strong feature fusion ability, and fast operation speed, which can aggregate wheat ear feature information in a large receptive field. In addition, this module can make full use of the information of surrounding pixels to improve the expression ability of wheat ear features, which will help to obtain higher quality wheat ear feature maps.

2.4. LWDNet Model

To reduce the storage capacity of the regression counting network model based on the wheat ear density map, enhance the expression of the characteristic diversity and complexity of mutually occluded wheat ears, and increase the feature response of micro-wheat ears, a lightweight wheat ear density map regression counting network (LWDNet) was used. The structure of LWDNet is shown in Figure 5. H represents Height, W represents Width, and 3 represents the number of channels.

The LWDNet model is mainly divided into front network, end network, and regression counting network. CA-MobileNetV3 is the feature extraction network in the front network, which contains 3×3 convolution operation, 11-layer CA-MobileNetV3 block, and 1×1 convolution operation. It is used to extract the deep features of wheat ears. The regression counting network includes CARAFE upsampling module and 3 continuous convolution modules, which combine the applyColorMap function in OpenCV and the color mapping algorithm to visualize the output wheat ear density map.

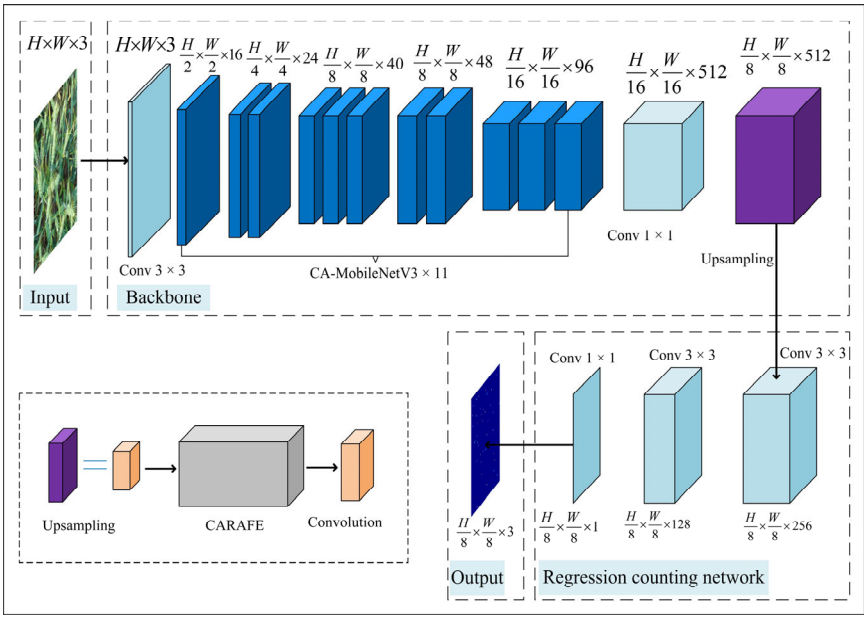

Figure 5. Structure diagram of LWDNet network model.

The density map is denoted as p, which reflects the spatial distribution information of wheat ears. The number of wheat ears N can be obtained by integrating the density values in the density map. Its definition is shown in Formula (3), where x represents any position of the density map and corresponds to the density value $\mathrm{p}(\mathrm{x})$.

$$N = \sum_{\mathrm{x} \in \mathbf{p}} \mathrm{p}(\mathrm{x}) \tag{3}$$

The specific parameters of the LWDNet model in this study are shown in Table 2. Among them, Input represents the input feature matrix, Operator represents the operation or module, Output represents the number of feature matrix channels finally output by each Operator, CA represents collaborative attention mechanism, NL represents the type of nonlinear activation function used, where HS stands for h-swish, RE stands for ReLU, S stands for stride.

Table 2. Parameters of the LWDNet model.

Input	Operator	Output	CA	NL	S
1024 × 1024 × 3	Conv2d, 3 × 3	3	-	-	1
1024 × 1024 × 3	CA-MobileNetV3	16	√	RE	2
512 × 512 × 16	CA-MobileNetV3	24	-	RE	2
256 × 256 × 24	CA-MobileNetV3	24	-	RE	1
256 × 256 × 24	CA-MobileNetV3	40	√	HS	2
128 × 128 × 40	CA-MobileNetV3	40	√	HS	1
128 × 128 × 40	CA-MobileNetV3	40	√	HS	1
128 × 128 × 40	CA-MobileNetV3	48	√	HS	1
128 × 128 × 48	CA-MobileNetV3	48	√	HS	1
128 × 128 × 48	CA-MobileNetV3	96	√	HS	2
64 × 64 × 96	CA-MobileNetV3	96	√	HS	1
64 × 64 × 96	CA-MobileNetV3	96	√	HS	1
64 × 64 × 96	Conv2d, 1 × 1	512	-	-	1
64 × 64 × 512	Upsample	512	-	-	-
128 × 128 × 512	Conv2d, 3 × 3	256	-	-	1
128 × 128 × 256	Conv2d, 3 × 3	128	-	-	1
128 × 128 × 128	Conv2d, 1 × 1	1	-	-	1

Note: - indicates that the CA module is not included, and √ indicates that the CA module is included.

2.5. Regression Count of Wheat Ear Density Map Based on Lwdnet Model

2.5.1. Overall Technical Route

Figure 6 shows the technical route of this study. Firstly, the wheat images used as the training set were manually annotated. Secondly, the constructed LWDNet model was trained, and the optimal wheat ear counting model was obtained by optimizing parameters. Finally, the model was tested. The predicted density map of wheat ears was obtained through the LWDNet model. The predicted density map of wheat ears was obtained through the LWDNet model, and the sum of the density values in the density map was the predicted number of wheat ears, which reflected the approximate spatial location and distribution of wheat ears. We also evaluated the model on the basis of the ground truth and predicted ears.

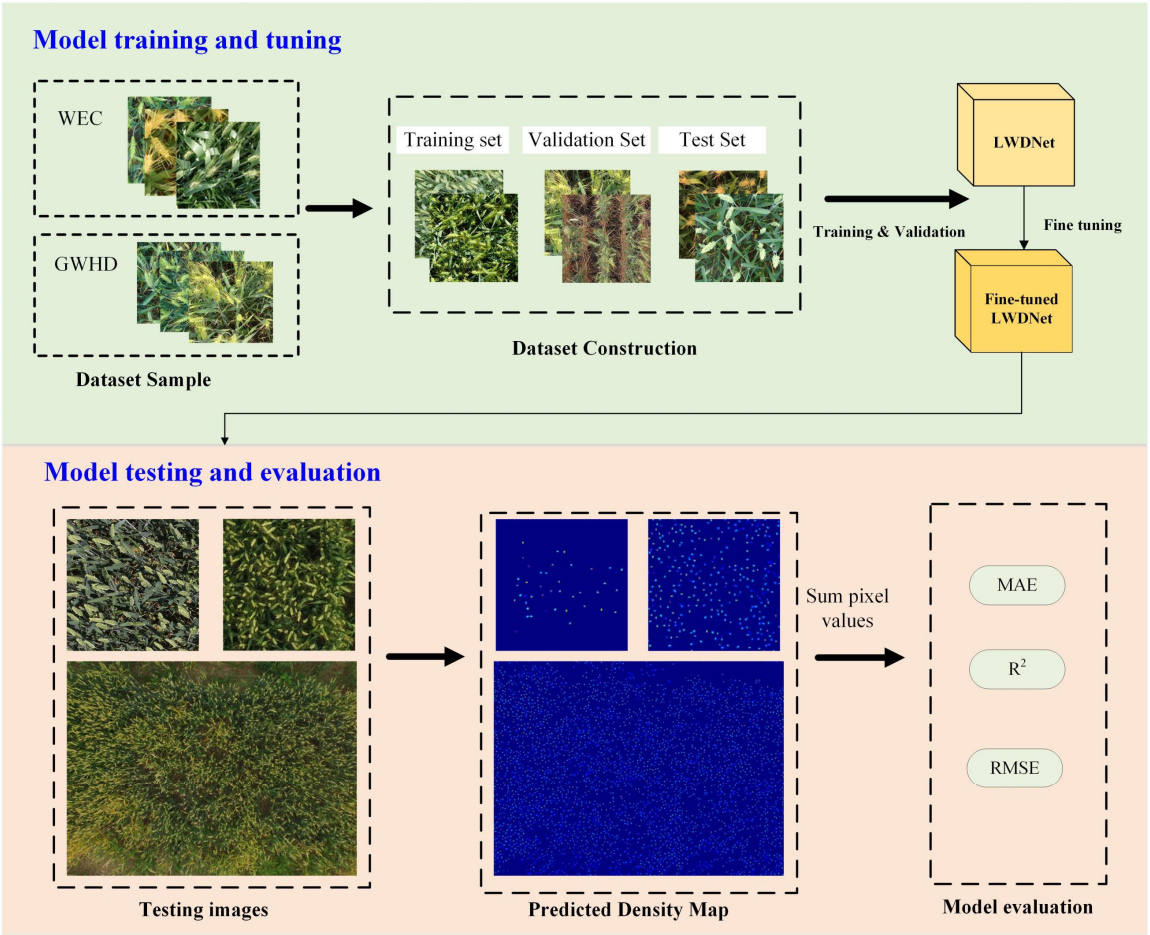

Figure 6. Overall technology roadmap.

2.5.2. Design of Loss Function

The loss function is generally used to estimate the difference between the predicted value of the model and the ground truth from the density map. In complex scenes, the scale of dense wheat ears varies, and it is difficult to select appropriate Gaussian checkpoint annotations for smoothing, resulting in the generation of density maps of wheat fields with additional errors. To obtain good generalization performance of the LWDNet model, we constructed the loss function L_{Total} of LWDNet using the count loss L_C, the optimal transmission loss L_{OT}, and the overall variation loss L_{TV}. Among them, L_{Total} is used to measure the total difference between the ground truth density map g and the predicted density map p, and L_{OT} and L_{TV} are used to calculate the distribution difference between

g and p. L_C is used to supervise the relationship between the total number of wheat ears and the total number of predictions, and its definition is shown in Equation (4).

$$L_C = |P - G| \tag{4}$$

Among them, P and G, respectively, represent the predicted number of wheat ears and the actual number of wheat ears obtained by integrating and summing the density values according to the density map of predicted p and ground truth g.

The optimal transmission loss L_{OT} is defined as shown in Equation (5).

$$L_{OT} = \left\langle \alpha^*, \frac{\mathrm{p}}{\|\mathrm{p}\|_1} \right\rangle + \left\langle \beta^*, \frac{\mathrm{g}}{\|\mathrm{g}\|_1} \right\rangle \tag{5}$$

where $||*||_1$ represents the L1 norm of the vector, and α^* and β^* are the solutions based on the Monge–Kantarovich Optimal Transport formulation.

The definition of the total loss function is shown in Formula (6):

$$L_{TV} = \frac{1}{2} \| \frac{\mathrm{p}}{\|\mathrm{p}\|_1} - \frac{\mathrm{g}}{\|\mathrm{g}\|_1} \|_1 \tag{6}$$

The total loss function L_{Total} is the weighted sum of the above three losses, and to ensure that the count loss L_C and the total variation loss L_{TV} have the same scale, multiply the total change loss L_{TV} by $\|\mathrm{g}\|_1$.

$$L_{Total} = L_C + \lambda_1 L_{OT} + \lambda_2 \|\mathrm{g}\|_1 L_{TV} \tag{7}$$

where λ_1 and λ_2 are adjustable hyperparameters used to balance the optimal transmission loss L_{OT} and the overall variation loss L_{TV}. In this study, λ_1 and λ_2 are set to 0.1 and 0.01, respectively.

2.5.3. Generation of Ground Truth Density Maps

The density map regression method was used to obtain the number of wheat ears, and the density map g of the ground truth of wheat ears was generated according to the point labeling. On the marked wheat ear image, according to the position of the marked wheat ear, the size of the wheat ear at the position is estimated, and the coverage area of the wheat ear is obtained. Assuming that there is a wheat ear $i(i = 1, 2, .., N)$ at the position of pixel x_i in the wheat image, the probability that the pixel point at position x on the density map belongs to the wheat ear is represented by the impulse function $\delta(x - x_i)$. Therefore, the total probability value $\mathrm{g(x)}$ of the pixel at position x_i belonging to the wheat ear in an image with N wheat ear labeled points is expressed as:

$$\mathrm{g(x)} = \sum_{i=1}^{N} \delta(x - x_\mathrm{i}) \tag{8}$$

2.5.4. The Evaluation Index of The Model

To effectively appraise the proposed model, we used the mean absolute error (MAE) and the root mean squared error ($RMSE$) are used, and the coefficient of determination (R^2). The formulas are as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^{n} \left| x_i^{GT} - x_i^P \right| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{n} \left| x_i^{GT} - x_i^P \right|^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(x_i^{GT} - x_i^{P}\right)}{\sum_{i=1}^{N}\left(x_i^{GT} - \overline{x}_i^{P}\right)} \tag{11}$$

MAE is the average of the absolute values of all forecast deviations, which can avoid the offset of positive and negative errors due to the calculation of the absolute values. Therefore, MAE can directly reflect the accuracy of the wheat ear counting model, and $RMSE$ can better reflect the robustness of the wheat ear counting model. R^2 represents the degree of fit between the estimated value of ear count and the ground truth.

In addition, we also used the number of parameters (Parameter), floating point operations (FLOPs), the Model Size, and FPS (Frames Per Second) to measure the counting efficiency of our model.

3. Results

3.1. Model Training

In this study, all models were trained and tested on the same device, and the hardware and software details of the device are shown in Table 3.

Table 3. Details of experimental hardware and software parameters.

Configuration Name	Parameter
Operating system	Windows 10 Professional 64-bit
Code execution Environment	Python 3.7
Deep Learning Framework	Pytorch 1.80
GPU model	NVIDIA GeForce RTX 2080
Processor	Intel Core i7-8700 CPU @ 3.20 GHz

The LWDNet model was trained and validated using the wheat dataset shown in Table 1. The training set and validation set were used to build the LWDNet model. In this study, the following initial parameters are used to train the LWDNet model: Learning rate = 0.00001, Weight decay = 0.0001, Max Epoch = 5000, Batch size = 2.

When the Epoch reaches about 4000, the value of Loss gradually stabilizes, and the corresponding MAE and $RMSE$ also tend to be stable. Among them, when the Epoch is 4845, when the value of Loss reaches the lowest value, that is, 0.40, the corresponding MAE and $RMSE$ also reach 0.32 and 0.53, respectively. The corresponding LWDNet model is the optimal model, which will be used for subsequent testing.

3.2. Counting Results of Wheat Ears for Ground System

An example of counting results for wheat ears is shown in Figure 7. Among them, Figure 7a,b showed the predicted ear density map and ear number of wheat images from the GWHD and WEC datasets in the test set, respectively. The top row shows the RGB image obtained from different platforms, the second row shows the ground truth density map generated according to the point annotation, the third row represents the predicted density map based on our model, and the fourth row represents the superimposed image of the predicted density map and the original image.

As can be seen from Figure 7, the distribution of the wheat ear density map obtained by our model is not much different from that of the ground truth density map. In addition, from the predicted value in Figure 7, we also found that the predicted value is close to the ground truth of wheat ears. Therefore, the model has a good counting effect on images of different periods, different densities, and different regions, which verifies the strong generalization ability of the model.

Moreover, to verify the wheat ear prediction results of the LWDNet model, three datasets were used for testing, and relevant evaluations were made on the test results, as shown in Table 4. From the results in Table 4, it was obvious that the counting effects of LWDNet models tested with different datasets performed well. Among them, the model's

MAE were 1.36, 1.82, and 1.39; *RMSE* were 1.59, 2.57, and 1.71, respectively. In particular, the R^2 of the LWDNet model reached 0.9672 in 16 images of the GWHD, including 7204 wheat ears. In the 50 images of the WEC, including 12,721 wheat ears, the LWDNet model R^2 reached 0.9787. In the 66 images of the WEC, 19,925 wheat ears were included. The LWDNet model R^2 achieved 0.9792.

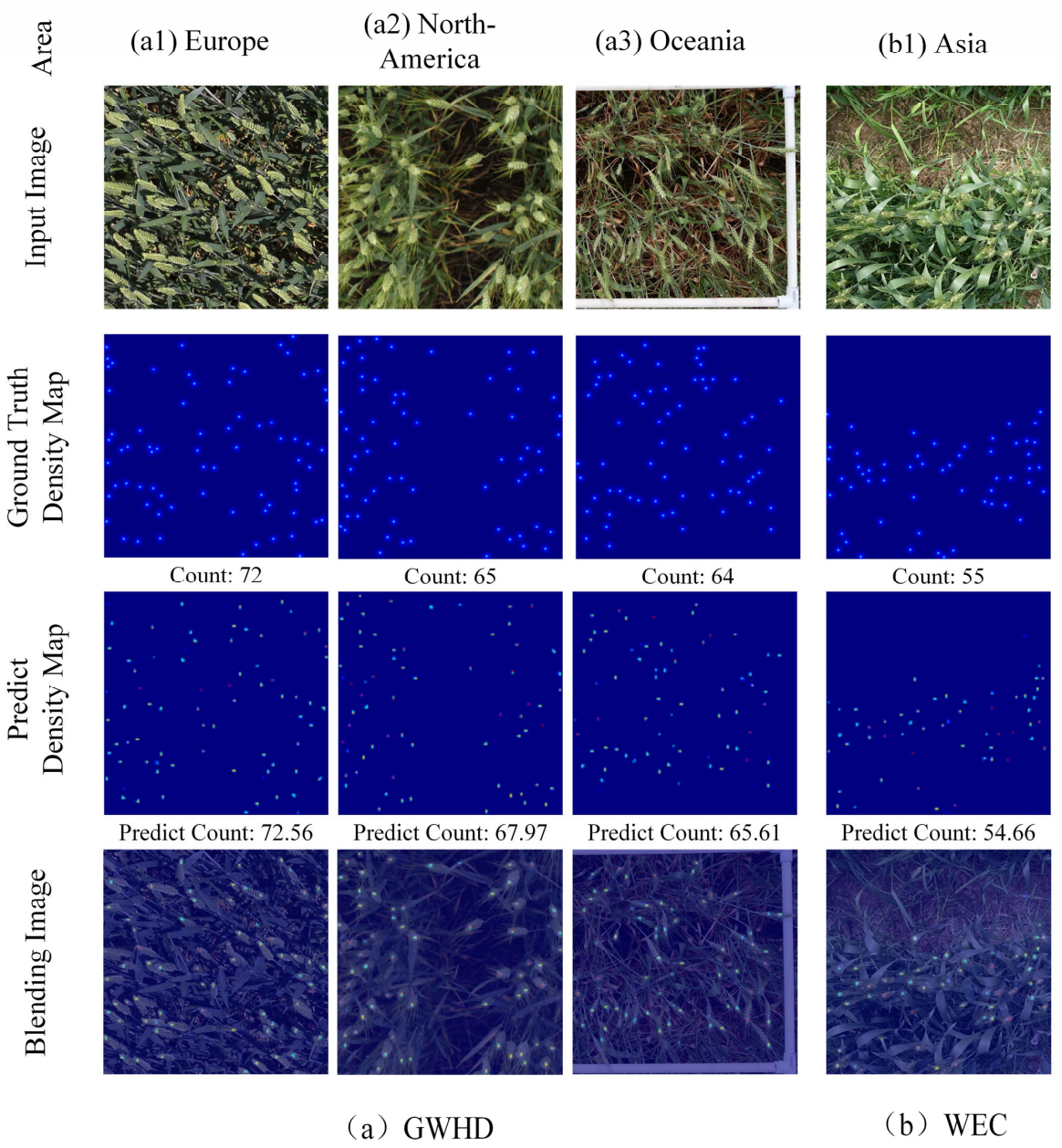

Figure 7. Example of wheat ear count results for Ground Systems.

Table 4. Density regression count results for the test set.

Different Data	Image Number	*MAE*	*RMSE*	R^2
WEC	50	1.36	1.59	0.9672
GWHD	16	1.82	2.57	0.9787
WEC + GWHD	66	1.39	1.71	0.9792

In addition, the *MAE* of the count results of the test set WEC constructed from images collected in this region increased by 25.2% compared to the dataset GWHD with images

collected in different continents and increased by 2.2% compared to the mixed dataset WEC + GWHD. The WEC test set results were better than the GWHD test set results, probably because the GWHD dataset comes from four different continents, and the size, shape, and color of the wheat ears were quite different. However, the wheat varieties in the WEC dataset were relatively single, and different ears of wheat had a certain degree of similarity, which led to the fact that the test effect on GWHD was inferior to that on WEC.

3.3. Counting Results of Wheat Ears for UAV Platform

To verify whether the LWDNet model can effectively count wheat ears in a dense wheat scene and evaluate the application ability and potential promotion value of the LWDNet model. The model was tested using the UAV-WC dataset, which was acquired from the UAV platform. Twenty denser wheat images in UAV-WC were selected for testing, and the total number of wheat ears in each image exceeded 200. The results of automatic counting using our model are shown in Figure 8. In Figure 8, the first row represents the RGB image of wheat aerially taken from the UAV platform, and the second to third rows represent the density map of wheat ears, which are the ground truth density map generated based on the point labeling and the predicted density using the model proposed in this study. The fourth row represents the overlay of the two density maps mentioned above.

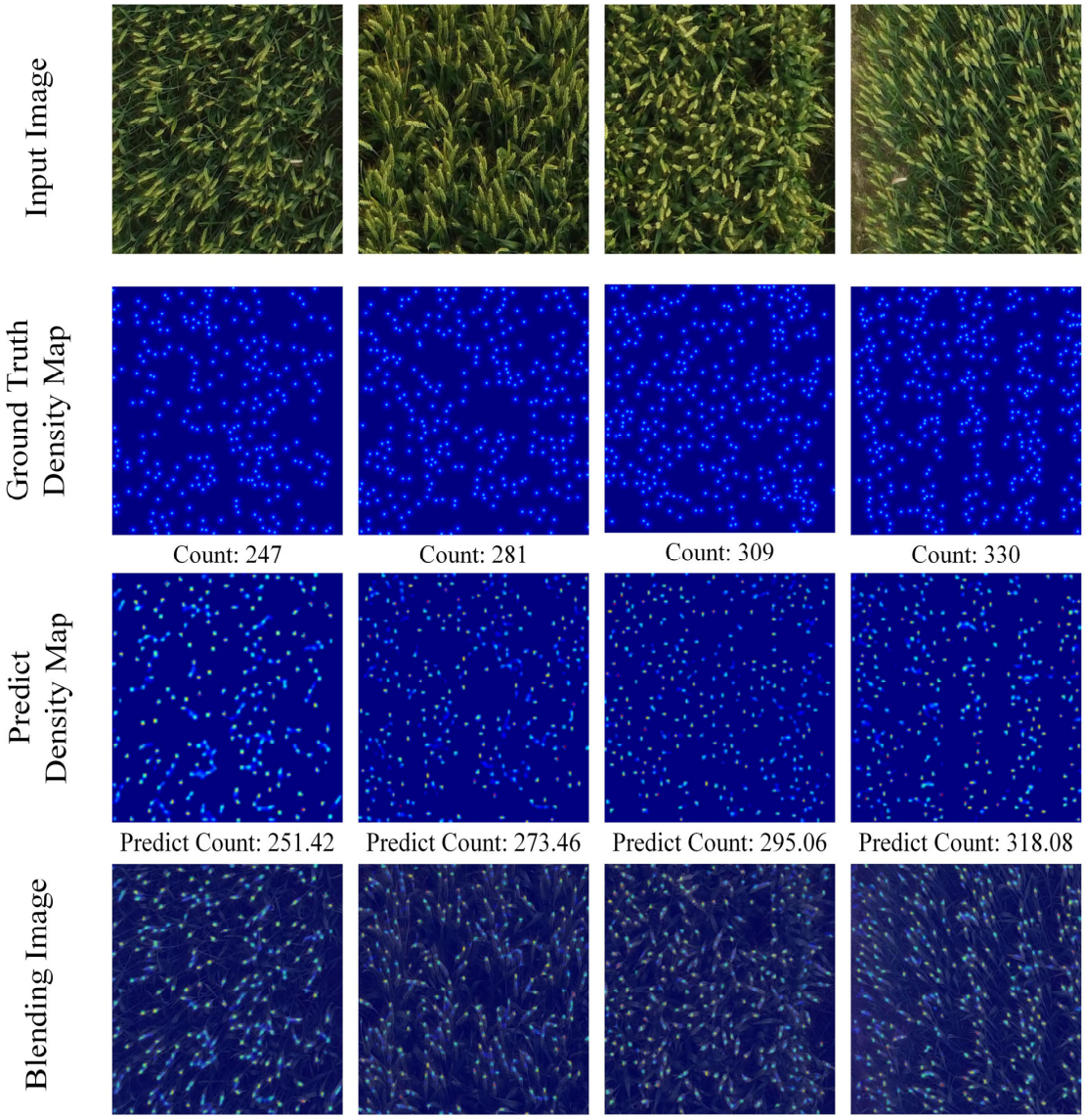

Figure 8. Low-density wheat counting results obtained for UAV platform.

From Figure 8, we found that the wheat ear counting model LWDNet had a better effect. Although some tiny wheat ears were ignored in the wheat image, the predicted density map could reflect the spatial information of most ears in the original wheat image, and the predicted number of wheat ears was close to that of the ground truth.

4. Discussion

4.1. Comparison of Counting Results for Ground System Based on Different Models

To examine the counting performance of our model, self-built test set was used to compare the results of typical density map counting models, including multi-column convolutional neural network (MCNN) [23], congested scenes recognition network (CSRNet) [24], distribution matching for crowd counting (DM-Count) [25]. Among them, the software and hardware environment obtained by the inference speed indicator FPS: Windows10 operating system, Intel Core i7-8700 CPU @ 1.80 GHz, NVIDIA GeForce GTX 2080 GPU, memory 8 G.

As can be seen from Table 5, R^2 based on LWDNet is 83.8%, 76.1%, and 1.9% higher than that of the MCNN, CSRNet, and DM-Count, respectively. The *MAE* based on LWDNet is 80.4%, 76.6%, and 32.2% lower than that of MCNN, CSRNet, and DM-Count. The *RMSE* based on LWDNet is 85.9%, 83%, and 35.2% lower than that of MCNN, CSRNet, and DM-Count. Therefore, the experimental results showed that our model outperformed other models in multiple evaluation indicators.

Table 5. Regression count results of wheat ear density for different models.

Method	*MAE*	*RMSE*	R^2
MCNN	7.09	12.09	0.1578
CSRNet	5.93	10.08	0.2327
DM-Count	2.05	2.64	0.9546
LWDNet	1.39	1.71	0.9726

In addition, it can be seen from Table 6 that the Parameter, Model Size, and FPS of LWDNet proposed in this study reached 2.38 million, 9.24 MB, and 58.82, respectively. Compared with CSRNet and DM-Count, LWDNet reduced by 13.88 million and 19.12 million for Parameter, 52.81 MB and 72.78 MB for Model Size, and increased by 49.9 and 49.56 for FPS. Further, the Parameter, Model Size, and FPS of MCNN and LWDNet were close. However, the counting accuracy of LWDNet was significantly better than that of MCNN. Table 6 shows a comparison of model evaluation metrics. In terms of model size, the LWDNet model was 85.1% and 88.7% less than CSRNet and DM-Count. In terms of parameters, the LWDNet model was 85.4% and 88.9% less than CSRNet and DM-Count. The FLOPs of the LWDNet model were 92.3% and 92.2% lower than both CSRNet and DM-Count. Moreover, the FPS of the LWDNet model was not the best. However, the density estimation accuracy based on MCNN performed the worst among the four models [26]. In conclusion, LWDNet outperformed the other three models in terms of counting accuracy and efficiency.

Table 6. Comparison results of performance indicators of different counting models.

Method	Parameter (Million)	Model Size (MB)	FLOPs (G)	FPS
MCNN	0.13	0.52	28.23	18.7
CSRNet	16.26	62.05	433.36	6.01
DM-Count	21.5	82.02	432.16	6.4
LWDNet	2.38	33.58	9.24	58.82

Figure 9 shows the test results on different datasets. It can be seen from Figure 10 that MCNN and CSRNet have low accuracy for wheat ear statistics in severe occlusion

scenarios. It may be because the loss function of the density map engendered based on the MCNN and CSRNet models is determined according to the Gaussian kernel. However, it is difficult to determine the appropriate size of Gaussian kernel due to the different sizes and shapes of ears in wheat images [27].

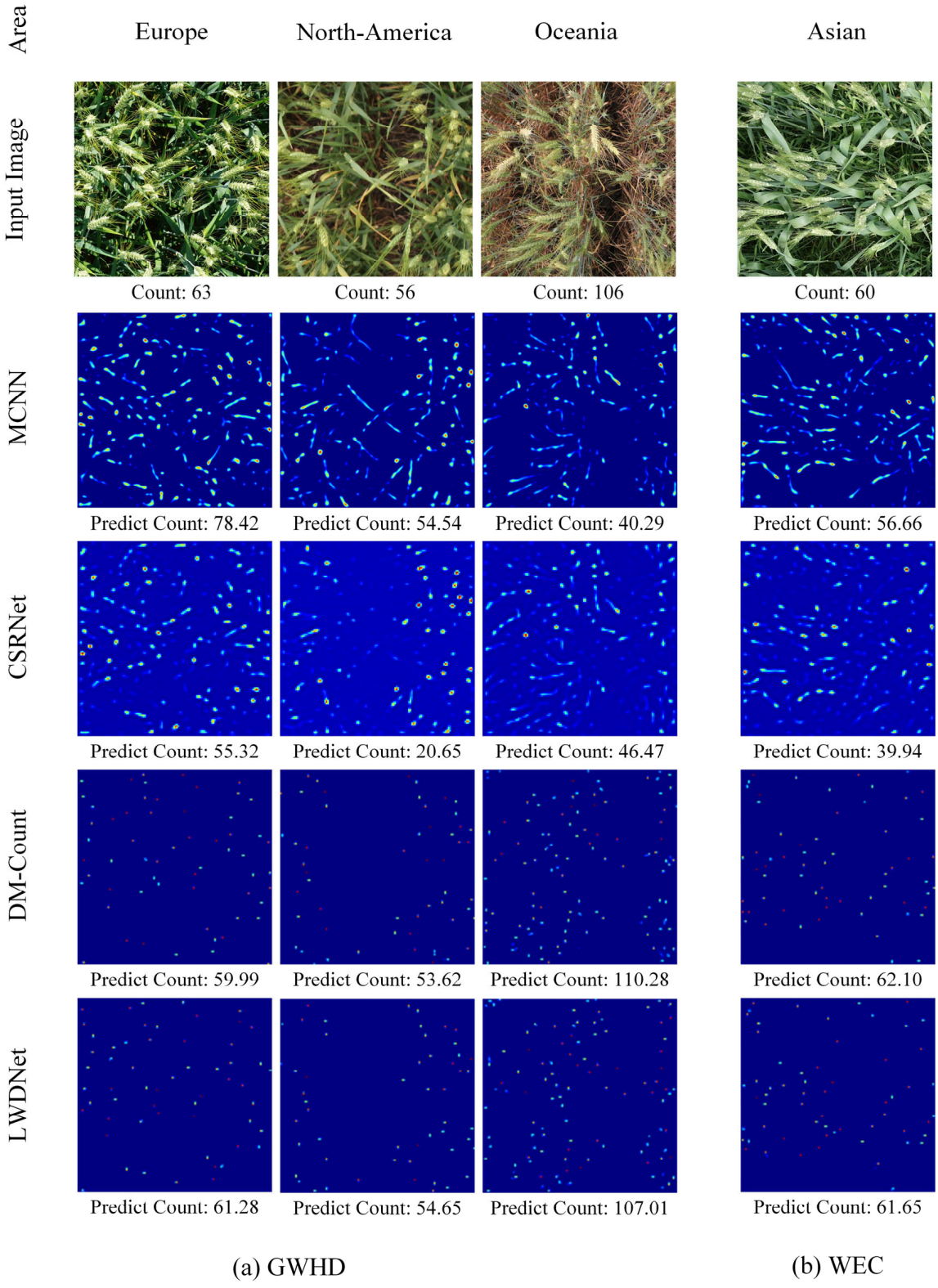

Figure 9. Counting results of wheat ears based on different models.

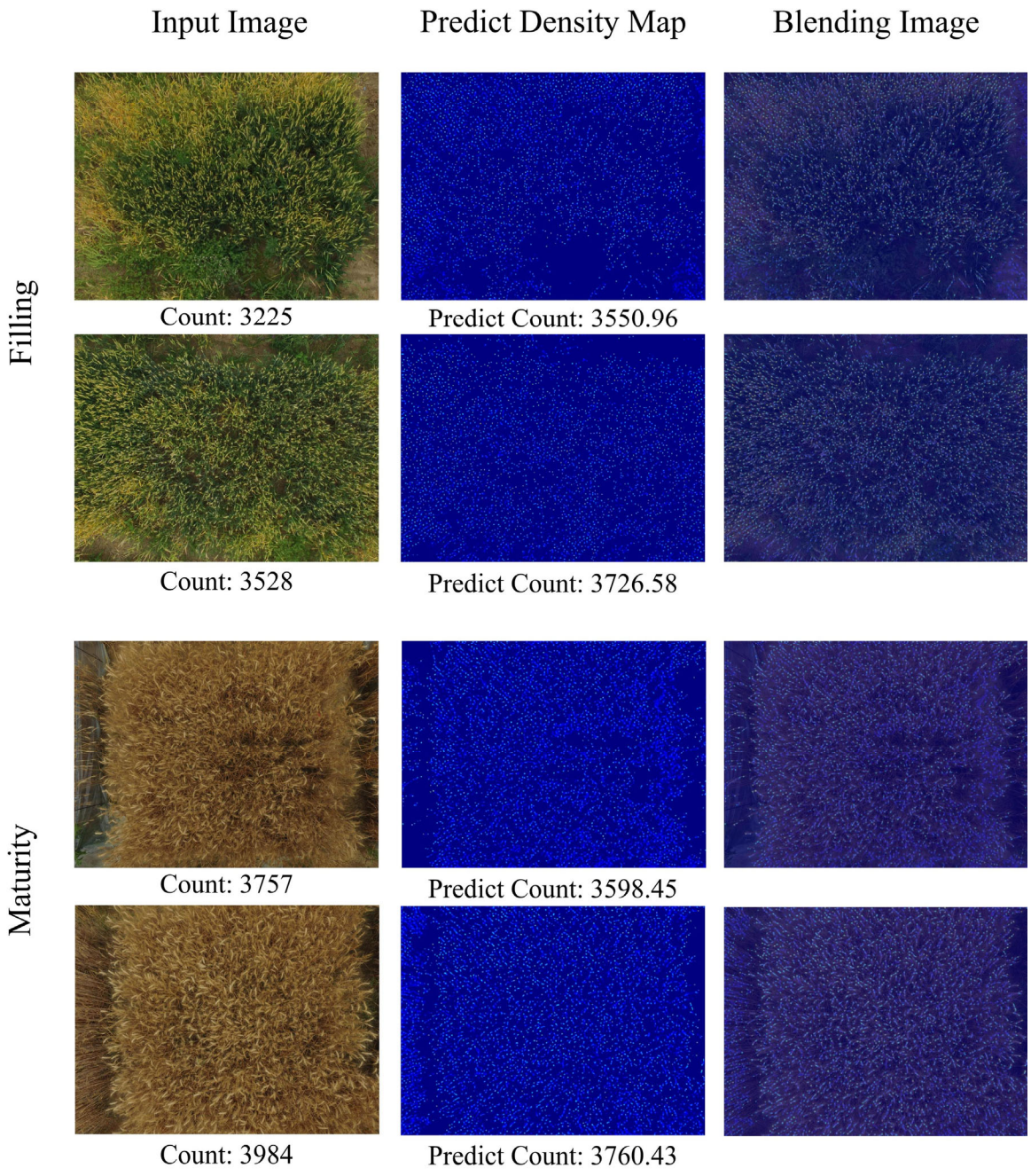

Figure 10. High-density wheat counting results obtained for UAV platform.

4.2. Counting Results of High-Density Wheat Ears for UAV Platform

Actually, the density distribution and overlapping degree of wheat ears in the images acquired by different platforms are different, and each image contains about 50–100, 200–400, 3000–4000 wheat ears, respectively, as shown in Figures 7, 8 and 10. Therefore, in order to verify the robustness of the proposed LWDNet model, four RGB images of wheat obtained from drones in the UAV-WC dataset with high-density and large-scale were used for testing, including the filling period and the maturity period, the result of counting, as shown in Figure 10. From the results in Figure 10, it can be found that the wheat ears with high-density distribution in the filling stage are estimated to be 4.2–5.6% higher than the ground truth, and the estimated values of wheat ears in the mature stage are 5.6% and 10.1% lower than the ground truth. The possible reason is that the wheat in the filling period is growing vigorously, and the wheat ears and wheat leaves have similar

characteristics, which may easily lead to false detection. However, the tips of wheat ears in the mature stage are all vertical to the ground, resulting in missed detection.

Although the density map predicted by the LWDNet model still had a few ears that were not completely consistent with the ground truth, the LWDNet model could effectively distinguish the irrelevant background from wheat, and the predicted density map could generally reflect the spatial distribution and count of wheat ears captured by drones at low altitudes. The visualization of experimental results is beneficial to better understand how well the model is performing. Through the test results, it can be found that the LWDNet model proposed in this study can generate high-quality wheat ear density maps for wheat ear images of different densities obtained by ground systems and UAV platforms, and the error of density estimation is small.

In addition, when the UAV flies at a low altitude, the airflow generated by the rotation of its blades will cause the leaf structure of the crop canopy to be unstable, which increases the difficulty of phenotypic analysis and detection. Therefore, it is very important to choose the best aerial photography altitude without affecting the image quality, aiming to obtain as high-resolution images as possible. In addition, while providing high resolution, the capture throughput of the UAV remote sensing system is controlled as much as possible to improve flight work efficiency.

4.3. Existing Counting Challenges and Visualization of Existing Techniques

Recently, to accurately identify and count ears of wheat in field wheat images taken under natural conditions, many researchers built different semantic segmentation models using a deep learning framework. For example, Ma et al., built a network model based on the combination of a deep convolutional neural network and a fully convolutional network (FCN) and realized the counting of wheat ears through pixel-level semantic segmentation [28]. Sadeghi-Tehran et al., built the model with simple linear iterative clustering, which can successfully segment the wheat ears from the complex background, thereby realizing the counting of wheat ears [29]. Wang et al., constructed the SSRNet model to achieve field wheat ear segmentation, which provided a good foundation for wheat ear counting under field conditions [30]. Misra et al., constructed SpikeSegNet to segment and count spikes in wheat plants [31]. The above studies have shown that the deep learning model can effectively segment wheat ears for accurate counting. However, it is seriously affected by the occlusion of wheat ears and leaves, changes in the scale of wheat ears, and the background interference of wheat fields. There are still many challenges to achieving high accuracy in predicting wheat ear counts in a single image. We improved the YOLOv4 model to better realize the counting of wheat ears by detecting wheat ears and counting the number of detection frames [7]. However, in the face of wheat field scenes of different scales, different densities, and different complex backgrounds, the robustness of detection boxes is not as good as that of counting wheat ears based on density map estimation. Moreover, in this study, we have improved the regression estimation method based on the density map, which has good performance in the detection and counting of wheat ears with severe occlusion and different scales, including ground scale and UAV scale.

To better illustrate the superiority of the proposed model, the class activation map (CAM) was used to verify the robustness of the proposed model in this study. Class activation map (CAM) is an intuitive method in convolutional neural network (CNN) interpretation, and it was usually generated by the last convolutional layer of a CNN, which can highlight different regions of the target class in the input image. To verify the contribution of CA-MobileNetV3 and CARAFE upsampling module to the count of wheat ears, an object localization method (Grad-CAM++) [32] was used to construct important regions of multiple objects on the image through gradients with convolutional layers for feature visualization. In this study, LWDNet and DM-Count were selected for feature visualization comparison, and the original wheat ear map, the wheat ear heat map of

DM-Count and LWDNet, and the wheat ear visualization class heat activation map, were displayed, respectively, as shown in Figure 11.

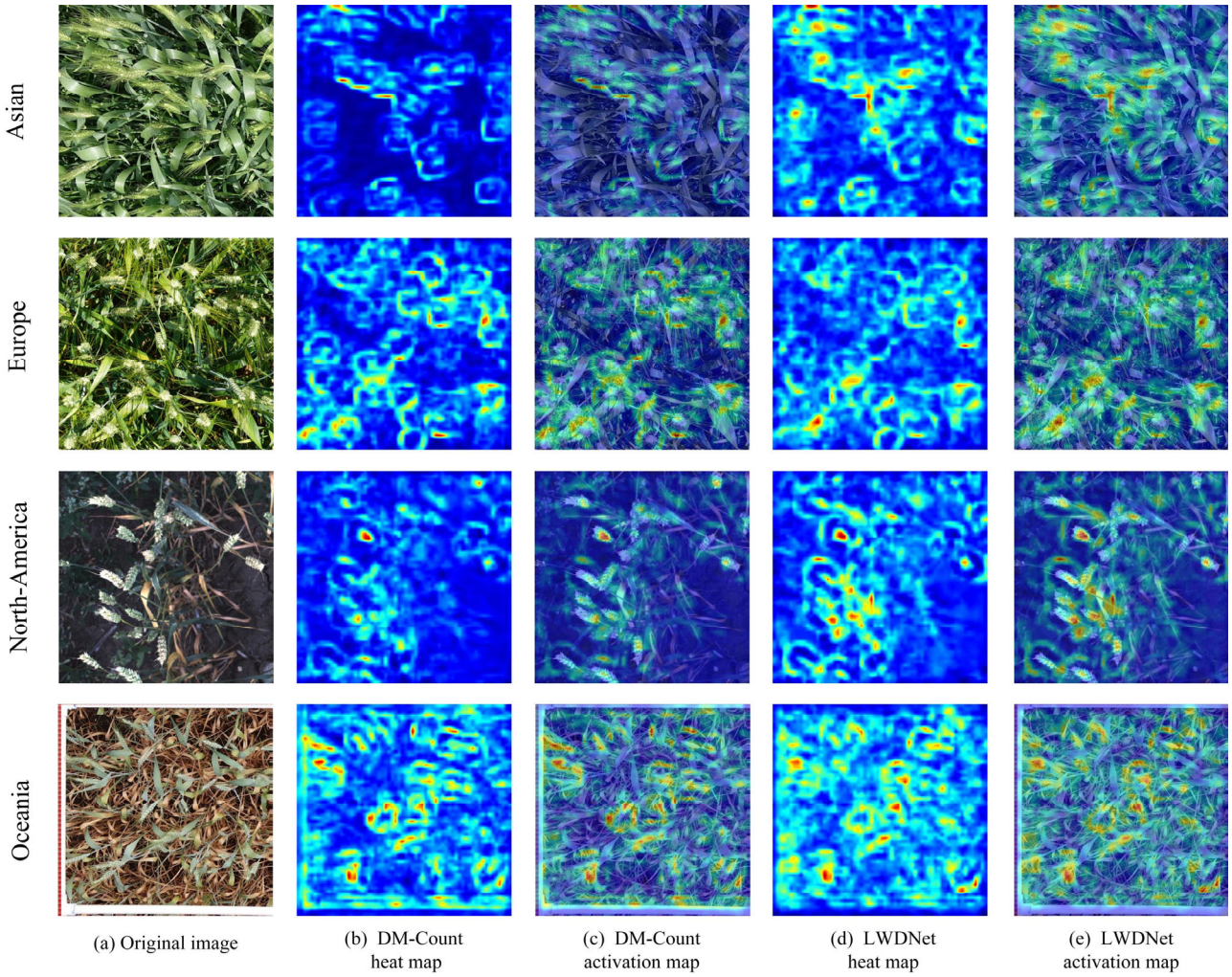

Figure 11. Visualization of feature activations using Grad-CAM++.

In Figure 11b–e, the red and light blue regions in the feature map represented the regions activated by the network of LWDNet, and the dark blue background showed the inactivated parts by any network. Among them, the redder part indicated that the influence of the network on the target recognition of ears of wheat was greater. In particular, in Figure 11b,d, we found that DM-Count pays more attention to background information than wheat ears. Obviously, the model is not enough to focus on the characteristics of small target wheat ears. We noticed from Figure 11c,e that the LWDNet network model paid more attention to the characteristics of wheat ears more finely and accurately. It could be seen that the CA-MobileNetV3 and CARAFE upsampling module used in the proposed LWDNet network model could better learn the characteristic information of wheat ears so that the counting effect of the ear density map is good.

5. Conclusions

To achieve accurate and efficient counting of wheat ears from different platforms, including ground systems and UAV platforms, a lightweight network model LWDNet is proposed in this study. This structure not only introduces a collaborative attention mechanism to improve the lightweight neural network but also uses the CARAFE upsampling module to optimize the feature fusion layer, thereby overcoming the problems of dense distribution of wheat ears at different scales, serious overlap, small size, and complex

background information. The experimental results show that the model proposed in this study realizes cross-platform wheat ear counting, which is of great significance for wheat field monitoring and wheat yield estimation in different spatial scale application scenarios, and also provides a reference for the estimation of other scenarios.

Author Contributions: Writing, Resources, Writing—review & editing: B.Y., Methodology, Software: M.P. and Z.G., Data curation and Validation: M.P., H.Z. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Major Science and Technology Projects in Anhui Province (202203a06020007), the Select the Best Candidates to Undertake Key Research Project of common technologies in Hefei City (GJ2022QN03), the Open Project of Jiangsu Key Laboratory of Information Agriculture (Grant No. 15266).

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

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

References

1. Hasan, M.M.; Chopin, J.P.; Laga, H.; Miklavcic, S.J. Detection and analysis of wheat spikes using convolutional neural networks. *Plant Methods* **2018**, *14*, 100. [CrossRef]
2. Du, S.; Li, Y.; Yao, M.; Ding, Q.; He, R. Counting method of grain number based on wheat ear spikelet image segmentation. *J. Nanjing Agric. Univ.* **2018**, *41*, 742–751.
3. Wang, D.; Fu, Y.; Yang, G.; Yang, X.; Zhang, D. Combined use of FCN and Harris corner detection for counting wheat ears in field conditions. *IEEE Access* **2019**, *7*, 178930–178941. [CrossRef]
4. Madec, S.; Jin, X.; Lu, H.; De, S.; Liu, S.; Duyme, F.; Heritier, E.; Baret, F. Ear density estimation from high resolution RGB imagery using deep learning technique. *Agric. For. Meteorol.* **2019**, *264*, 225–234. [CrossRef]
5. Wang, Y.; Qin, Y.; Cui, J. Occlusion robust wheat ear counting algorithm based on deep learning. *Front. Plant Sci.* **2021**, *12*, 1139. [CrossRef]
6. Yang, Y.; Huang, X.; Cao, L.; Chen, L.; Huang, K. Field wheat ears count based on YOLOv3. In Proceedings of the IEEE 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Dublin, Ireland, 16–18 October 2019; pp. 444–448.
7. Yang, B.; Gao, Z.; Gao, Y.; Zhu, Y. Rapid Detection and Counting of Wheat Ears in the Field Using YOLOv4 with Attention Module. *Agronomy* **2021**, *11*, 1202. [CrossRef]
8. Dandrifosse, S.; Ennadifi, E.; Carlier, A.; Gosselin, B.; Dumont, B.; Mercatoris, B. Deep learning for wheat ear segmentation and ear density measurement: From heading to maturity. *Comput. Electron. Agric.* **2022**, *199*, 107161. [CrossRef]
9. Zang, H.; Wang, Y.; Ru, L.; Zhou, M.; Chen, D.; Zhao, Q.; Zhang, J.; Li, G.; Zheng, G. Detection method of wheat spike improved YOLOv5s based on the attention mechanism. *Front. Plant Sci.* **2022**, *13*, 993244. [CrossRef]
10. Pound, M.P.; Atkinson, J.A.; Wells, D.M.; Pridmore, T.P.; French, A.P. Deep learning for multi-task plant phenotyping. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2055–2063.
11. Lempitsky, V.; Zisserman, A. Learning to count objects in images. In Proceedings of the 23rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–9 December 2010; Curran Associates Inc.: Red Hook, NY, USA, 2010; pp. 1324–1332.
12. Sindagi, V.A.; Patel, V.M. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognit. Lett.* **2018**, *107*, 3–16. [CrossRef]
13. Ma, J.; Li, Y.; Du, K.; Zheng, F.; Zhang, L.; Gong, Z.; Jiao, W. Segmenting ears of winter wheat at flowering stage using digital images and deep learning. *Comput. Electron. Agric.* **2021**, *168*, 105159. [CrossRef]
14. Khaki, S.; Safaei, N.; Pham, H.; Wang, L. Wheatnet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting. *Neurocomputing* **2022**, *489*, 78–89. [CrossRef]
15. Sadeghi-Tehran, P.; Sabermanesh, K.; Virlet, N.; Hawkesford, M.J. Automated method to determine two critical growth stages of wheat: Heading and flowering. *Front. Plant Sci.* **2017**, *8*, 252. [CrossRef]
16. Zhou, C.; Liang, D.; Yang, X.; Xu, B.; Yang, G. Recognition of wheat spike from field based phenotype platform using multi-sensor fusion and improved maximum entropy segmentation algorithms. *Remote Sens.* **2018**, *10*, 246. [CrossRef]
17. Zhu, Y.; Cao, Z.; Lu, H.; Li, Y.; Xiao, Y. In-field automatic observation of wheat heading stage using computer vision. *Biosyst. Eng.* **2016**, *143*, 28–41. [CrossRef]
18. Zhaosheng, Y.; Tao, L.; Tianle, Y.; Chengxin, J.; Chengming, S. Rapid Detection of Wheat Ears in Orthophotos from Unmanned Aerial Vehicles in Fields Based on YOLOX. *Front. Plant Sci.* **2022**, *13*, 1272. [CrossRef] [PubMed]

19. Sun, J.; Yang, K.; Chen, C.; Shen, J.; Yang, Y.; Wu, X.; Norton, T. Wheat head counting in the wild by an augmented feature pyramid networks-based convolutional neural network. *Comput. Electron. Agric.* **2022**, *193*, 106705. [CrossRef]
20. Fernandez-Gallego, J.A.; Lootens, P.; Borra-Serrano, I.; Derycke, V.; Haesaert, G.; oldán-Ruiz, I.; Araus, J.L.; Kefauver, S.C. Automatic wheat ear counting using machine learning based on RGB UAV imagery. *Plant J.* **2020**, *103*, 1603–1613. [CrossRef] [PubMed]
21. Petti, D.; Li, C. Weakly-supervised learning to automatically count cotton flowers from aerial imagery. *Comput. Electron. Agric.* **2022**, *194*, 106734. [CrossRef]
22. Tarek, H.; Aly, H.; Eisa, S.; Abul-Soud, M. Optimized Deep Learning Algorithms for Tomato Leaf Disease Detection with Hardware Deployment. *Electronics* **2022**, *11*, 140. [CrossRef]
23. Yan, L.; Zhang, L.; Zheng, X.; Li, F. Deeper multi-column dilated convolutional network for congested crowd understanding. *Neural Comput. Appl.* **2022**, *34*, 1407–1422. [CrossRef]
24. He, J.; Liu, Y.; Qiao, Y.; Dong, C. Conditional sequential modulation for efficient global image retouching. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 679–695.
25. Wang, B.; Liu, H.; Samaras, D.; Nguyen, M.H. Distribution matching for crowd counting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1595–1607.
26. Cao, J.; Zhang, Z.; Luo, Y.; Zhang, L.; Zhang, J.; Li, Z.; Tao, F. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. *Eur. J. Agron.* **2021**, *123*, 126204. [CrossRef]
27. Zeng, L.; Xu, X.; Cai, B.; Qiu, S.; Zhang, T. Multi-scale convolutional neural networks for crowd counting. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 465–469.
28. Ma, J.; Li, Y.; Liu, H.; Wu, Y.; Zhang, L. Towards improved accuracy of UAV-based wheat ears counting: A transfer learning method of the ground-based fully convolutional network. *Expert Syst. Appl.* **2022**, *191*, 116226. [CrossRef]
29. Sadeghi-Tehran, P.; Virlet, N.; Ampe, E.M.; Reyns, P.; Hawkesford, M.J. DeepCount: In-Field Automatic Quantification of Wheat Spikes Using Simple Linear Iterative Clustering and Deep Convolutional Neural Networks. *Front. Plant Sci.* **2019**, *10*, 1176. [CrossRef]
30. Wang, D.; Zhang, D.; Yang, G.; Xu, B.; Luo, Y.; Yang, X. SSRNet: In-field counting wheat ears using multi-stage convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
31. Misra, T.; Arora, A.; Marwaha, S.; Chinnusamy, V.; Rao, A.R.; Jain, R.; Sahoo, R.N.; Ray, M.; Kumar, S.; Raju, D. SpikeSegNet: A deep learning approach utilizing encoder-decoder network with hourglass for spike segmentation and counting in wheat plant from visual imaging. *Plant Methods* **2020**, *16*, 40. [CrossRef]
32. Inbaraj, X.A.; Villavicencio, C.; Macrohon, J.J.; Jeng, J.H.; Hsieh, J.G. Object Identification and Localization Using Grad-CAM++ with Mask Regional Convolution Neural Network. *Electronics* **2021**, *10*, 1541. [CrossRef]