agronomy

MDPI

Article

Improved U-Net for Growth Stage Recognition of In-Field Maize

Tianyu Wan [1,2,3], Yuan Rao [1,2,3,*], Xiu Jin [1,2,3], Fengyi Wang [1,2,3], Tong Zhang [1,2,3], Yali Shu [1,2,3] and Shaowen Li [1,2,3]

[1] College of Information and Computer Science, Anhui Agricultural University, Hefei 230036, China; wantianyu991116@stu.ahau.edu.cn (T.W.); jinxiu123@ahau.edu.cn (X.J.); shwli@ahau.edu.cn (S.L.)
[2] Key Laboratory of Agricultural Sensors, Ministry of Agriculture and Rural Affairs, Hefei 230036, China
[3] Anhui Provincial Key Laboratory of Smart Agricultural Technology and Equipment, Hefei 230036, China
* Correspondence: raoyuan@ahau.edu.cn; Tel.: +86-17756008850

Abstract: Precise recognition of maize growth stages in the field is one of the critical steps in conducting precision irrigation and crop growth evaluation. However, due to the ever-changing environmental factors and maize growth characteristics, traditional recognition methods usually suffer from limitations in recognizing different growth stages. For the purpose of tackling these issues, this study proposed an improved U-net by first using a cascade convolution-based network as the encoder with a strategy for backbone network replacement to optimize feature extraction and reuse. Secondly, three attention mechanism modules have been introduced to upgrade the decoder part of the original U-net, which highlighted critical regions and extracted more discriminative features of maize. Subsequently, a dilation path of the improved U-net was constructed by integrating dilated convolution layers using a multi-scale feature fusion approach to preserve the detailed spatial information of in-field maize. Finally, the improved U-net has been applied to recognize different growth stages of maize in the field. The results clearly demonstrated the superior ability of the improved U-net to precisely segment and recognize maize growth stage from in-field images. Specifically, the semantic segmentation network achieved a mean intersection over union (mIoU) of 94.51% and a mean pixel accuracy (mPA) of 96.93% in recognizing the maize growth stage with only 39.08 MB of parameters. In conclusion, the good trade-offs made in terms of accuracy and parameter number demonstrated that this study could lay a good foundation for implementing accurate maize growth stage recognition and long-term automatic growth monitoring.

Keywords: recognition; growth stage; semantic segmentation; precision agriculture; U-net; dilation path

Citation: Wan, T.; Rao, Y.; Jin, X.; Wang, F.; Zhang, T.; Shu, Y.; Li, S. Improved U-Net for Growth Stage Recognition of In-Field Maize. *Agronomy* **2023**, *13*, 1523. https://doi.org/10.3390/agronomy13061523

Academic Editors: Borja Espejo-García, Spyros Fountas and Georgios Leontidis

Received: 28 April 2023
Revised: 26 May 2023
Accepted: 29 May 2023
Published: 31 May 2023

1. Introduction

Maize is one of the main food crops in the world with a wide range of applications such as feed, energy, food and chemical [1,2]. The growth stage information of field crops is not only crucial basic data for analyzing the relationship between crop growth process and agrometeorological conditions but also holds significant value for multiple facets of precision agriculture [3–5]. Accurate recognition of maize growth stages in the field plays a key role in estimating yields and conducting agricultural activities. Maize needs sufficient nutrients during the seedling stage, appropriate attention should be paid to the work of irrigation during the jointing stage, and the small trumpet stage is susceptible to insect pests. Proper timing of fertilizer, irrigation, cultivation, harvest, insect, weed and disease control can help in significantly improving yields [6]. However, manual recognition of the growth stage of maize is a time-consuming, labor-intensive, subjective and discontinuous process [7]. Furthermore, improper operation during manual recognition can introduce human error and potentially lead to crop damage [8]. Therefore, a recognition method of the maize growth stage is particularly critical in a more efficient, continuous and automatic way.

Deep learning pertains to the domain of machine learning that specifically focuses on the utilization of intricate artificial neural networks (ANNs) characterized by significant depth [9]. Deep learning entails the exploitation of deep network architectures to extract high-level representations from input data, enabling enhanced performance in various learning tasks [10–12]. Nowadays, deep learning has been playing a significant role in the new era of agriculture [13–18]. It offers an automated solution that is both low-cost and lightweight [19]. Semantic segmentation occupied the frontline as a technology that deals with how deep learning could achieve crop phenotypic characteristics from digital images [20]. Through semantic segmentation technology, objects can be effectively separated from their background. Among the most notable and scalable models, U-net stands out with its unique U-shape structure, which consists of an encoder and decoder [21]. The encoder component extracts image features and performs down-sampling, while the decoder restores image resolution to ultimately produce the result [22]. Yu et al. conducted a comprehensive investigation on the potential of the U-net model for segmenting maize tassels. Their study demonstrated exceptional segmentation accuracy even in complex scenarios [23]. Due to the intricacies of agricultural work, some researchers made improvements to the original U-net model. In [24], an Attention feature fusion U-net (AFFU-Net) was proposed to achieve fast and accurate crack segmentation of winter jujubes in complex environments, and it provided guidance for the quality assessment of winter jujubes. Zou et al. reduced the computational burden of the original U-net for segmenting crops and weeds, which achieved precise weeding and reduced herbicide pollution [25]. Moreover, an Enhanced U-net (En-UNet) model was also constructed and trained to segment the rotten portion present in apples. The successful segmentation of previous work based on U-net provided a reference for maize segmentation and growth stage recognition in fields. However, since maize is a natural connectivity crop with slender stems, the color similarity between weeds and maize, and because it grows in a complex environment, it is difficult to recognize the growth stage for conducting irrigation and crop growth evaluation. To address these pending issues, the primary way is to replace the backbone network to improve the feature extraction ability of the U-net on the color, texture and morphology of maize.

Many of the proposed high-performance backbone networks have been applied to implementing the tasks of agricultural image segmentation, detection and classification because of their potential to enhance the feature extraction ability of deep learning models. In recent years, the VGG16 network was used for transfer learning after fine-tuning to identify and classify the seed images [26]. Ayhan and Kwan utilized DeepLabV3+ with Xception to classify forest and grassland [27]. A U-net model with a ResNet backbone was selected for classifying irrigation systems at a regional scale using remote sensing imagery [28]. Roy and Bhaduri proposed a real-time object detection framework, Dense-YOLOv4, by including DenseNet in the YOLOv4 backbone to optimize feature transfer and reuse [29]. Chen et al. used a dual path network as a feature extraction network to extract richer small target semantic features for detecting cherry tomatoes [30]. Wang employed EfficientNet as the backbone of the recognition model, which effectively facilitated data augmentation and recognition performance of practical cucumber leaf diseases [31]. However, since the structure of U-net has no effective mechanism to learn the characteristics of a specific region, it tends to extract features on the whole range of the image rather than focus on specific objects. As a result, a part of the background surrounding the weeds and debris is often inevitably involved when the original U-net is implementing maize growth stage recognition.

In visual tasks of agriculture, attention mechanisms played a positive role in highlighting critical local regions and extracting more discriminative features and the ability of deep learning models. For the purpose of effectively dealing with the common challenge of a complex environment, numerous applications of the attention mechanism were carried out. Gong et al. proposed a model based on Squeeze-and-Excitation (SE) attention so that the model paid attention to the relationship between channels and could automatically learn the importance of different channel features and solve the problem of plant roots segmentation with strong noise in the background [31]. Kang et al. introduced Convolutional

Block Attention Module (CBAM) modules to DeepLabv3+ to recognize the root system of cotton located in dark and closed soil, which put more weight on the segmentation and recognition of fine roots and root hairs, reducing the extraction of soil environment information [32]. Wang et al. integrated Coordinate Attention (CA) block into the Swin Transformer to distinguish the regions in the background where the color and texture are similar to the diseased spots and then applied it to the recognition of practical cucumber leaf diseases [33]. Despite the excellent results of the aforementioned methods, the recognition performance of the maize growth stage was still limited due to the natural connectivity of maize and the long span of the stem.

In the past years, dilated convolution has been proposed for adjusting the reception fields of feature points without decreasing the resolution of feature maps [12]. It can be used as a cascade model [34] and parallel model [35], both of which have a strong ability to preserve spatial information and increase segmentation accuracy. It was widely used recently; S. Chen et al. applied dilated convolution to the backbone network, achieving high segmentation accuracy and efficiency for the filed grape bunches [13]. Ma et al. segmented the winter wheat ears from canopy images by integrating the dilated convolution, significantly improving the segmentation accuracy [36].

To address these issues and enhance the performance of in-field maize growth stage recognition, this study proposed an improved U-net. Different from the original U-net, this study first designed a backbone network replacement strategy to substitute the encoder of the original U-net for enhancing the model's generalization and feature extraction ability. To encourage the network to focus on the important features, coordinate attention modules were introduced to the decoder part. Subsequently, the dilation path was constructed by integrating dilated convolution layers using a multi-scale feature fusion approach, which preserved the detailed spatial information and considered the natural properties of in-field maize. Finally, the adaptability of the model in different environments was verified to outperform the state-of-the-art DeepLabv3+ [35], SegFormer [37], UperNet [38], PSPNet [39], FCN [40] and original U-net in terms of the maize growth stage recognition results on the test set. Other sections in this paper are organized as follows: The dataset and methods are described in Section 2. The experiments conducted using the proposed approach are described in Section 3. Discussions and conclusions are given in Section 4.

2. Materials and Methods

2.1. Data Acquisition

This study focused on recognizing the growth stages of maize, which aimed at growth status monitoring and precision irrigation conducted in smart agriculture. All images used in this study were acquired from Nongcuiyuan Experimental Station in Anhui Agricultural University, which were photographed between 7:30–17:30 from 10 September to 28 October 2021. The experimental station was set up with different cultivars, namely Zhengdan 958, Nongda 108, Jinhai No. 5 and Denghai 3662, each planted in separate plots. In each plot, the maize plants were arranged in six rows with a row spacing of 60 cm and a within-row plant spacing of 25 cm. The planting density was 65,000 plants per hectare. Other field management measures were implemented in accordance with local standards. During image collection, maize samples were randomly selected from the experimental field and captured from one or more of the following angles: horizontal, 45° inclination and vertical for each individual plant. To be more specific, Figure 1 shows the schematic diagram of the image collection process in this study. The experimenter stood at a distance of 0.5 m from the same maize plant and used a mobile phone (HUAWEI or iPhone) to photograph it from three angles: horizontal, 45° inclination and vertical. The phone was about 0.3 m away from the maize plant. This distance was employed to ensure that each image contained a clear maize plant. Furthermore, this employed distance can help in avoiding the incident of damaging the maize plant by the experimenter during data collection.

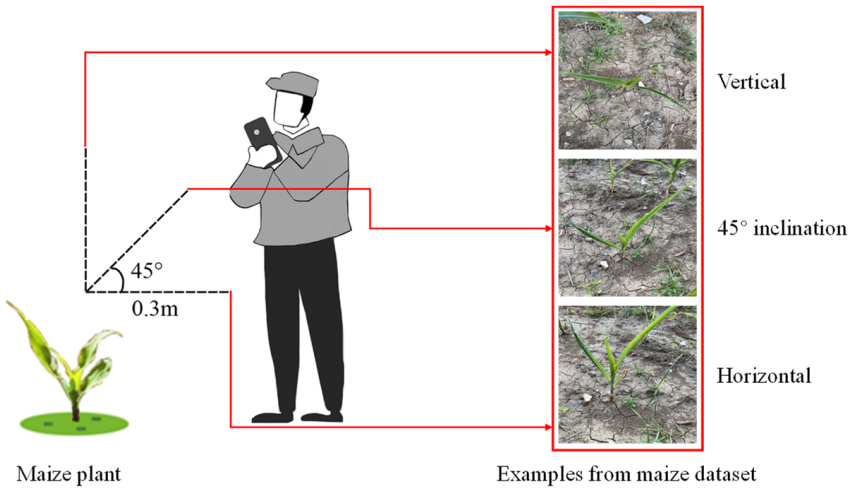


Figure 1. Schematic diagram of the image collection process.

Maize growth stages are commonly recorded using international standards that define both vegetative and reproductive stages [41]. The vegetative stages are further classified into several sub-stages according to number of expanded leaves, including VE (germinal sheath exposed to the ground), V1 (the first leaf fully expanded), V(n) (the nth leaf fully expanded) and VT (tasseling stage), among others [42]. In this study, VE to V5 was designated as the seedling stage [43], V6 to V8 was assigned the jointing stage label [44] and V9 to V12 was classified as the small trumpet stage [45]. The subjects in this study were maize and its seedling stage, jointing stage and small trumpet stage. We conducted three data collection procedures at each growth stage, with 150 images collected at each growth stage, resulting in a total of 450 images being collected.

2.2. Data Preprocessing

We resized the maize images into 512 × 512 pixels to improve the model inference speed. LabelMe software was used to annotate all images manually. Figure 2 shows the annotated images and the corresponding original images, and the annotated images were saved in the format of png. The label color scheme was seedling stage maize RGB = (128,0,0), jointing stage maize RGB = (0,128,0) and small trumpet maize RGB = (128,128,0).

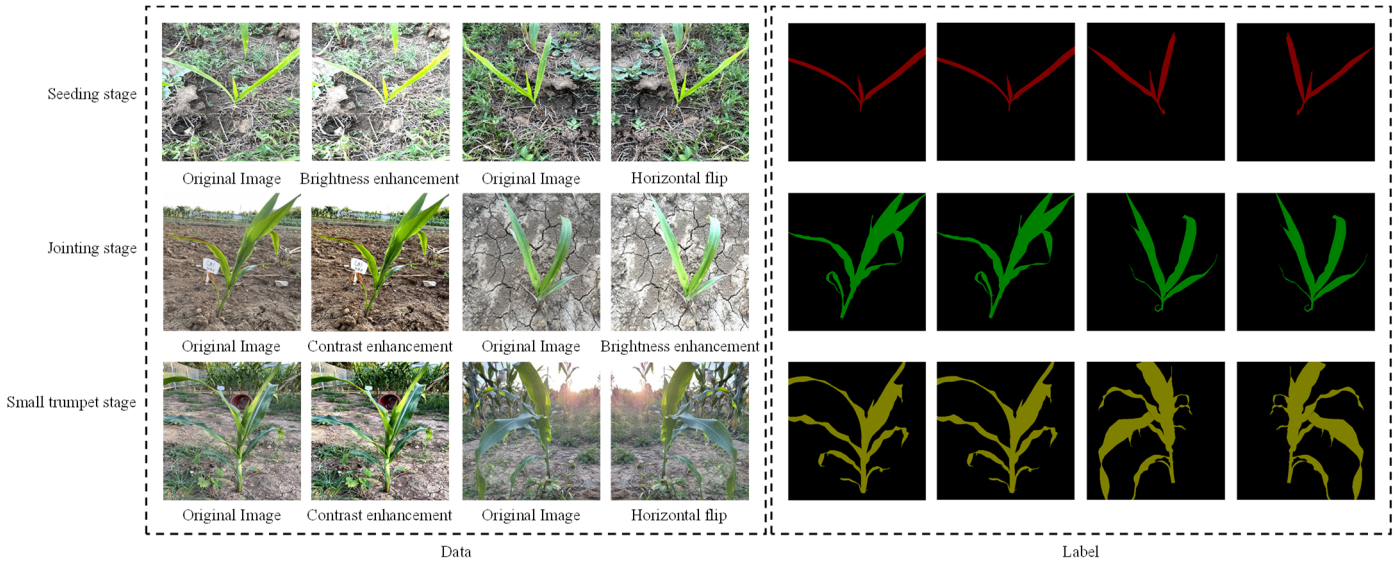


Figure 2. Data annotation and augmentation.

Data augmentation is a proven approach to constructing a robust deep-learning model. [46,47]. In the field of smart agriculture, researchers often use geometric transformation and pixel intensity shift (color shift) for data augmentation [48–50]. As shown in Figure 2, data augmentation processes were randomly performed by horizontal flip, contrast enhancement and brightness enhancement to all 450 images for simulating the effects of different weather and light intensities, with the aim of improving the robustness and the generalization ability of the recognition model. Additionally, the data augmentation method utilized in this study can effectively augment the dataset with minimal cost. The reason for this is that by simply flipping the label horizontally or leaving it unchanged, we are able to generate the corresponding label for generated images. As a result, the number of images was expanded to 900 by means of the aforementioned data augmentation operation. Subsequently, all images were randomly divided into 70%, 20% and 10% as the training set (630 images), validation set (180 images) and test set (90 images). Finally, the maize dataset was generated according to the PASCAL VOC 2012 format.

2.3. Construction and Improvement of U-Net Segmentation Model

2.3.1. Baseline U-Net Model

The original U-net model used the Encoder-Decoder structure, with the encoder on the left and the decoder on the right. The encoder part was divided into five stages according to four max pooling operations. Correspondingly, the decoder part was divided into five stages according to four up-sampling operations as well. The encoder part consisted of the repeated two convolution layers, each followed by a rectified linear unit (ReLU) and a max-pooling layer for down-sampling. Each stage of the decoder included the up-sampling process of the feature map, which was matched and fused with the feature map from the corresponding stage of the encoder part. The shallower high-resolution layers in the U-net network were used to solve the problem of pixel positioning, while the deeper layers were used to solve the problem of pixel classification. This model was widely used in image segmentation because it was trained in an end-to-end way and performed well with limited amounts of data on a small dataset [25].

However, the in-field maize images acquired by smartphones were severely affected by lighting variations, complex backgrounds, image noise and so on [51]. In this study, one semantic segmentation model for maize growth stage based on the improved U-net model was constructed by optimizing the existing U-net model to tackle such challenging problems. Compared with the original U-net, three main improvements were made in the encoder and the decoder: (1) the feature extraction ability of the U-net was enhanced by replacing an outperformed backbone network loading pre-trained weights on the ImageNet dataset according to our replacement strategy; (2) three coordinate attention blocks were embedded in the decoder after the skip connection of stages 2, 3 and 4; and (3) the dilation path was constructed by integrating four sub-paths and each sub-path consisting of many dilated convolution layers, which preserved the detailed spatial information and considered natural properties of in-field maize. In summary, the improved U-net comprehensively considered the in-field environment and natural maize characteristics, enhancing the performance of the model to recognize the maize growth stage in the field. The motivations and details regarding the above three modifications were presented in detail.

2.3.2. Backbone Network Replacement Strategy

Due to the original aim for biomedical image segmentation, the encoder of the U-net loaded with pre-trained weights on the medical dataset was not suitable for the recognition of maize growth stage. Different from the single feature of the medical dataset, the in-field image of maize was usually affected by factors such as background, light and maize posture, which inherently led to the strong requirements of feature learning ability of the backbone network. Hence, the backbone networks played a crucial role during the processing of maize growth stage recognition, which resulted in the replacement operation of the backbone with strong feature extraction ability. Additionally, in our case, backbone

networks were initialized with pre-trained weights of ImageNet datasets for improving model performance on semantic segmentation [52].

Previous studies showed that many backbone networks, such as VGG, Xception, DPN, DenseNet, ResNet and EfficientNet, were worth considering for the task in this paper [26–30,33]. However, the U-net architecture offered a unique property (i.e., skip connections), which set it apart from other encoder-decoder models. At each spatial resolution level of the U-net model, the skip connection copies the activation outputs to the same resolution level in the decoder [28]. Therefore, the number of skip connections was equal to the number of down-sampling and up-sampling. As we know, U-net had four skip connections, corresponding to four pairs of up-sampling and down-sampling operations. This means that each backbone network needed to be divided into five stages according to the down-sampling operation for fitting into U-net. However, ResNet, DenseNet, EfficientNet, Xception and DPN had five down-sampling operations; the division according to the down-sampling operations might result in six stages. Therefore, these aforementioned backbone networks faced a common problem, which led to the inability to be directly used as feature extraction networks of the original U-net.

For the purpose of enhancing the feature extraction ability, we divided six backbone networks into two categories according to the number of down-sampling operations and proposed a replacement strategy to employ them to the encoder of the original U-net. In Figure 3, the replacement strategy is divided into three distinct parts: A, B and C. These three parts are, respectively, referred to as the input and pre-operations stage, the center stage and the output and post-operation stage [12]. Each part of the replacement strategy will be described in the following paragraph.

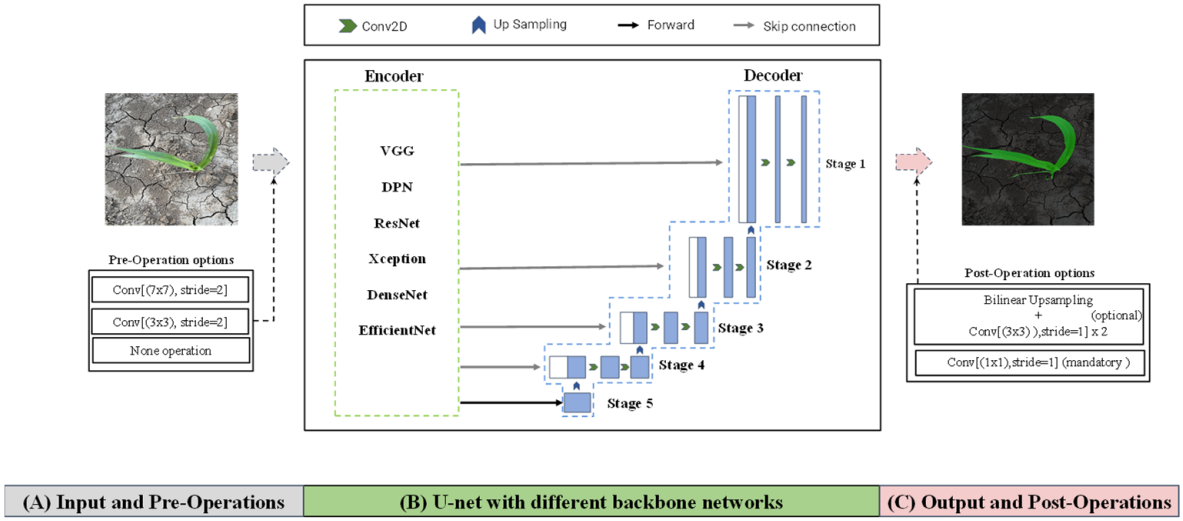


Figure 3. The replacement strategy for different backbone networks.

ResNet, DPN, DenseNet and Xception, EfficientNet used 7 × 7 convolution operation and 3 × 3 convolution operation for the first down-sampling. Therefore, as shown in part (A) of Figure 3, we defined the first down-sampling operation of these five backbone networks as Pre-Operation, including three options. To achieve consistency between the input and output feature map resolutions, we introduced a Post-Operation step in part (C) of our approach. Specifically, this step adjusts the feature map resolution following the decoder output of the U-net [53]. The Post-Operation step contained two options; the final feature vector mapping was mandatory, but the other one was optional. When the number of down-sampling operations of the backbone network was greater than four, the Post-Operation part used bilinear up-sampling and double 3 × 3 convolution operations, which was similar to the decoder of the U-net, restoring the resolution of the feature map to 512 × 512. At the mandatory option of the Post-Operation part, a 1 × 1 convolution was employed to map feature vector to four [54], which indicated the seedling stage, jointing

stage, small trumpet stage and background in this study. The U-net architecture, depicted in Part (B), omits the Pre-Operation and divides the remaining backbone network into five Encoder-Blocks based on four max pooling operations [55]. As a result, backbone networks were able to be employed as the encoder for enhancing the feature extraction ability of the original U-net. In summary, the proposed strategy in this study could allow the feature maps generated by each stage in backbone networks to be concatenated with the corresponding decoder stage using a skip connection. As a consequence, all of the aforementioned six different backbone networks were adapted to the U-net and did not change the original U-net structure.

2.3.3. Coordinate Attention Module

The feature maps obtained from the backbone network were transmitted to the decoder to restore the original resolution and segment image. The objects of this study were maize plants in the field, which had strongly different characteristics due to the surrounding background. Extensive research has been devoted to attention mechanisms, which have been widely implemented to enhance the performance of contemporary deep neural networks [56,57]. Therefore, the coordinate attention mechanism was added to the decoder to augment the feature representations of the model for the recognition of maize growth stage.

Figure 4 demonstrates that the coordinate attention mechanism serves as a computational unit with the objective of boosting the expressive capacity of network features [58]. By incorporating positional information into channel attention, this technique enhances model performance by allowing networks to attend to interest regions without incurring excessive computational costs [59].

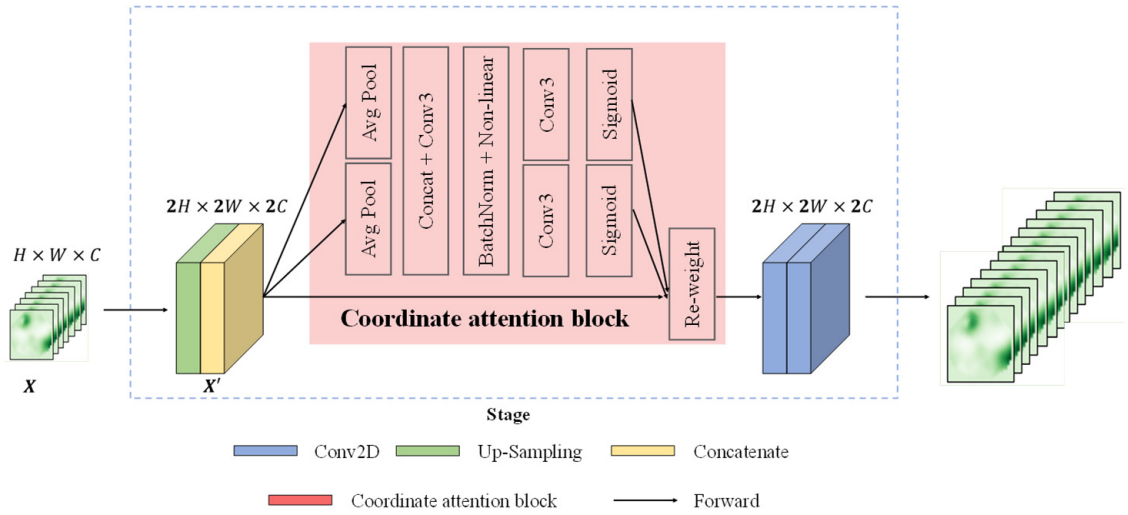


Figure 4. A stage in the decoder with coordinate attention mechanism.

For the incoming feature map X, firstly, the size of X by up-sampling operation was doubled. Subsequently, the result of the up-sampling operation was contacted with the feature map X′ output from corresponding stage in the encoder. Then, it was transferred to the CA block for processing to obtain G, which allowed our improved U-net to locate the exact position of the object of interest more accurately and hence helped the whole model to recognize maize growth stage better. Finally, the output feature maps G were processed by using two convolution layers with 3 × 3 kernels. The operation of the stage of the decoder can be formulated as follows:

$$G = CA(Concat(X', Up(X))), X \in R^{(H,W,C)}, X' \in R^{(2H,2W,C)}, G \in R^{(2H,2W,2C)} \quad (1)$$

$$G' = DConv(G),\ G = [G^1; G^2; G^3; \ldots; G^d] \quad (2)$$

where CA denoted the operation of CA block. *Concat* represented the processing of copy and crop in U-net. *DConv* was an operation consisting of two convolution layers with 3×3 kernels. The G' indicated output features of a stage in decoder.

2.3.4. The Dilation Path

Due to the maize's narrow branching, plant connectivity, background complexity and similarity to weeds, it is not only important to increase the receptive field of feature points of the semantic segmentation model but also to keep the detailed information. The receptive field of feature points could be multiply increased by using pooling layers. However, pooling layers may reduce the resolution of center feature maps and drop spatial information. Inspired by Zhou et al. [12], dilated convolution layer based on multi-resolution features could be a desirable alternative to the pooling layer.

We constructed the dilation path to connect the encoder part and the decoder part and the structure, as shown in Figure 5. The dilation path composited of four sub-paths, from top to bottom, named $\{P_1, P_2, P_3, P_4\}$, and each path consists of stacked dilated convolution layers with different dilation rates. The dilation rates of the top path were set to 1, 2 and 4, respectively. From top to bottom path, the number of dilated convolution layers decreased sequentially, causing the receptive field of each path to be different. Therefore, the semantic segmentation model could combine features from different scales by adding results of all paths. In this way, the receptive field was increased by cascading dilated convolution layers, and the advantage of multi-resolution features was taken by the parallel mode of the dilation path.

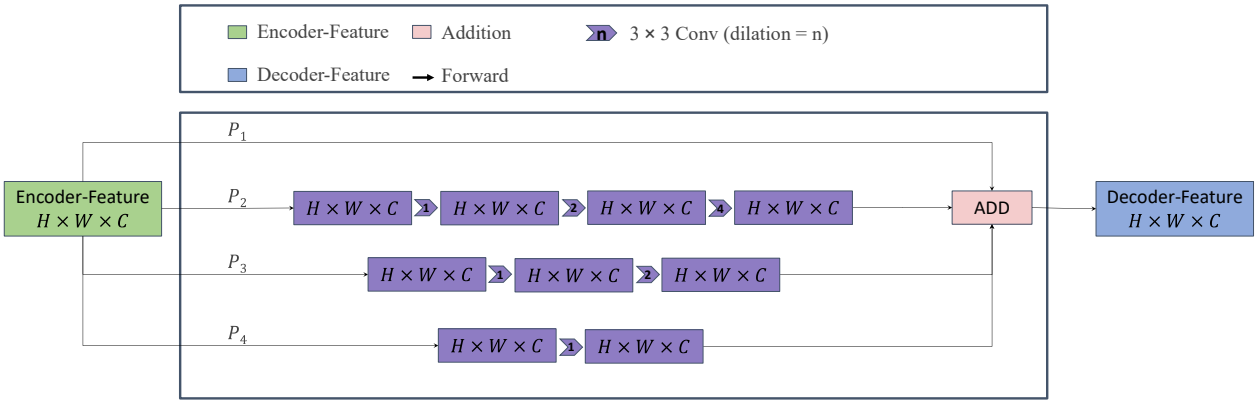


Figure 5. The structure of the dilation path.

2.3.5. The Improved U-Net

The structure of the improved U-net model for recognizing in-field maize growth stage is shown in Figure 6, with the same encoder and decoder branches as original U-net. The structure consisted of three parts: a backbone network with outperformed feature extraction ability, a dilation path fusing multi-scale features and a decoder embedded with three coordinate attention modules. The former different backbone networks were divided into five stages according to our replacement strategy and represented by five encoder blocks; each stage consisted of an encoder block and a down-sampling operation. The feature maps obtained from the backbone network were transmitted to the decoder through the dilation path to preserve the detailed spatial information of in-field maize. The latter three coordinate attention blocks were embedded in the decoder after the skip connection of stages 2, 3 and 4. In this way, accurate coordinate information and channel relationships could be obtained, allowing the improved U-net to acquire knowledge over a larger area and enhance feature information.

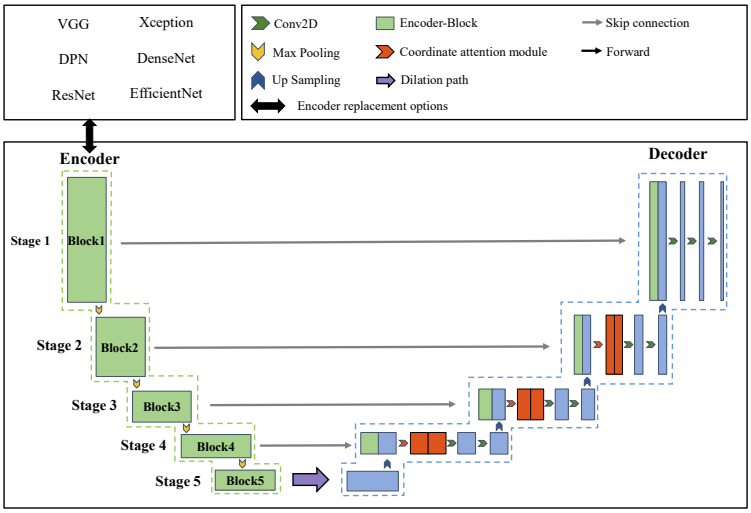


Figure 6. The structure of the improved U-net for maize growth stage recognition.

2.4. Experimental Platform and Training Strategy

Transfer learning aims at eliminating the need for a vast number of labeled datasets by transferring the weights of a pre-trained network and fine-tuning it for another task using new image datasets. Fine-tuning the whole network with transfer learning is generally much faster and easier than training the network from scratch with randomly initialized weights [60].

As shown in Figure 7, the training process was divided into two stages: the freezing stage and the unfreezing stage. The backbone networks were pre-trained on the ImageNet dataset to achieve a relatively optimal parameter space. In this study, representative hyperparameters batch size, learning rate and epoch were considered when training the improved U-net. In the freezing stage, all parameters of the backbone network were frozen, and the decoder of the improved U-net was trained. Adjusting fewer model parameters in the freezing stage and the learning rate was set slightly larger to jump out of the optimal local solution. Therefore, the learning rate was set to 1×10^{-4}. The batch size of training was set to 4, and epochs were set to 100. In the unfreezing stage, all parameters of the backbone networks were unfrozen and participated in training. The learning rate was set smaller to ensure the stability of the model training. The learning rate of the unfreezing stage was set to $1 \times 10^{-5,}$ and other parameters were the same as in the freezing stage. In this way, pre-trained backbone networks could be better adapted to maize datasets for potentially achieving good performance and reducing the data demand.

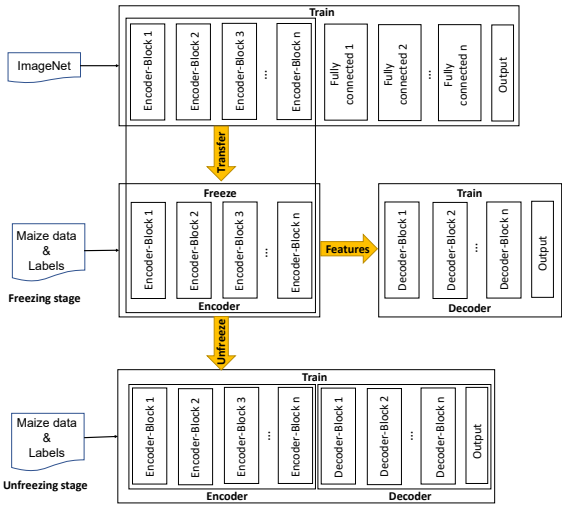


Figure 7. Training strategy of transfer learning.

All models used in this study were compiled with GPU support. All experimental studies were conducted on a Conda environment on Ubuntu Desktop 18.04.4 LTS operating system running on Intel(R) Xeon (R) Gold 5118 CPU @2.30GHz CPU and Nvidia GeForce RTX 2080Ti 11GB GPU.

2.5. Evaluation of the Performance of Semantic Segmentation Model

In this study, three standard semantic segmentation metrics were used to evaluate the segmentation results [61]. The overall intersection over union (IoU), mIoU, pixel accuracy (PA) and mPA were used to assess and compare the segmentation performance (Equations (3)–(6)). The IoU and PA were averaged over all images in the test set to obtain mIoU and mPA, respectively. Intuitively, the mIoU calculates the degree of spatial alignment between segmentation results and ground truth, while the mPA measures the number of correctly classified pixels relative to the total number of pixels [51]. The model parameters were also measured to assess the complexity of segmentation network.

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \times 100\% \tag{3}$$

$$mPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \times 100\% \tag{4}$$

$$IoU = \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \times 100\% \tag{5}$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \times 100\% \tag{6}$$

where p_{ii} was the number of pixels that were actually in category I and were predicted to be in category I. p_{ij} was the number of pixels actually in category I but were predicted to be in category J. p_{ji} was the number of pixels actually in category J but were predicted to be in category I. k indicated the number of different categories in the dataset, which value was 4 in this study.

3. Results and Discussion

3.1. Comparison with Different Network Architectures

To improve the feature extraction ability of the original U-net, the replacement strategy (see Section 2.3.2) was used to employ different backbone networks and the training results of U-net-VGG16, U-net-DenseNet, U-net-DPN92, U-net-EfficientNet-B5, U-net-ResNet34, U-net-Xception and the original U-net were compared. The training set was applied to conduct the training of maize growth stage recognition based on the above network architectures. Additionally, the test set of 90 images was used to assess the performance of the different backbone network architectures.

The loss of the original U-net with different backbone network compositions on the maize image train data set in the field is shown in Figure 8. It could be seen that with the increase in training iterations, the losses of the train set kept gradually decreasing from the start and achieved stability in the end. Obviously, the training loss of the original U-net was the highest and reached saturation around 45 epochs. In contrast, the losses of the rest encoders reached saturation at a faster rate at approximately 30 epochs, which was attributed to the overall learning distinguishability of features learned by the network architectures used in the encoder. It is noteworthy that U-net-VGG16 demonstrated superior performance compared to other models, as evidenced by the convergence of the loss curve at approximately 20 epochs. This observation suggested that the feature extraction network plays a crucial role in enhancing the ability of semantic segmentation for maize growth stage recognition [62]. Moreover, the cascade convolutional structure of U-net-VGG16 exhibits higher efficiency and stability during the training process [47,63].

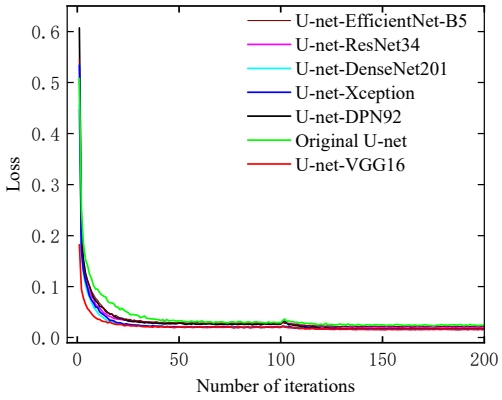
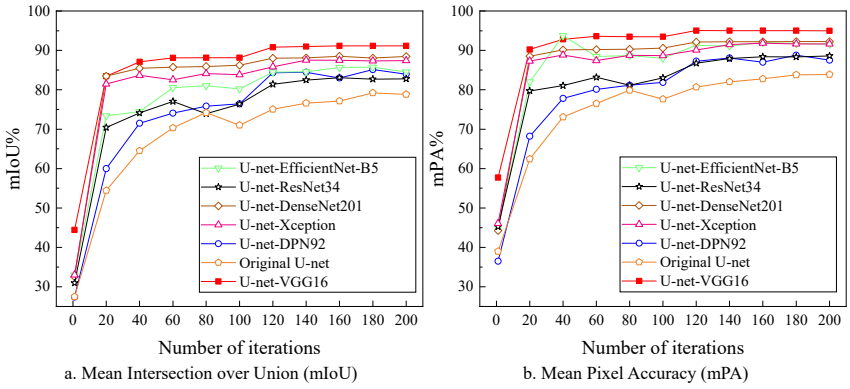


Figure 8. Loss curve of different backbone networks.

Subsequently, the models were verified once every 20 iterations during training and the validation results of the different encoders were analyzed, as shown in Figure 9. In general, significant improvement of the mIoU and mPA among the encoders was found. It could be seen that the U-net-ResNet34, U-net-DPN92 and original U-net generally kept an upward trend until the peak was reached, in spite of some small fluctuations. For the rest of the encoders, the trend of mIoU and mPA curves was almost identical. Specifically, with the increase of training iterations, the mIoU and mPA gradually increased. The U-net-VGG16 model exhibited exceptional performance in terms of mIoU. However, it is important to acknowledge that U-net-EfficientNet-B5 outperformed U-net-VGG16 in terms of mPA at 40 iterations. This discrepancy can be attributed to the limited number of training iterations, which resulted in the instability of the training weights. Nonetheless, it is crucial to recognize the overall advantages of U-net-VGG16, as this minor deviation should not overshadow its strengths. As a result, the U-net-VGG16 model exhibited the highest performance in terms of mPA among the backbone networks. This finding further underscored the advantage of employing cascade convolution in the recognition of the maize growth stage.



a. Mean Intersection over Union (mIoU) b. Mean Pixel Accuracy (mPA)

Figure 9. Metrics of different backbone networks.

The mIoU, mPA and parameters of the maize growth stage recognition model with different encoders on the test set were presented in Table 1. It could be observed U-net-Densenet201, U-net-Xception and U-net-VGG16 achieved mIoU and mPA both of over 85%, and they reached high accuracy in the testing phase. Among them, U-net-VGG16 performed best among all models, and the number of parameters in U-net-VGG16 was 24.89 M, which was lower than the average of several encoders. Moreover, compared with the original U-net, the U-net-VGG16 increased in mIoU by 10.89% and 8.1% in mPA, which proved the effectiveness of U-net-VGG16 in improving the ability of feature extraction of the maize. For the maize dataset used in this study, the position of maize in the image was obvious but with the interference of environmental factors. VGG16 enhanced accuracy through

the utilization of smaller convolution kernels, appropriate network depth, and a reduced parameter count [64]. These design choices contributed to the stability of feature extraction for maize images and alleviated the computational burden during network training [65]. As the results mentioned before, when using the CNN model to extract confused features of images through transfer learning, the VGG network performed better than convolutional networks represented by residual and inception structures [66]. The U-net-VGG16 network basically met the requirements of primary segmentation of the recognition of the maize growth stage. Therefore, the VGG16 was employed as the backbone network of the maize growth stage recognition model, and the U-net-VGG16 was used as a further improved baseline model.

Table 1. The mIoU and mPA of different backbone networks.

Backbone	mIoU (%)	mPA (%)	Parameters (M)
U-net-Densenet201	88.43	92.35	28.58
U-net-Resnet34	82.84	88.18	24.43
U-net-Xception	87.86	91.82	28.76
U-net-VGG16	**91.15**	**94.99**	**24.89**
U-net-EfficientNet-B5	84.53	91.6	31.21
U-net-DPN-92	84.52	87.98	46.95
Original-U-net	80.26	86.89	38.02

3.2. Performance Comparison of Different Attention Mechanisms

Based on the U-net-VGG16 model employed by the backbone network replacement strategy, a comparative experiment was conducted to assess the performance of maize growth stage recognition models using different attention mechanisms (CA, CBAM and SE). In this study, we embedded attention mechanisms after skipping the connection of stages 2, 3 and 4 of the decoder and analyzed its impact on the recognition results. The trends of training loss over 200 epochs for all three attention blocks are illustrated in Figure 10. As inferred from Figure 10, the loss values decreased as the number of epochs increased and were generally stable when the number of epochs reached 25, which indicated that all models achieved low loss values in the training phase. Note that the loss of the U-net-VGG16-CA network was the lowest among all three attention blocks, and the model quickly reached its optimum point.

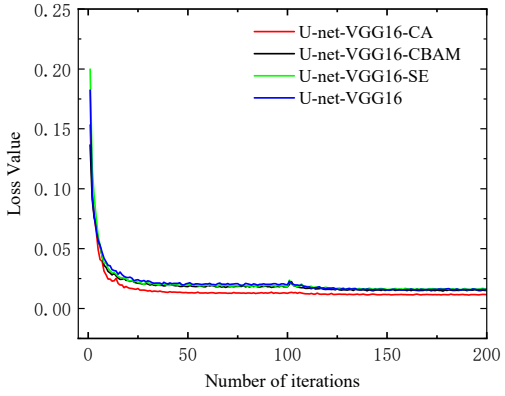


Figure 10. Loss curve of different attention mechanisms.

After completion of the training processes, the experiment results in the validation set were obtained, as shown in Figure 11. All models could recognize the growth stage of maize with the mIoU and the mPA, both of around 90%. The performance of segmentation localization can be enhanced through the utilization of attention mechanisms when compared to the U-net-VGG16 model. However, both U-net-VGG16-SE and U-net-VGG16-CBAM demonstrated comparable mPA values to that of U-net-VGG16, indicating that the integration of SE and CBAM attention blocks did not contribute to the enhancement of

pixel classification accuracy in the model. In summary, U-net-VGG16-CA generally offered the best localization ability and growth stage recognition accuracy among the three models.

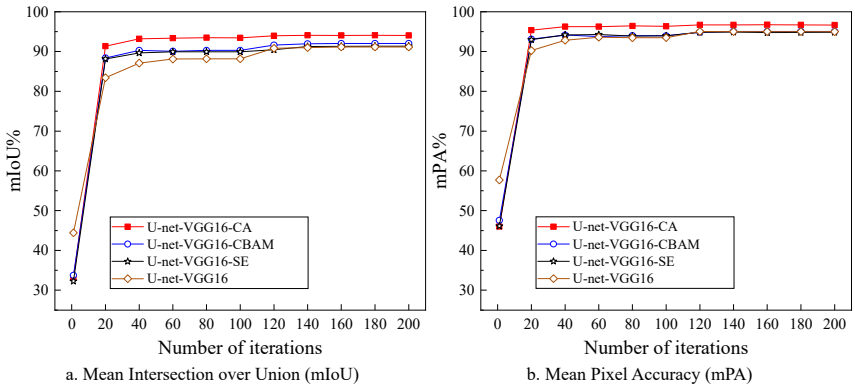


Figure 11. Metrics of different attention mechanisms.

The results from the different attention mechanisms are summarized in Table 2. It could be found from Table 2 that the U-net-VGG16 with CA reached the mIoU (94.11%) and mPA (96.8%) on the test dataset, which of both was higher than that of CBAM and SE. Compared with the U-net-VGG16, the U-net-VGG16-CA had a 2.96% increase in mIoU and a 1.81% increase in mPA, which proved the effectiveness of coordinate attention block in improving the recognition of the maize growth stage. Moreover, with the integration of coordinate attention blocks, there was a mere increase of 0.03M in the model parameters. This indicated that the performance of the U-net-VGG16 model was enhanced with little additional training and inference overhead. Compared to SE and CBAM attention mechanisms, the CA mechanism could increase the model applicability to recognize the growth stage of maize, and other models did not hold a satisfying performance based on the result. It was because the coordinate attention embedded positional information into channel attention to enable the maize growth stage recognition model to focus on the maize-connected area [67], which enhanced the representation of maize location information while suppressing feature extraction of background area in images [68]. In summary, the U-net-VGG16 model introducing the CA mechanism yielded the highest mIoU and mPA, providing empirical validation for the effectiveness of integrating spatial and channel dimensions within the attention mechanism for maize growth stage recognition [69].

Table 2. Comparisons of different attention methods when taking VGG16 as the baseline.

Model	mIoU (%)	mPA (%)	Parameters (M)
U-net-VGG16	91.15	94.99	24.89
+CBAM	$92.01_{+0.86}$	$95.05_{+0.06}$	$24.97_{+0.08}$
+SE	$93.22_{+2.07}$	$96.19_{+1.2}$	$24.93_{+0.04}$
+CA	$\mathbf{94.11_{+2.96}}$	$\mathbf{96.8_{+1.81}}$	$\mathbf{24.92_{+0.03}}$

3.3. Effects of Multi-Scale Dilation Path

In this study, the dilation path of the improved U-net was constructed by integrating four sub-paths, named {P_1, P_2, P_3, P_4}. While the utilization of dilated convolution was advantageous, it resulted in an increase in the number of model parameters [70]. Therefore, we performed ablation experiments to verify the effectiveness of the proposed dilation path. Additionally, the balance between accuracy and the number of parameters was discussed.

As shown in Figure 5, the dilation rates of P_2 were set to 1, 2 and 4, respectively. In this section, P_2 was used as a basis to ensure the receptive field and discussed the effect of different structures on the dilation path. The experiment results of different dilation path structures on the testing set are shown in Table 3. It could be seen that expanding the receptive field using a dilation path could really enhance the model's ability to preserve

spatial information and improve segmentation accuracy [71]. However, the achieved improvements by {P_1, P_2} and {P_1, P_2, P_3} were very limited because of the semantic gap between feature maps with different scales of the receptive field [72]. In addition, the {P_1, P_2, P_3} increased the number of parameters, while mIoU and mPA did not effectively improve. The maize growth stage recognition performance of the improved U-net was the best among the different dilation path structures and U-net-VGG16-CA. Although the number of parameters was increased, a considerable improvement in maize growth stage recognition was achieved. Finally, we named the improved U-net for the U-net-VGG16-CA model with the dilation path.

Table 3. Comparisons of different dilation path structures.

Model	mIoU%	mPA%	Parameters (M)
U-net-VGG16-CA	94.11	96.8	24.89
+{P_1, P_2}	$94.21_{+0.1}$	$96.87_{+0.07}$	$32.00_{+7.11}$
+{P_1, P_2, P_3}	$94.20_{+0.03}$	$96.85_{+0.02}$	$36.72_{+11.83}$
+{P_1, P_2, P_3, P_4}(improved U-net)	**$94.51_{+0.4}$**	**$96.93_{+0.13}$**	$39.08_{+14.19}$

3.4. Comparison with Other Semantic Segmentation Models

The performance of the improved U-net model was evaluated through a comparative analysis with state-of-the-art semantic segmentation models, including FCN, PSPNet, UperNet and DeepLabv3+ and PSPNet, as well as the original U-net. Table 4 shows the comparative results based on performance metrics mIoU, mPA and the number of parameters. Moreover, the prediction results of different segmentation models are shown in Figure 12, and the differences between the effect of each model were highlighted with red rectangles.

Table 4. Metrics of different semantic segmentation models.

Models	mIoU (%)	mPA (%)	Parameters (M)
Improved U-net	**94.51**	**96.93**	39.08
DeepLabv3+	89.97	93.31	54.71
SegFormer	83.06	86.98	39.89
Original U-net	80.26	86.89	**38.02**
UperNet	78.35	88.87	126.07
PSPNet	74.19	84.44	49.07
FCN	72.61	81.91	134.28

The results from Table 4 revealed that both FCN and PSPNet exhibited the worst performance, with their mIoU values falling below 75%. Figure 12h illustrates their deficiency in segmenting densely distributed and intertwined maize plants, leading to the significant omission of maize regions and severely impacting the model's effectiveness. Moreover, Figure 12d demonstrated the difficulty of PSPNet dealing with seeding maize images that contained a lot of weeds in the background. UperNet and original U-net also exhibited subpar performance, as evidenced by a significant number of misclassified pixels, even in scenarios with a simple background (refer to Figure 12a,c). The SegFormer model achieved an mIoU of 83.06% and an mPA of 86.98%, demonstrating its superior accuracy in maize morphology segmentation and indicating fewer misclassification errors compared to the previous four models. However, Figure 12c revealed the presence of certain misclassifications, suggesting that further improvement is needed in the classification accuracy. Additionally, Figure 12a,b indicated that there was still much room for improvement in the delineation of maize segmentation results. We noticed that the DeepLabv3+ obtained a good result, even though it did not make any improvements. However, this architecture exhibited certain limitations, the most prominent being its increased complexity compared to other networks such as DeepLabv3+. Although DeepLabv3+ demonstrated superior

segmentation results and excelled in handling complex segmentation tasks, it necessitated a larger volume of training data to achieve optimal performance [25]. Furthermore, as depicted in Figure 12f, all previous models, including DeepLabv3+, exhibited limited robustness to variations in lighting conditions. Specifically, under strong illumination, these models demonstrated suboptimal performance in segmenting maize plant leaves.

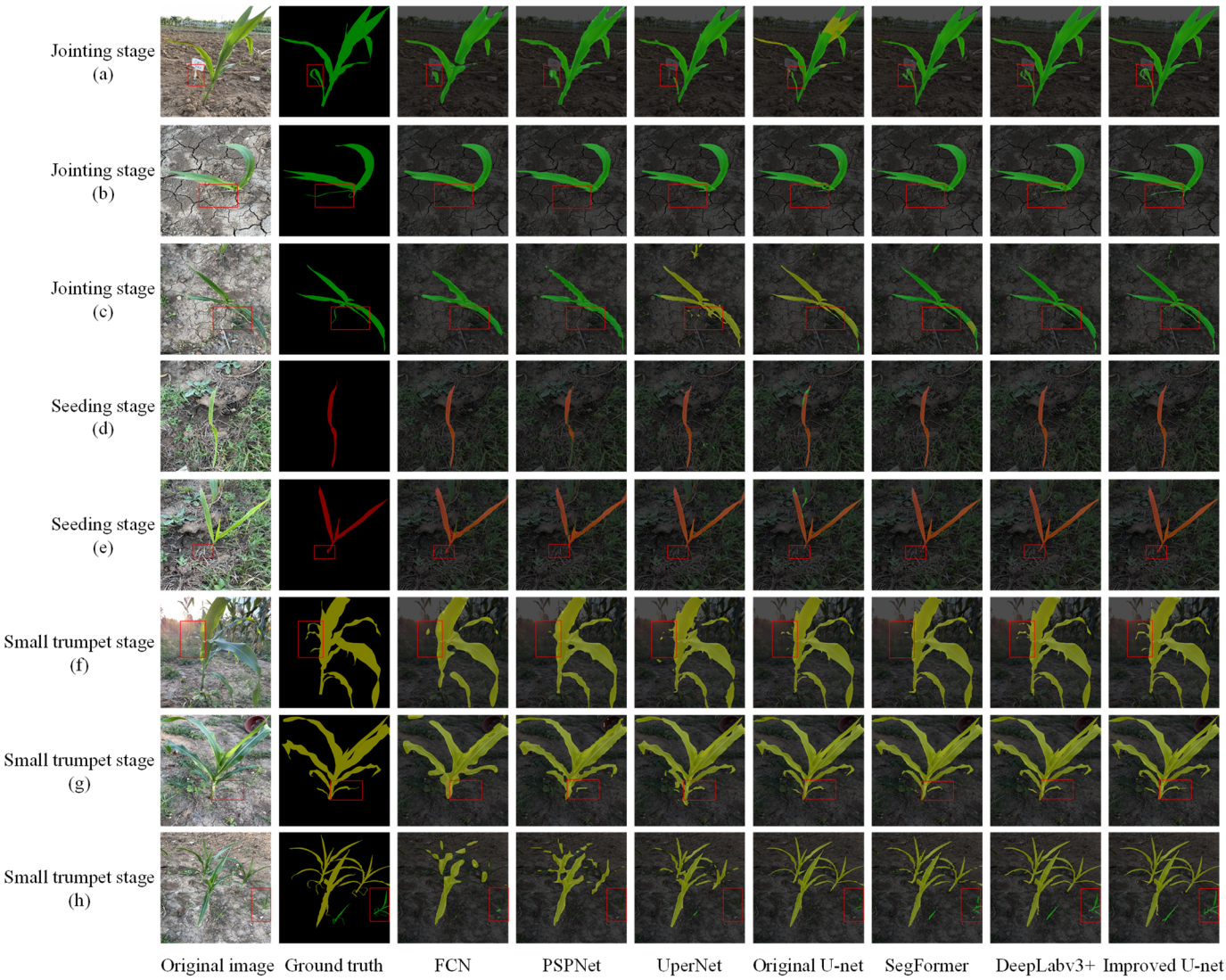


Figure 12. The results of different models.

As for the improved U-net, the mIoU and mPA were 94.51% and 96.93%, respectively, and both were better than the other six models. The mIoU and mPA increased by 4.54%, by 3.62%, respectively, compared with DeepLabv3+. Meanwhile, the simple network structure also made a small number of parameters for improved U-net [73], which was 39.08 MB. As depicted in Figure 12, it became evident that the improved U-net model effectively achieved precise maize segmentation across diverse backgrounds, exhibiting an accurate classification of nearly all pixels. This notable achievement serves to underscore the model's efficacy in accurately recognizing maize growth stages. It could be observed that the segmentation accuracy of the original U-net model showed significant improvement due to the stable and efficient feature extraction ability of the cascaded convolutional structure [74]; the upgraded decoder with embedded CA module that focuses on connected regions while suppressing background feature expression [75]; and the dilation path that effectively preserves spatial information [76]. This finding indicated that the improved U-net model achieved optimal ac-

curacy in recognition of the maize growth stage and obtained a satisfactory result, successfully recognizing the maize growth stage in most cases.

4. Conclusions

Due to the characteristics of the field environment and natural properties of maize, the original U-net was unable to efficiently extract maize features because of the lack of efficiency in utilizing channel and spatial information. Therefore, the performance of U-net was still limited when dealing with the recognition of the maize growth stage. Based on the Encoder-Decoder architecture, this study proposed the improved U-net by means of enhancing feature extraction ability and optimizing the model representation of channel and spatial information. Through experiments, it was demonstrated that the backbone networks with different structures had great potential for optimizing maize extraction features in maize growth stage recognition. The upgraded decoder, integrated with coordinate attention modules, had a good ability to focus on maize-connected areas by means of reducing the interference of complex backgrounds. Moreover, the dilation path further contributed to the improvement of maize growth stage recognition performance by fusing preserved spatial information. It was found that the improved U-net could not only effectively and accurately recognize the growth stage of maize with different sizes, maize with interlaced and uneven illuminations on the leaves but also outperforms other state-of-the-art segmentation models under different conditions, demonstrating a high degree of robustness to illumination, weeds and debris.

In the future, we will collect different varieties and growth stages of maize images under various conditions to expand the dataset and explore the methods to further simplify the network structure and improve the segmentation accuracy.

Author Contributions: Conceptualization, Y.R. and X.J.; data curation, T.W., Y.R. and X.J.; formal analysis, Y.R.; funding acquisition, Y.R.; investigation, T.W., Y.R. and X.J.; methodology, T.W., Y.R., X.J., F.W., T.Z., S.L. and Y.S.; project administration, Y.R. and X.J.; resources, Y.R., F.W., T.Z., S.L. and Y.S.; software, T.W., Y.R. and X.J.; supervision, Y.R.; validation, Y.R., X.J., F.W., T.Z., S.L. and Y.S.; visualization, T.W., Y.R., X.J., F.W., T.Z., S.L. and Y.S.; writing—original draft, T.W.; writing—review and editing, Y.R. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: The study was funded by the Natural Science Research Key Project of Anhui Provincial University (No. 2022AH040125), the Anhui Provincial Key Laboratory of Smart Agricultural Technology and Equipment (No. APKLSATE2021X004), the Natural Science Foundation of Anhui Province (No. 2008085MF203) and the Key Research and Development Plan of Anhui Province (No. 201904a06020056 & 202104a06020012 & 202204c06020022).

Data Availability Statement: Given that the data used in this study were self-collected, the dataset is being further improved. Thus, the dataset is unavailable at present.

Acknowledgments: The authors would like to thank the research farms that participated in this study for providing the study environment.

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

References

1. Long, N.V.; Assefa, Y.; Schwalbert, R.; Ciampitti, I.A. Maize Yield and Planting Date Relationship: A Synthesis-Analysis for US High-Yielding Contest-Winner and Field Research Data. *Front. Plant Sci.* **2017**, *8*, 2106. [CrossRef] [PubMed]
2. Shu, M.; Zhou, L.; Gu, X.; Ma, Y.; Sun, Q.; Yang, G.; Zhou, C. Monitoring of Maize Lodging Using Multi-Temporal Sentinel-1 SAR Data. *Adv. Space Res.* **2020**, *65*, 470–480. [CrossRef]
3. Bannayan, M.; Sanjani, S. Weather Conditions Associated with Irrigated Crops in an Arid and Semi Arid Environment. *Agric. For. Meteorol.* **2011**, *151*, 1589–1598. [CrossRef]
4. Kherif, O.; Keskes, M.I.; Pansu, M.; Ouaret, W.; Rebouh, Y.-N.; Dokukin, P.; Kucher, D.; Latati, M. Agroecological Modeling of Nitrogen and Carbon Transfers between Decomposer Micro-Organisms, Plant Symbionts, Soil and Atmosphere in an Intercropping System. *Ecol. Model.* **2021**, *440*, 109390. [CrossRef]
5. Latati, M.; Dokukin, P.; Aouiche, A.; Rebouh, N.Y.; Takouachet, R.; Hafnaoui, E.; Hamdani, F.Z.; Bacha, F.; Ounane, S.M. Species Interactions Improve Above-Ground Biomass and Land Use Efficiency in Intercropped Wheat and Chickpea under Low Soil Inputs. *Agronomy* **2019**, *9*, 765. [CrossRef]

6. Li, Q.; Dong, B.; Qiao, Y.; Liu, M.; Zhang, J. Root Growth, Available Soil Water, and Water-Use Efficiency of Winter Wheat under Different Irrigation Regimes Applied at Different Growth Stages in North China. *Agric. Water Manag.* **2010**, *97*, 1676–1682. [CrossRef]
7. Omari, M.K.; Lee, J.; Faqeerzada, M.A.; Joshi, R.; Cho, B.-K. Digital Image-Based Plant Phenotyping: A Review. *Korean J. Agric. Sci.* **2020**, *47*, 119–130.
8. Yu, Z.; Cao, Z.; Wu, X.; Bai, X.; Qin, Y.; Zhuo, W.; Xiao, Y.; Zhang, X.; Xue, H. Automatic Image-Based Detection Technology for Two Critical Growth Stages of Maize: Emergence and Three-Leaf Stage. *Agric. For. Meteorol.* **2013**, *174–175*, 65–84. [CrossRef]
9. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
10. Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; 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. [CrossRef]
12. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE: Salt Lake City, UT, USA, 2018; pp. 182–186.
13. Chen, S.; Song, Y.; Su, J.; Fang, Y.; Shen, L.; Mi, Z.; Su, B. Segmentation of Field Grape Bunches via an Improved Pyramid Scene Parsing Network. *Int. J. Agric. Biol. Eng.* **2021**, *14*, 185–194. [CrossRef]
14. Wu, J.; Wen, C.; Chen, H.; Ma, Z.; Zhang, T.; Su, H.; Yang, C. DS-DETR: A Model for Tomato Leaf Disease Segmentation and Damage Evaluation. *Agronomy* **2022**, *12*, 2023. [CrossRef]
15. Zhang, J.; Guo, H.; Guo, J.; Zhang, J. An Information Entropy Masked Vision Transformer (IEM-ViT) Model for Recognition of Tea Diseases. *Agronomy* **2023**, *13*, 1156. [CrossRef]
16. Chen, Z.; Su, R.; Wang, Y.; Chen, G.; Wang, Z.; Yin, P.; Wang, J. Automatic Estimation of Apple Orchard Blooming Levels Using the Improved YOLOv5. *Agronomy* **2022**, *12*, 2483. [CrossRef]
17. Feng, J.; Yu, C.; Shi, X.; Zheng, Z.; Yang, L.; Hu, Y. Research on Winter Jujube Object Detection Based on Optimized Yolov5s. *Agronomy* **2023**, *13*, 810. [CrossRef]
18. Zhang, S.; Ban, X.; Xiao, T.; Huang, L.; Zhao, J.; Huang, W.; Liang, D. Identification of Soybean Planting Areas Combining Fused Gaofen-1 Image Data and U-Net Model. *Agronomy* **2023**, *13*, 863. [CrossRef]
19. Li, Y.; Rao, Y.; Jin, X.; Jiang, Z.; Wang, Y.; Wang, T.; Wang, F.; Luo, Q.; Liu, L. YOLOv5s-FP: A Novel Method for In-Field Pear Detection Using a Transformer Encoder and Multi-Scale Collaboration Perception. *Sensors* **2022**, *23*, 30. [CrossRef]
20. Liu, L.; Du, Y.; Li, X.; Liu, L.; Mao, E.; Guo, D.; Zhang, Y. Trailer Hopper Automatic Detection Method for Silage Harvesting Based Improved U-Net. *Comput. Electron. Agric.* **2022**, *198*, 107046. [CrossRef]
21. Zhang, S.; Zhang, C. Modified U-Net for Plant Diseased Leaf Image Segmentation. *Comput. Electron. Agric.* **2023**, *204*, 107511. [CrossRef]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
23. Yu, X.; Yin, D.; Nie, C.; Ming, B.; Xu, H.; Liu, Y.; Bai, Y.; Shao, M.; Cheng, M.; Liu, Y.; et al. Maize Tassel Area Dynamic Monitoring Based on Near-Ground and UAV RGB Images by U-Net Model. *Comput. Electron. Agric.* **2022**, *203*, 107477. [CrossRef]
24. Zheng, Z.; Hu, Y.; Yang, H.; Qiao, Y.; He, Y.; Zhang, Y.; Huang, Y. AFFU-Net: Attention Feature Fusion U-Net with Hybrid Loss for Winter Jujube Crack Detection. *Comput. Electron. Agric.* **2022**, *198*, 107049. [CrossRef]
25. Zou, K.; Chen, X.; Wang, Y.; Zhang, C.; Zhang, F. A Modified U-Net with a Specific Data Argumentation Method for Semantic Segmentation of Weed Images in the Field. *Comput. Electron. Agric.* **2021**, *187*, 106242. [CrossRef]
26. Tu, K.; Wen, S.; Cheng, Y.; Zhang, T.; Pan, T.; Wang, J.; Wang, J.; Sun, Q. A Non-Destructive and Highly Efficient Model for Detecting the Genuineness of Maize Variety 'JINGKE 968' Using Machine Vision Combined with Deep Learning. *Comput. Electron. Agric.* **2021**, *182*, 106002. [CrossRef]
27. Ayhan, B.; Kwan, C. Tree, Shrub, and Grass Classification Using Only RGB Images. *Remote Sens.* **2020**, *12*, 1333. [CrossRef]
28. Raei, E.; Akbari Asanjan, A.; Nikoo, M.R.; Sadegh, M.; Pourshahabi, S.; Adamowski, J.F. A Deep Learning Image Segmentation Model for Agricultural Irrigation System Classification. *Comput. Electron. Agric.* **2022**, *198*, 106977. [CrossRef]
29. Roy, A.M.; Bhaduri, J. Real-Time Growth Stage Detection Model for High Degree of Occultation Using DenseNet-Fused YOLOv4. *Comput. Electron. Agric.* **2022**, *193*, 106694. [CrossRef]
30. Chen, J.; Wang, Z.; Wu, J.; Hu, Q.; Zhao, C.; Tan, C.; Teng, L.; Luo, T. An Improved Yolov3 Based on Dual Path Network for Cherry Tomatoes Detection. *J. Food Process Eng.* **2021**, *44*, e13803. [CrossRef]
31. Gong, L.; Du, X.; Zhu, K.; Lin, C.; Lin, K.; Wang, T.; Lou, Q.; Yuan, Z.; Huang, G.; Liu, C. Pixel Level Segmentation of Early-Stage in-Bag Rice Root for Its Architecture Analysis. *Comput. Electron. Agric.* **2021**, *186*, 106197. [CrossRef]
32. Kang, J.; Liu, L.; Zhang, F.; Shen, C.; Wang, N.; Shao, L. Semantic Segmentation Model of Cotton Roots In-Situ Image Based on Attention Mechanism. *Comput. Electron. Agric.* **2021**, *189*, 106370. [CrossRef]
33. Wang, F.; Rao, Y.; Luo, Q.; Jin, X.; Jiang, Z.; Zhang, W.; Li, S. Practical Cucumber Leaf Disease Recognition Using Improved Swin Transformer and Small Sample Size. *Comput. Electron. Agric.* **2022**, *199*, 107163. [CrossRef]
34. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.

35. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
36. Ma, J.; Li, Y.; Liu, H.; Du, K.; Zheng, F.; Wu, Y.; Zhang, L. Improving Segmentation Accuracy for Ears of Winter Wheat at Flowering Stage by Semantic Segmentation. *Comput. Electron. Agric.* **2020**, *176*, 105662. [CrossRef]
37. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Proceedings of the Advances in Neural Information Processing Systems*; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 12077–12090. [CrossRef]
38. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified Perceptual Parsing for Scene Understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
39. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
40. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
41. Martin, K.L.; Girma, K.; Freeman, K.W.; Teal, R.K.; Tubańa, B.; Arnall, D.B.; Chung, B.; Walsh, O.; Solie, J.B.; Stone, M.L.; et al. Expression of Variability in Corn as Influenced by Growth Stage Using Optical Sensor Measurements. *Agron. J.* **2007**, *99*, 384–389. [CrossRef]
42. Shaver, T.M.; Khosla, R.; Westfall, D.G. Evaluation of Two Ground-Based Active Crop Canopy Sensors in Maize: Growth Stage, Row Spacing, and Sensor Movement Speed. *Soil Sci. Soc. Am. J.* **2010**, *74*, 2101–2108. [CrossRef]
43. Quan, L.; Feng, H.; Lv, Y.; Wang, Q.; Zhang, C.; Liu, J.; Yuan, Z. Maize Seedling Detection under Different Growth Stages and Complex Field Environments Based on an Improved Faster R–CNN. *Biosyst. Eng.* **2019**, *184*, 1–23. [CrossRef]
44. Qiu, R.; Zhang, M.; He, Y. Field Estimation of Maize Plant Height at Jointing Stage Using an RGB-D Camera. *Crop J.* **2022**, *10*, 1274–1283. [CrossRef]
45. Zhou, Y.; Li, Y.; Liu, X.; Wang, K.; Muhammad, T. Synergistic Improvement in Spring Maize Yield and Quality with Micro/Nanobubbles Water Oxygation. *Sci. Rep.* **2019**, *9*, 5226. [CrossRef]
46. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
47. Kusrini, K.; Suputa, S.; Setyanto, A.; Agastya, I.M.A.; Priantoro, H.; Chandramouli, K.; Izquierdo, E. Data Augmentation for Automated Pest Classification in Mango Farms. *Comput. Electron. Agric.* **2020**, *179*, 105842. [CrossRef]
48. Vayssade, J.-A.; Jones, G.; Gée, C.; Paoli, J.-N. Pixelwise Instance Segmentation of Leaves in Dense Foliage. *Comput. Electron. Agric.* **2022**, *195*, 106797. [CrossRef]
49. Astani, M.; Hasheminejad, M.; Vaghefi, M. A Diverse Ensemble Classifier for Tomato Disease Recognition. *Comput. Electron. Agric.* **2022**, *198*, 107054. [CrossRef]
50. Picon, A.; San-Emeterio, M.G.; Bereciartua-Perez, A.; Klukas, C.; Eggers, T.; Navarra-Mestre, R. Deep Learning-Based Segmentation of Multiple Species of Weeds and Corn Crop Using Synthetic and Real Image Datasets. *Comput. Electron. Agric.* **2022**, *194*, 106719. [CrossRef]
51. Tassis, L.M.; Tozzi de Souza, J.E.; Krohling, R.A. A Deep Learning Approach Combining Instance and Semantic Segmentation to Identify Diseases and Pests of Coffee Leaves from In-Field Images. *Comput. Electron. Agric.* **2021**, *186*, 106191. [CrossRef]
52. De Melo, M.J.; Gonçalves, D.N.; Gomes, M. de N.B.; Faria, G.; Silva, J. de A.; Ramos, A.P.M.; Osco, L.P.; Furuya, M.T.G.; Marcato Junior, J.; Gonçalves, W.N. Automatic Segmentation of Cattle Rib-Eye Area in Ultrasound Images Using the UNet++ Deep Neural Network. *Comput. Electron. Agric.* **2022**, *195*, 106818. [CrossRef]
53. Zhou, J.; Lu, Y.; Tao, S.; Cheng, X.; Huang, C. E-Res U-Net: An Improved U-Net Model for Segmentation of Muscle Images. *Expert Syst. Appl.* **2021**, *185*, 115625. [CrossRef]
54. Su, Z.; Li, W.; Ma, Z.; Gao, R. An Improved U-Net Method for the Semantic Segmentation of Remote Sensing Images. *Appl. Intell.* **2022**, *52*, 3276–3288. [CrossRef]
55. Zhou, Y.; Huang, W.; Dong, P.; Xia, Y.; Wang, S. D-UNet: A Dimension-Fusion U Shape Network for Chronic Stroke Lesion Segmentation. *IEEE ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 940–950. [CrossRef]
56. Dong, X.; Yan, S.; Duan, C. A Lightweight Vehicles Detection Network Model Based on YOLOv5. *Eng. Appl. Artif. Intell.* **2022**, *113*, 104914. [CrossRef]
57. Zeng, W.; Li, H.; Hu, G.; Liang, D. Lightweight Dense-Scale Network (LDSNet) for Corn Leaf Disease Identification. *Comput. Electron. Agric.* **2022**, *197*, 106943. [CrossRef]
58. Zhou, K.; Zhang, M.; Wang, H.; Tan, J. Ship Detection in SAR Images Based on Multi-Scale Feature Extraction and Adaptive Feature Fusion. *Remote Sens.* **2022**, *14*, 755. [CrossRef]
59. Ming, Q.; Miao, L.; Zhou, Z.; Song, J.; Yang, X. Sparse Label Assignment for Oriented Object Detection in Aerial Images. *Remote Sens.* **2021**, *13*, 2664. [CrossRef]
60. Abdalla, A.; Cen, H.; Wan, L.; Rashid, R.; Weng, H.; Zhou, W.; He, Y. Fine-Tuning Convolutional Neural Network with Transfer Learning for Semantic Segmentation of Ground-Level Oilseed Rape Images in a Field with High Weed Pressure. *Comput. Electron. Agric.* **2019**, *167*, 105091. [CrossRef]
61. Sun, J.; Zhou, J.; He, Y.; Jia, H.; Liang, Z. RL-DeepLabv3+: A Lightweight Rice Lodging Semantic Segmentation Model for Unmanned Rice Harvester. *Comput. Electron. Agric.* **2023**, *209*, 107823. [CrossRef]

62. Peng, H.; Zhong, J.; Liu, H.; Li, J.; Yao, M.; Zhang, X. ResDense-Focal-DeepLabV3+ Enabled Litchi Branch Semantic Segmentation for Robotic Harvesting. *Comput. Electron. Agric.* **2023**, *206*, 107691. [CrossRef]
63. Wspanialy, P.; Moussa, M. A Detection and Severity Estimation System for Generic Diseases of Tomato Greenhouse Plants. *Comput. Electron. Agric.* **2020**, *178*, 105701. [CrossRef]
64. Jiang, Z.; Dong, Z.; Jiang, W.; Yang, Y. Recognition of Rice Leaf Diseases and Wheat Leaf Diseases Based on Multi-Task Deep Transfer Learning. *Comput. Electron. Agric.* **2021**, *186*, 106184. [CrossRef]
65. Zhang, J.; Karkee, M.; Zhang, Q.; Zhang, X.; Yaqoob, M.; Fu, L.; Wang, S. Multi-Class Object Detection Using Faster R-CNN and Estimation of Shaking Locations for Automated Shake-and-Catch Apple Harvesting. *Comput. Electron. Agric.* **2020**, *173*, 105384. [CrossRef]
66. Altuntaş, Y.; Cömert, Z.; Kocamaz, A.F. Identification of Haploid and Diploid Maize Seeds Using Convolutional Neural Networks and a Transfer Learning Approach. *Comput. Electron. Agric.* **2019**, *163*, 104874. [CrossRef]
67. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. [CrossRef]
68. Yang, L.; Zhang, F.; Wang, P.S.-P.; Li, X.; Meng, Z. Multi-Scale Spatial-Spectral Fusion Based on Multi-Input Fusion Calculation and Coordinate Attention for Hyperspectral Image Classification. *Pattern Recognit.* **2022**, *122*, 108348. [CrossRef]
69. Zha, M.; Qian, W.; Yi, W.; Hua, J. A Lightweight YOLOv4-Based Forestry Pest Detection Method Using Coordinate Attention and Feature Fusion. *Entropy* **2021**, *23*, 1587. [CrossRef]
70. Wu, H.; Zhang, J.; Huang, K.; Liang, K.; Yu, Y. FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation. *arXiv* **2019**, arXiv:1903.11816.
71. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds. Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29. [CrossRef]
72. Pang, Y.; Li, Y.; Shen, J.; Shao, L. Towards Bridging Semantic Gap to Improve Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4229–4238. [CrossRef]
73. Liu, J.; Wang, X.; Wang, T. Classification of Tree Species and Stock Volume Estimation in Ground Forest Images Using Deep Learning. *Comput. Electron. Agric.* **2019**, *166*, 105012. [CrossRef]
74. Zaji, A.; Liu, Z.; Xiao, G.; Bhowmik, P.; Sangha, J.S.; Ruan, Y. Wheat Spike Localization and Counting via Hybrid UNet Architectures. *Comput. Electron. Agric.* **2022**, *203*, 107439. [CrossRef]
75. Zhang, Y.; Duan, H.; Liu, Y.; Li, Y.; Lin, J. Converge of Coordinate Attention Boosted YOLOv5 Model and Quantum Dot Labeled Fluorescent Biosensing for Rapid Detection of the Poultry Disease. *Comput. Electron. Agric.* **2023**, *206*, 107702. [CrossRef]
76. Wan, H.; Zeng, X.; Fan, Z.; Zhang, S.; Kang, M. U2ESPNet—A Lightweight and High-Accuracy Convolutional Neural Network for Real-Time Semantic Segmentation of Visible Branches. *Comput. Electron. Agric.* **2023**, *204*, 107542. [CrossRef]