agronomy

MDPI

Article

Weed Identification in Maize Fields Based on Improved Swin-Unet

Jiaheng Zhang [1,*], Jinliang Gong [1,*], Yanfei Zhang [2], Kazi Mostafa [3] and Guangyao Yuan [1]

[1] School of Mechanical Engineering, Shandong University of Technology, Zibo 255049, China; loki971125@163.com
[2] School of Agricultural Engineering and Food Science, Shandong University of Technology, Zibo 255049, China; zyfwing@foxmail.com
[3] School of Intelligent Manufacturing Ecosystem, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; kazi.mostafa@xjtlu.edu.cn
* Correspondence: wan5480556@163.com (J.Z.); gjlwing@sdut.edu.cn (J.G.)

Abstract: The maize field environment is complex. Weeds and maize have similar colors and may overlap, and lighting and weather conditions vary. Thus, many methods for the automated differentiation of maize and weeds achieve poor segmentation or cannot be used in real time. In this paper, a weed recognition model based on improved Swin-Unet is proposed. The model first performs semantic segmentation of maize seedlings and uses the resulting mask to identify weeds. U-Net acts as the semantic segmentation framework, and a Swin transformer module is introduced to improve performance. DropBlock regularization, which randomly hides some blocks in crop feature maps, is applied to enhance the generalization ability of the model. Finally, weed areas are identified and segmented with the aid of an improved morphological processing algorithm. The DeepLabv3+, PSANet, Mask R-CNN, original Swin-Unet, and proposed models are trained on a dataset of maize seedling images. The proposed Swin-Unet model outperforms the others, achieving a mean intersection over union of 92.75%, mean pixel accuracy of 95.57%, and inference speed of 15.1 FPS. Our model could be used for accurate, real-time segmentation of crops and weeds and as a reference for the development of intelligent agricultural equipment.

Keywords: crop; target recognition; target segmentation; semantic segmentation; weed recognition

Citation: Zhang, J.; Gong, J.; Zhang, Y.; Mostafa, K.; Yuan, G. Weed Identification in Maize Fields Based on Improved Swin-Unet. *Agronomy* **2023**, *13*, 1846. https://doi.org/10.3390/agronomy13071846

Academic Editor: Sara Di Lonardo

Received: 29 May 2023
Revised: 5 July 2023
Accepted: 11 July 2023
Published: 13 July 2023

1. Introduction

Weeds can greatly affect the yield and quality of crops [1]. Weeds not only compete with crops for nutrients and delay seedling development, but they also attract pests and diseases; hence, weeding is a necessary task in the field management of seedling crops [2,3]. At present, chemical weed control, which is easy and efficient, is typically used in maize fields [4], but the irregular use of herbicides damages soil, crops, and human health [5–7]. In recent years, precision agriculture methods have been proposed, leading numerous scholars to investigate automatic and accurate weed identification systems, such as those based on machine vision and images [8,9].

Conventional machine vision methods detect features such as the color, shape, texture, and spatial geometry of a target; the recognition and classification of the target are performed with methods including feature fusion, wavelet transforms, and support vector machines (SVMs) [10–12]. Although these recognition methods are simple to implement, they may be ineffective for some crops with different sizes and growth conditions, and they cannot analyze nonstructural environmental factors, such as dead grass and rocks. Hence, recognition models that rely on target-specific features have poor robustness and low accuracy.

With the recent development of deep learning methods, convolutional neural networks (CNNs) have been widely and successfully applied for machine vision tasks [13].

Jiang et al. [14] established graph CNNs for the classification and identification of multiple crops and their associated weeds. Zhou et al. [15] used Moderate Resolution Imaging Spectroradiometer (commonly abbreviated as MODIS) satellite data to construct a CNN model for the complex task of predicting winter wheat yields. Peng et al. [16] trained a deep CNN on a weed dataset and used stochastic gradient descent (SGD) to optimize the model; the resulting VGG16-SGD network had the highest recognition accuracy of the tested methods, with an average F1 score of 0.977. Meng et al. [17] improved a single-shot multibox detector (SSD) by constructing a lightweight antecedent base network, and also fused the information from different feature layers to improve the model's recognition accuracy while reducing its number of parameters. Wang et al. [18] proposed a maize field weed identification method based on Swin transformer. They improved the backbone of the model and adjusted network parameters to generate four model variants. By combining a morphological processing algorithm, they achieved accurate identification and segmentation of crops and weeds.

These studies have revealed that, in contrast to conventional image processing methods, deep learning models need not rely on specific features for weed recognition; moreover, they have higher accuracy. However, deep learning methods have various problems. For example, CNNs often extract irrelevant features from pixels depicting dead grass, rocks, and other background objects. Moreover, feature layers may have no interaction mechanism, rendering the extraction of board contextual information challenging for a model. These factors limit the recognition accuracy and inference speed of CNN models.

This paper presents a recognition method based on an enhanced Swin-Unet network. During object recognition, this model extracts more targeted information, reducing the interference of redundant information on inference speed. The proposed model in this paper captures more global visual information through its self-attention mechanism during feature extraction. It also enables interaction between the extracted features. In addition, the method used in this paper only requires annotations for crop targets, greatly reducing the difficulty of obtaining samples.

2. Data Processing

2.1. Data Acquisition

In this study, images of maize seedlings at the three-to-five-leaf stage were collected from actual plots that had not been weeded. The images were captured in a maize test field in Zibo, Shandong Province, China, using an iPhone XS with a maximum resolution of 1920 × 1080 pixels. The images were acquired parallel to the ground at a height of 60 cm. To ensure that the dataset was generalizable, images were captured at three times of day—morning (07:00–09:00), late morning (10:00–12:00), and evening (16:00–18:00)—to reflect the varied lighting of an actual crop field. To minimize image redundancy and maximize model stability, the images were screened; 1000 images depicting complex situations, such as target overlap and various lighting and growth environments, were selected.

The resolution of the collected images was first adjusted to 512 × 512 pixels. The labelme (v5.2.1) software was then used to manually label the maize seedlings in the images with the polygon labeling method by selecting and connecting a dense group of points on a target outline in an image to form a closed polygon bounding the seedlings (Figure 1b). The pixels within the closed polygon were labeled as maize, and the remaining pixels were labeled as background, as shown in Figure 1c. Only the maize seedlings needed to be manually labeled before image processing, limiting the labor required for sample acquisition and labeling.

The labeled images were used to produce a dataset in the PASCAL Visual Object Classes 2007 format and divided into training, test, and validation sets at a 7:2:1 ratio. The model was trained on the training set, while the test set was used to evaluate the generalization ability of the model after training, and the validation was used for hyperparameter tuning during training.

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

Figure 1. Image labeling process. (**a**) Original image, (**b**) annotated image, (**c**) generated label.

2.2. Data Enhancement

The collected dataset had few images and therefore was likely to produce an overfitted model. Hence, data augmentation was used to expand the training set by a factor of 5 through random adjustment of the saturation [19], brightness, contrast, orientation, and scale (scaling factor of 1–1.5) of the images. Such a process can improve model training, resulting in superior semantic segmentation accuracy.

3. Model Construction

3.1. Swin-Unet Semantic Segmentation Model

To accurately identify the crop seedlings and their morphology in a maize field, the Swin-Unet maize seedling semantic segmentation model was developed. The network backbone is that of U-Net, and Swin transformer blocks (Figure 2) are used for feature extraction and target segmentation [20,21]. The Swin-Unet model comprises four main components: an encoder, a decoder, a bottleneck, and skip connections. The key unit of the Swin-Unet network is the Swin transformer module. The encoder mainly performs patch partition, linear embedding, and patch merging functions with Swin transformer blocks. The encoder performs layer-by-layer downsampling on an input image to obtain feature information at different scales, and then gradually compresses and fuses the extracted feature information to obtain a high-level semantic feature representation. The decoder mainly performs patch expanding and linear projection functions with a Swin transformer block. The decoder's structure mirrors that of the encoder; using deconvolution, it gradually reduces the feature information extracted by the encoder and fuses shallow and deep feature information to increase the accuracy of the segmentation results. The bottleneck comprises two Swin transformer blocks, which reduce the dimensionality of the feature maps extracted by the encoder, reducing the required computation. The skip connection layer fuses the high-level feature maps from the encoder with the corresponding low-level feature maps from the decoder by using convolutional layers. This process improves the information flow and enables the use of feature information at various levels, effectively preventing gradient vanishing while improving semantic segmentation.

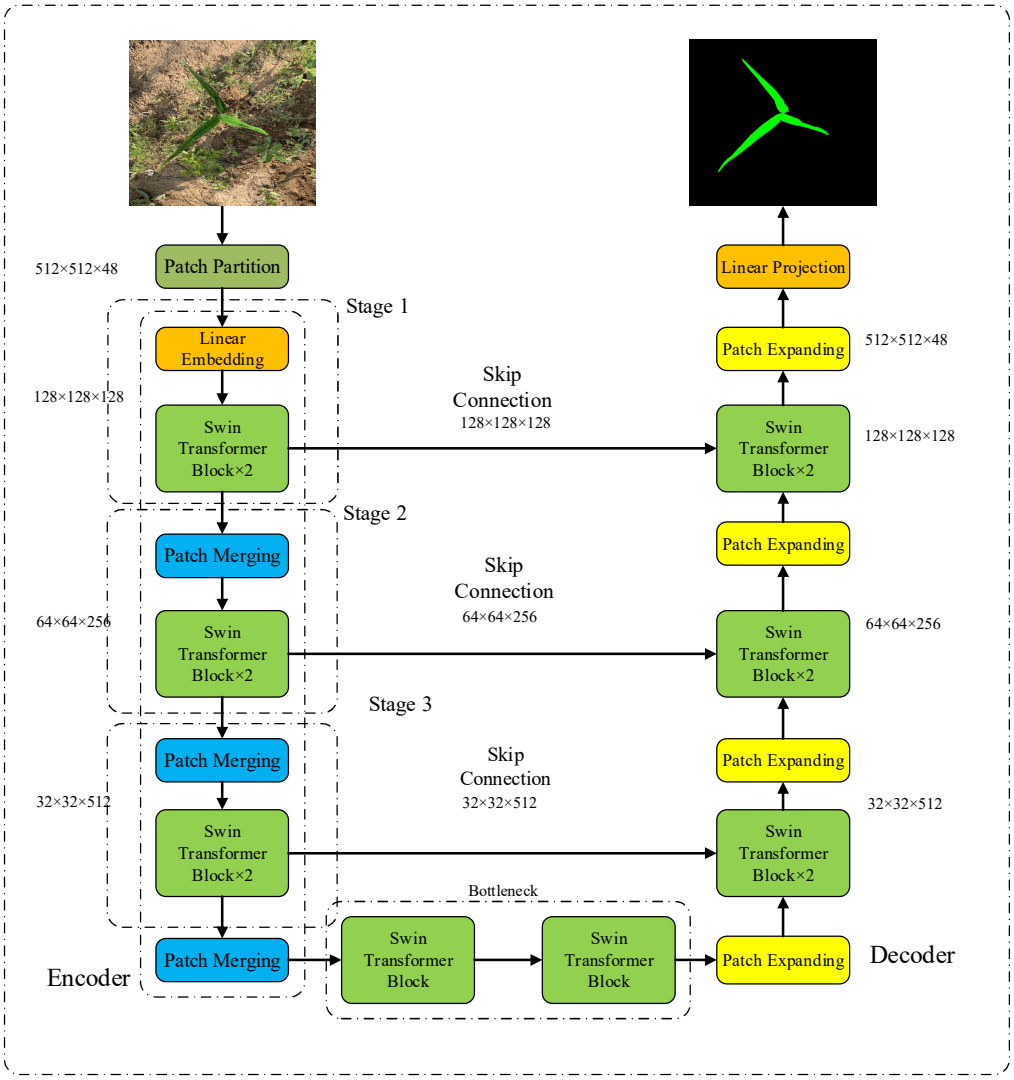


Figure 2. Swin-Unet network structure.

3.2. Swin Transformer Block

The Swin transformer block is the basic unit of the Swin-Unet network and executes window multi-head self-attention (W-MSA), multi-head self-attention based on a shift window (SW-MSA), multi-layer perceptron with a GELU nonlinear activation function (MLP), residual connectivity, and layer normalization functions [22]. The structure of the Swin transformer block is shown in Figure 3. The formula of the block can be expressed as follows [23]:

$$\hat{Z}^l = W - MSA(LN(Z^{l-1})) + Z^{l-1} \quad (1)$$

$$Z^l = MLP(LN(\hat{Z}^l)) + \hat{Z}^l \quad (2)$$

$$\hat{Z}^{l+1} = SW - MAS(LN(Z^l)) + Z^l \quad (3)$$

$$Z^{l+1} = MLP(LN(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \quad (4)$$

where $\hat{Z}^l$ represents the features output by the (S) W-MSA module, and Z^l represents the features output by the MLP module; l represents the number of blocks.

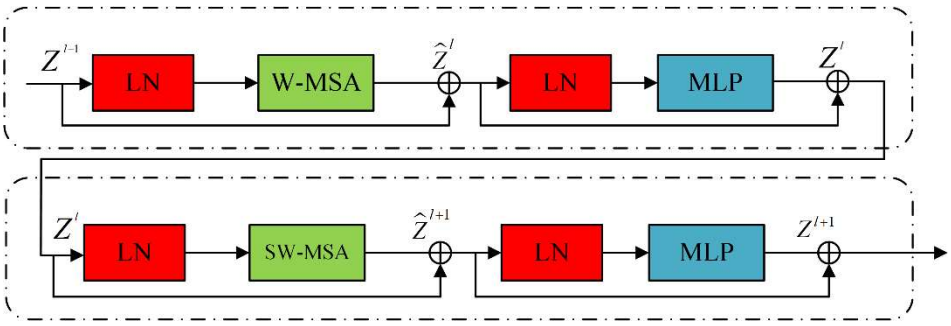


Figure 3. Structure of the Swin transformer block.

The Swin transformer block extracts high-dimensional features from the input image and generates feature maps of various scales. In the encoder of the proposed model, multiple Swin transformer blocks are stacked to capture more comprehensive and global semantic visual information. The block uses shift window division and a unidirectional cyclic mechanism to divide the acquired feature maps of various scales into disjoint windows, each with similar semantic features. The features within each window are weighted by W-MSA, and the attention weights are adaptively adjusted in accordance with bootstrap features selection. The features within each window are then weighted and scaled in accordance with the adjusted weights. SW-MSA performs information transfer among the sub-windows in combination with a patch-merging operation to achieve global attention. This process obtains a more comprehensive feature expression as follows:

$$Attention(Q, K, V) = \text{SoftMax}(\frac{QK^T}{\sqrt{d}} + B)V \quad (5)$$

where Q denotes the query vector, K denotes the key vector, V denotes the value vector, d denotes the dimensionality of the key–value vector (i.e., the hyperparameter), and B denotes the relative position bias. Attention is calculated independently for each sub-window.

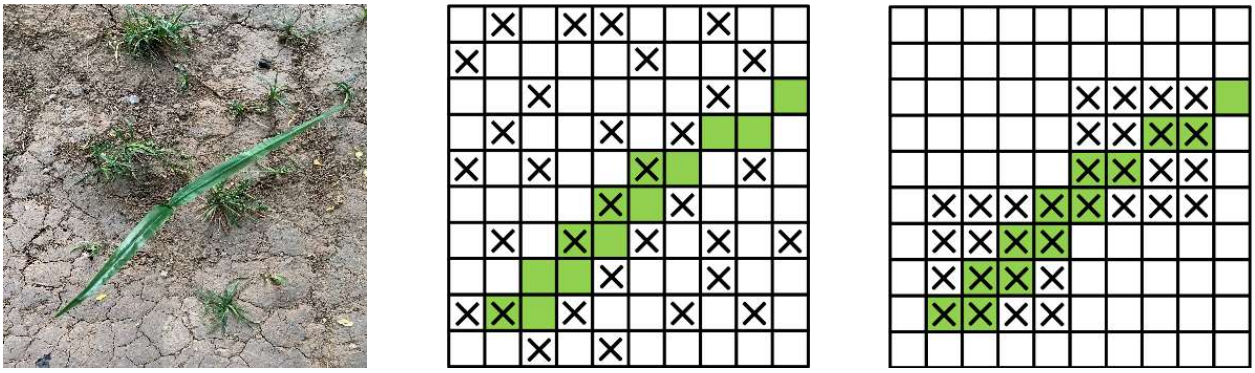
3.3. DropBlock Regularization

Dropout regularization is used in the original Swin-Unet network. This regularization approach reduces the risk of model overfitting by randomly hiding neurons in the fully connected layer [24]. However, dropout is ineffective in models with convolutional layers, because each feature element in the feature map has a corresponding perceptual field. The size of the extracted feature maps decreases as the size of perceptual fields and number of network layers increase, eventually resulting in the model being unable to learn the corresponding semantic information from the adjacent elements. This limits generalization ability and causes data explosion.

To avoid these problems, the DropBlock regularization method is used in our proposed model. DropBlock interferes with the learning of semantic information between neighboring blocks by randomly hiding the feature blocks, forcing the network to learn to use information from other feature blocks. Moreover, the remaining feature blocks are normalized to achieve regularization, reducing overfitting while improving model performance [25]. Figure 4 presents an example of this regularization process. Figure 4a shows the input image, while Figure 4b demonstrates the result of Dropout regularization. The green area represents the activated units of semantic information in the input image, and Dropout regularization randomly masks this semantic information. However, the convolutional layers exhibit spatial correlation among features, and the random Dropout mechanism cannot effectively mask the semantic information between neighboring blocks, thereby posing a risk of overfitting in the network. Figure 4c illustrates the result of DropBlock regularization, which randomly masks contiguous blocks in the feature map. This approach effectively prevents the flow of semantic information between adjacent blocks in the network, further reducing the risk of overfitting.

3.4. Weed Identification Model Based on Improved Swin-Unet

The collected images were 512 × 512 pixel, three-channel red, green, blue (RGB) images. We applied our network to these images using the process described as follows: first, the patch partition function segments an input image into nonoverlapping 4 × 4 image blocks; including the three channels, each image block contains 4 × 4 × 3 = 48 data points. The size of the segmentation blocks is selected on the basis of the desired size of the feature map to be output by each stage of the network. Small segmentation blocks result in segmented feature maps with low dimensionality and carrying capacity for local information. Moreover, more segmentation blocks increase the computational complexity of the model. By contrast, large segmentation blocks result in a low-resolution output feature map, which typically results in low recognition accuracy.



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

Figure 4. (**a**) Original image; (**b**) Dropout regularization results; (**c**) DropBlock regularization results.

Stage 1 of the model comprises the linear embedding function and a Swin transformer block. The linear embedding function maps the original feature dimension of the segmented image block to 128, and the Swin transformer block then performs feature extraction on the image. This process does not change the resolution or dimensions of the output feature map. Both stages 2 and 3 of the model perform the patch-merging function and contain a Swin transformer block. Patch merging halves the resolution of the feature map and increases the feature dimensionality, forming a hierarchical feature representation through a downsampling while reducing the required computations during learning. The process is illustrated in Figure 5 for an example of a single-channel image.

The image is then processed through the bottleneck to a size of 32 × 32 pixels, with a feature dimension of 512. The image is then inputted into the decoder, which has a layout opposite that of the encoder. Patch expansion is applied in which the original feature map is reconstructed from the inputted low-resolution feature map by an upsampling operation, which increases the size of the feature map and reduces the feature dimension. The skip connection mechanism in the U-Net network then concatenates the image features extracted by the encoder layer corresponding to the current decoder layer to mitigate the loss of spatial information caused by the downsampling operation. After patch expanding, the Swin transformer block then enhances the feature map representation by applying its unique self-attention mechanism, enabling the network to effectively capture the semantic information and spatial structure of the image.

The image is then upsampled four times through patch expansion, ultimately yielding a 512 × 512 feature map with a feature dimension of 48. It is then processed by a linear projection function to complete pixel classification and output the category mask, which is used to produce the segmentation map. This process enables pixel-level inference on a target image, resulting in fine-grained segmentation and accurate target recognition. The related details of the image processing pipeline for this model are illustrated in Figure 2.

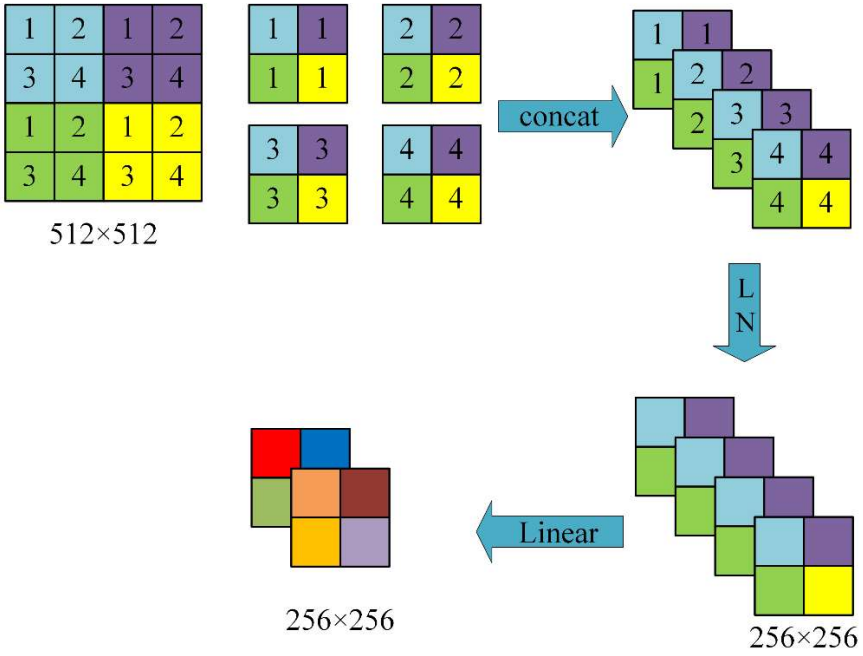


Figure 5. Single-channel patch-merging process.

Corn seedlings and weeds have similar colors in images; hence, distinguishing them is difficult. However, extracting all plant regions and stripping the background from the image is easier. This paper proposes a method in which a semantic segmentation model is combined with simple morphological processing to achieve rapid segmentation of foreground weeds. An overview of the method is visualized in Figure 6; the steps are as follows: (1) Original RGB images are segmented using the ExG super green feature algorithm (Excess Green). A binarized image is then obtained by using the super green feature component of the image as the gray value, and Otsu's algorithm (the maximum interclass variance algorithm) is applied to extract a mask of the plant region. (2) The crop area is deleted from the plant region mask. The crop mask is subjected to a mild dilation operation to clarify its boundaries, and the values of the pixels in the maize region of the segmented mask are set to 0 such that only the weed region is included in the mask. (3) Then, the weed mask is further optimized through a dilation operation to fill small holes, followed by subsequent etching to increase the adaptability and smoothness of the mask boundaries. Finally, the image is subjected to a closing operation to further eliminate fine voids and noise within the mask area and to optimize the morphology of the mask area. (4) Finally, a weed mask segmentation map is obtained.

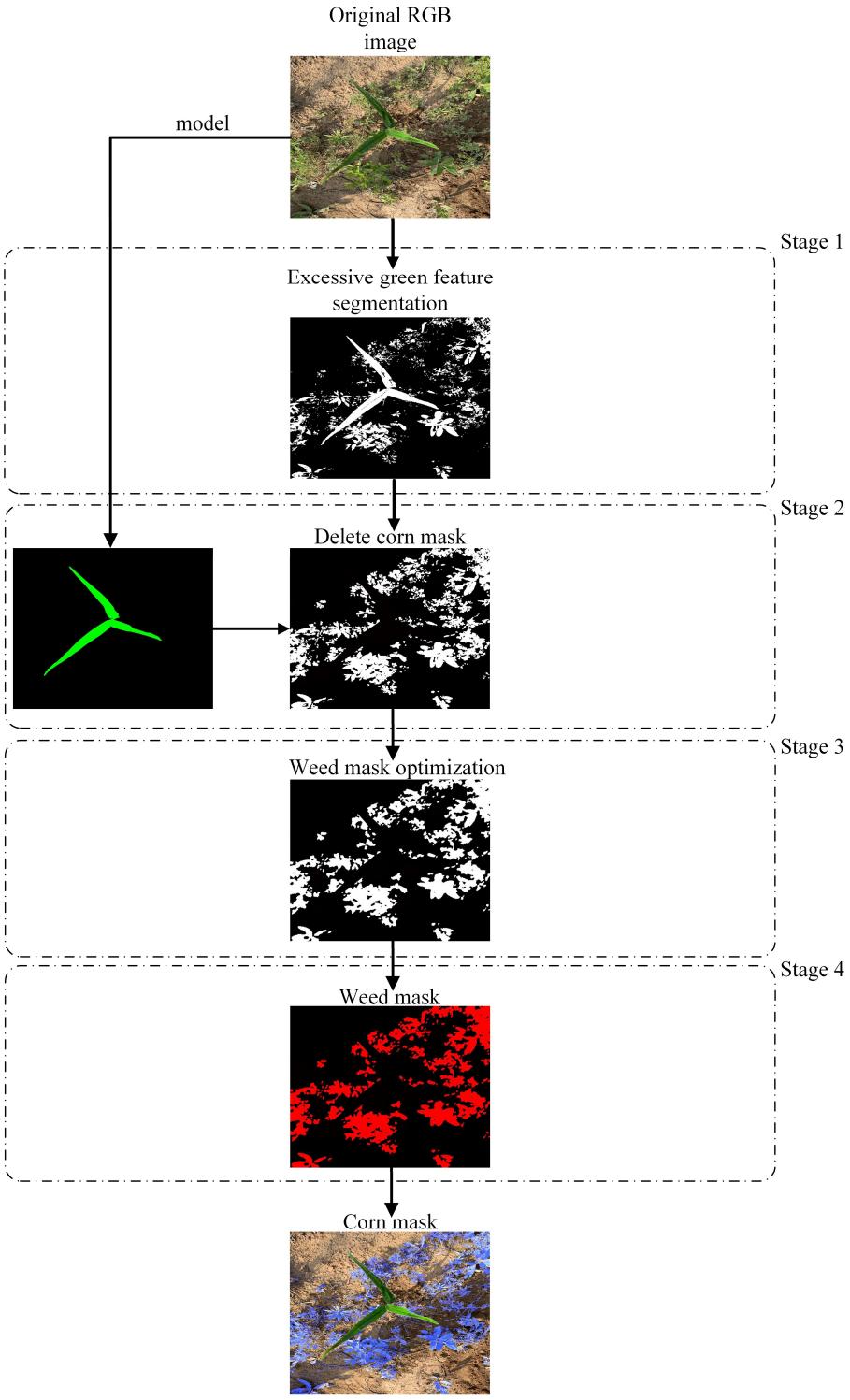


Figure 6. Flow of weed recognition algorithm.

4. Weed Recognition Test

4.1. Test Environment

Model testing and training were performed on an Ubuntu 20.04 system with the Pytorch1.6 deep learning framework, CUDA11.6 parallel computing architecture, cuDNN8.4.0 deep neural network GPU acceleration library, Python 3.9 programming language, and OpenCV4.5.1 vision library. The system had 64 GB of memory, an Intel Core i9-10900KF CPU running at 3.7 GHz with 20 threads, and an NVIDIA GeForce RTX3090 produced

using the 8-nm production process and with 24 GB of memory and a core frequency of 1695–1725 MHz.

4.2. Parameter Settings

The model was trained in an end-to-end manner; the input data were raw RGB images, and the outputs were the corresponding segmentation masks. The model was trained through transfer learning; that is, the network was first pretrained on the ImageNet large plant dataset to obtain the initial weights [26], and then tuned on the target dataset. ImageNet is a large dataset that is effective for transfer learning [27]. The pretrained model was then trained on the target dataset to fine-tune these parameters for the target recognition task. Such transfer learning was intended to improve the generalizability of the model.

To maximize the learning efficiency possible with the hardware capabilities, the following hyperparameters were selected: batch size of 32,110 epochs, and 11,000 total iterations. For the learning rate, a warm-up strategy was used; that is, the learning rate was initially low but increased with the number of iterations. Such an approach accelerates convergence during training. The AdamW [28] optimizer, linear learning rate decay, and linear warm-up for 880 iterations were used. The learning rate *lr* for each epoch can be calculated as follows:

$$lr = lr_0 \times \left[1 - \left(\frac{iter}{iter_{\max}}\right)^p\right] \tag{6}$$

where *lr* ($lr \geq 0$) is the learning rate for iteration number *iter*, and lr_0 is the initial learning rate. *iter* is the number of current iterations, $iter_{max}$ is the decay period, and *P* is the learning rate power. In this paper, $lr_0 = 6 \times 10^{-5}$, and $P = 1$. Moreover, weight decay was set to 0.01, and momentum was set to 0.9.

The loss considered was the cross-entropy loss, which is the distance between the predicted category probability distribution and the true label probability distribution and is calculated as follows:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} g(\alpha_i)\log(\beta_{ic}) \tag{7}$$

where *N* denotes the number of samples (i.e., pixels), *K* is the number of categories, *i* indicates the current sample, *c* indicates the current category, α_i is the true label category of sample *I*, *g* is the probability distribution function (equal to 1 if $\alpha_i = c$ and 0 otherwise), and β_{ic} is the probability that sample *i* belongs to category *c* as predicted by a sigmoid activation function. During training, the performance of the model is evaluated on the basis of the loss function's value, and the parameters are fine-tuned by back-propagation to reduce the distance between the predicted category and the true label to increase model accuracy.

4.3. Model Evaluation Metrics

In this study, mean intersection over union (mIoU), mean pixel accuracy (mPA), and inference speed were used as metrics for evaluating model performance.

5. Results and Analysis

5.1. Training Error

The Swin-Unet model was trained on the training dataset for 11,000 iterations; the loss function is presented in Figure 7. As the number of iterations increased, the loss gradually decreased, and the network converged with a final average loss value of approximately 1.2×10^{-2}, indicating that Swin-Unet had been trained successfully.

5.2. Model Comparison

Mask R-CNN, DeepLabv3+, PSANet, the original Swin-Unet, and the improved model presented in this paper were trained, validated, and tested on the collected dataset, and the mIoU, mPA, and inference speed of the models were compared. Each model was trained

three times with initial parameter weights generated from random seeds, and their results on the validation dataset were averaged. Moreover, the performance of each model during training was evaluated after every 1100 iterations. The variations in mIoU and mPA with the number of training iterations are presented in Figure 8.

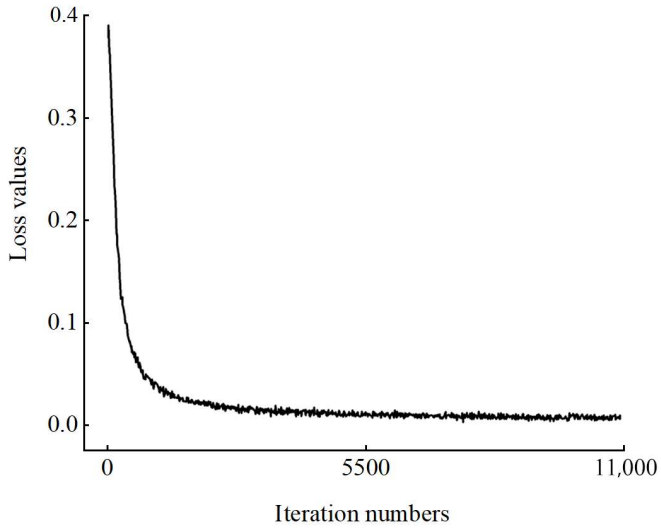
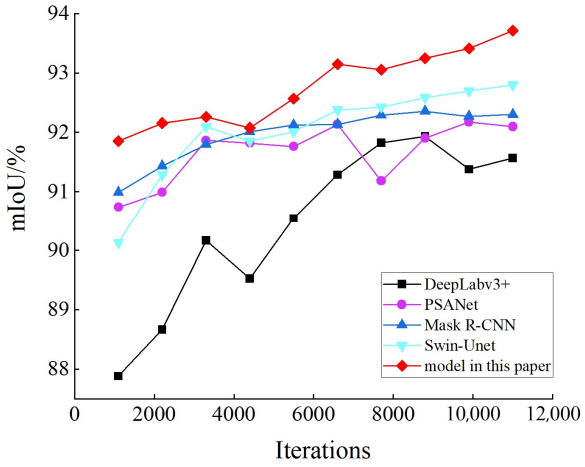
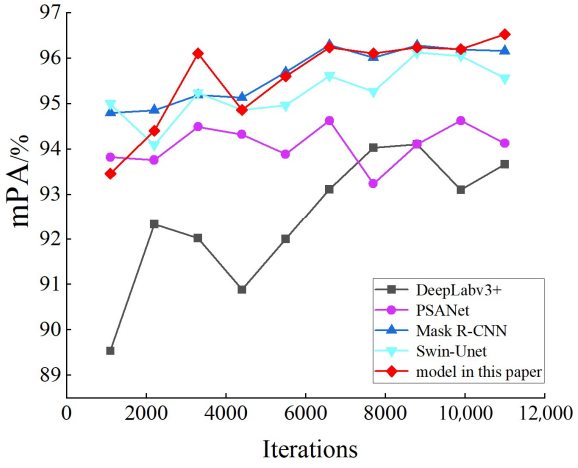


Figure 7. Model loss by number of iterations.



(**a**)



(**b**)

Figure 8. Training performance of each model. (**a**) mIoU of each model; (**b**) mPA of each model.

Figure 8 reveals that the mIoU of each network model tended to increase as training progressed; hence, the generalizability of these models is proportional to the number of iterations. The mIoU curves and mPA curves of Swin-Unet, DeepLabv3+, and PSANet have large fluctuations, indicating that these networks achieved poor segmentation and generalization ability after training. In the figure, the mIoU values of Mask R-CNN and the original Swin-Unet stabilize after 5000 iterations, whereas that of the network model proposed model increases linearly for approximately 7000 iterations before stabilizing at 93.71%, substantially higher than the mIoU of other models. After approximately 6000 iterations, the mPA of the proposed model stabilizes at 96.52% and is greater than the final value for Mask R-CNN of 96.15%, and compared to reference [29], there was a certain improvement in both the accuracy of identification and segmentation. Thus, the proposed model achieved the highest mIoU and mPA of all models tested and therefore the optimal performance. The trained models were tested on the test set; the results are listed in Table 1.

Table 1. Model performance on test set.

Networks	mIoU (%)	mPA (%)	Segmentation Rate (FPS)
DeepLabv3+	90.48	92.47	14.9
PSANet	91.67	94.09	14.3
Mask R-CNN	91.97	95.06	15.3
Swin-Unet	92.03	95.27	15.0
Model in this paper	92.75	95.57	15.1

Table 1 reveals that the mIoU of the proposed model (92.75%) was 2.27%, 1.08%, 0.78%, and 0.72% greater than those of the DeepLabv3+, PSANet, Mask R-CNN, and original Swin-Unet models, respectively (mean improvement of 1.21%). Similarly, the mPA of the proposed model (95.57%) was 3.10%, 1.48%, 0.51%, and 0.30% greater than those of the DeepLabv3+, PSANet, Mask R-CNN, and original Swin-Unet models, respectively (mean improvement of 1.35%). All models exhibited similar inference speed; the average inference speed was 14.92 FPS. The proposed model achieved both superior segmentation accuracy and inference speed to those achieved by the other models, and compared to reference [30], there was a certain improvement in the inference speed. Thus, the model can be used for detection of weeds and accurate weeding in the unstructured environments of maize fields.

5.3. Maize Identification and Segmentation

To test the recognition accuracy and segmentation of the proposed model, each model first performed inference on the test set, and the obtained maize masks were overlaid with the original RGB images for quantitative evaluation of recognition and segmentation ability. Some example images are displayed in Figure 9.

Figure 9a–c presents randomly selected original images and the corresponding segmentations by the proposed model DeepLabv3+. The images reveal that the proposed model accurately identified the maize seedling; the segmentation error was limited to a few pixels near the boundary. By contrast, DeepLabv3+ tended to mis-segment the ends of leaves and perform poorly in areas in which leaves and weeds overlapped. The results suggest that the model proposed can accurately identify and segment a target crop in a complex, unstructured environment more effectively than competing models.

A weed segmentation map based on the inference of proposed improved Swin-Unet network was generated. The results are presented in Figure 10.

The images in Figure 10 reveal that the proposed algorithm and model can recognize and segment areas containing weeds. In each of these randomly selected test-set images, the weed area is segmented completely with the target crop preserved. The maize and weed areas are independent, retain their respective morphological characteristics, and do not overlap. Hence, our model can be used to identify weeds in the unstructured environments of maize fields, despite crop overlap. The proposed model is simple and efficient and can be implemented in real time to provide reliable visual information for weeding robots.

Figure 9. Comparison of the segmentation results of the proposed models and DeepLabv3+.

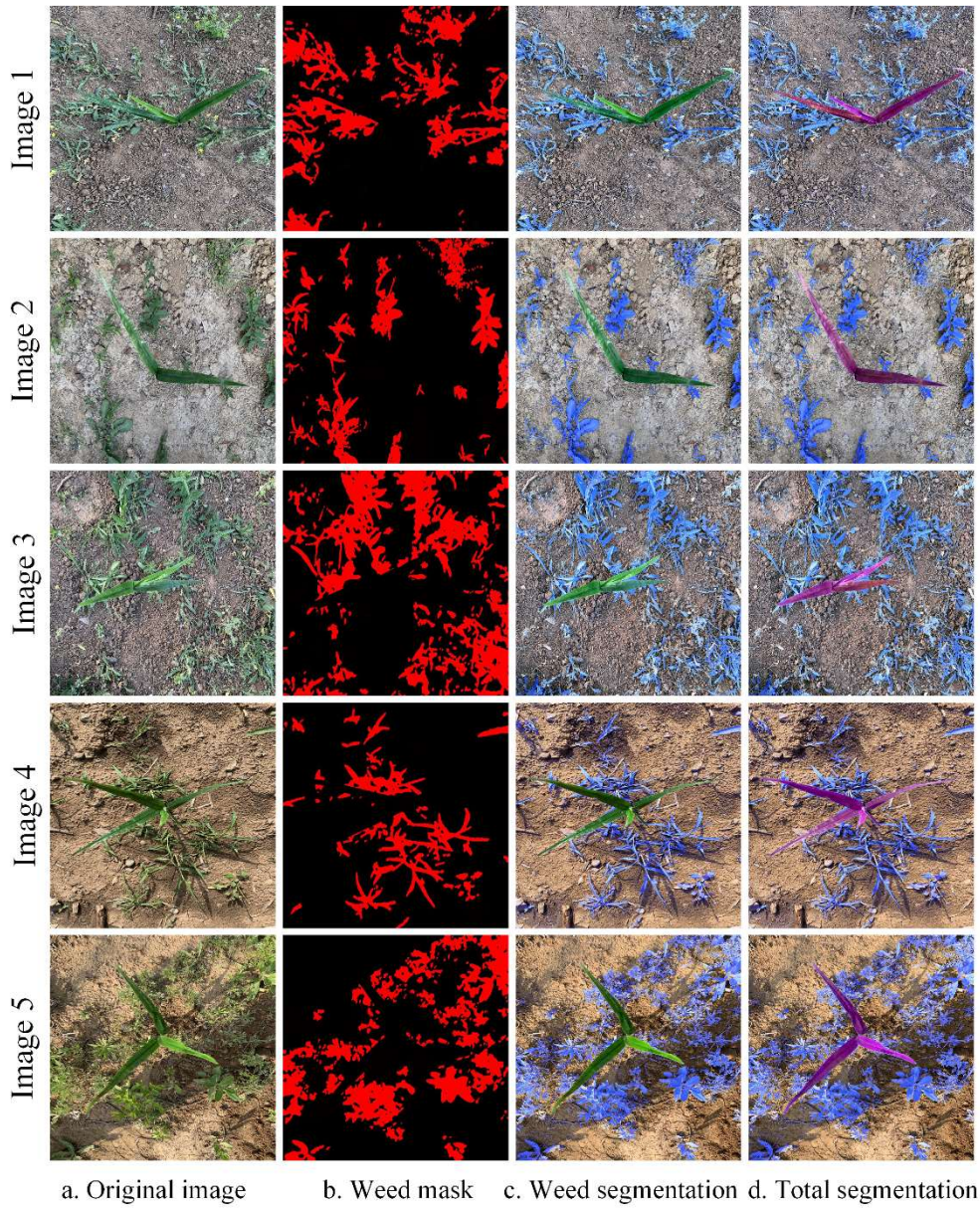


Figure 10. Weed recognition and segmentation effect.

6. Conclusions

To improve the accuracy of models for recognizing weeds in the complex environment of maize fields, this study developed an improved version of the Swin-Unet model. The model selectively extracts target features to reduce interference from redundant information on the inference speed. The proposed model in this paper can capture more global visual information and achieve an interaction between information through an SW-MSA mechanism. Additionally, this model focuses solely on the identification of crop targets without the requirement of identifying and classifying common companion weeds in the field. This greatly reduces the difficulty of sample collection compared to the literature [31,32]. The model was able to effectively recognize and segment target crops in a complex environment.

(1) The proposed model achieved up to 96.52% mPA and 93.71% mIoU, superior to those achieved by the DeepLabv3+, PSANet, Mask R-CNN, and original Swin-Unet models, indicating its effectiveness for target recognition and segmentation. The crop masks obtained through segmentation are used to obtain a weed mask through a morphological processing algorithm. Because the weed region can be obtained directly from the maize mask, only the maize seedlings must be labeled in the training

set, greatly reducing the labor required. The method can efficiently and accurately identify and segment maize and weeds in a complex maize field environments, even where crops and weeds in some overlap.

(2) The proposed model exhibited a higher inference speed than the original Swin-Unet, DeepLabv3+, and PSANet models; the processing time for each frame was 5.28×10^{-2} s. Hence, the proposed method is sufficiently fast for real-time data processing in applications such as vision for a weeding robot.

(3) The proposed model in this paper, compared to similar studies, shows a slight improvement but lacks significant advantages. In future research, we will further enhance the model structure to improve the practicality of the method and develop a more efficient and accurate weed identification approach.

Author Contributions: Conceptualization and methodology, J.G.; investigation, Y.Z.; writing—original draft preparation, J.Z.; grammar corrections, K.M.; test field support, G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for this study has been provided by the Key Research and Development Program of Shandong Province (Major Innovative Project in Science and Technology) (2020CXGC010804), and the Shandong Provincial Natural Science Foundation (ZR2021MC026).

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the anonymous reviewers for their critical comments and suggestions for improving the manuscript.

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

References

1. Zhang, Z.P. Development of chemical weed control and integrated weed management in China. *Weed Biol. Manag.* **2003**, *3*, 197–203. [CrossRef]
2. Li, X. Outstanding problems and management countermeasures in weed control on farmland in China in recent years. *Plant Prot.* **2018**, *44*, 77–84.
3. Machleb, J.; Peteinatos, G.G.; Kollenda, B.L.; Andújar, D.; Gerhards, R. Sensor-based mechanical weed control: Present state and prospects. *Comput. Electron. Agric.* **2020**, *176*, 105638. [CrossRef]
4. Zhang, S. Weed control technologies in major crop fields in China. *J. Weed Sci.* **2020**, *38*, 50–55.
5. Duan, X.; Han, J.; Ba, J.; Shi, J.; Zhang, Y.; Kang, L.; Wen, Y. Current status and development trend of chemical weed control in corn 417 fields. *Hortic. Seedl.* **2019**, *39*, 54–56.
6. Saha, D.; Cregg, B.M.; Sidhu, M.K. A review of non-chemical weed control practices in Christmas tree production. *Forests* **2020**, *11*, 554. [CrossRef]
7. Muola, A.; Fuchs, B.; Laihonen, M.; Rainio, K.; Heikkonen, L.; Ruuskanen, S.; Saikkonen, K.; Helander, M. Risk in the circular food economy: Glyphosate-based herbicide residues in manure fertilizers decrease crop yield. *Sci. Total Environ.* **2021**, *750*, 141422. [CrossRef]
8. Yuan, H.; Zhao, N.; Cheng, M. Research progress and prospect of weed identification in the field based on image processing. *J. Agric. Mach.* **2020**, *51* (Suppl. S2), 323–334.
9. Utstumo, T.; Urdal, F.; Brevik, A.; Dørum, J.; Netland, J.; Overskeid, Ø.; Berge, T.W.; Gravdahl, J.T. Robotic in-row weed control in vegetables. *Comput. Electron. Agric.* **2018**, *154*, 36–45. [CrossRef]
10. Huang, L.; Liu, W.; Huang, W.; Zhao, J.; Song, F. Remote sensing monitoring of powdery mildew in winter wheat by combining wavelet analysis and support vector machine. *J. Agric. Eng.* **2017**, *33*, 188–195.
11. Zhai, Z.; Xu, Z.; Zhou, X.; Wang, L.; Zhang, J. Identification of cotton blind toon weevil hazard classes based on plain Bayesian classifier. *J. Agric. Eng.* **2015**, *31*, 204–211.
12. Liang, X.; Chen, B.; Li, M.; Wei, C.; Feng, J. A dynamic counting method for cotton rows based on HOG features and SVM. *J. Agric. Eng.* **2020**, *36*, 173–181.
13. Chen, S.; Wu, S.; Yu, X. Identification of buckwheat diseases based on convolutional neural networks combined with image processing techniques. *J. Agric. Eng.* **2021**, *37*, 155–163.
14. Jiang, H.; Zhang, C.; Qiao, Y.; Zhang, Z.; Zhang, W.; Song, C. CNN feature based graph convolutional network for weed and crop recognition in smart farming. *Comput. Electron. Agric.* **2020**, *174*, 105450. [CrossRef]
15. Zhou, L.; Mu, H.; Ma, H.; Chen, G. Remote sensing yield estimation of winter wheat in northern China based on convolutional neural network. *J. Agric. Eng.* **2019**, *35*, 119–128.

16. Peng, W.; Lan, Y.; Yue, X.; Cheng, Z.; Wang, L.; Cen, Z.; Lu, Y.; Hong, J. Research on weed identification in rice fields based on deep convolutional neural network. *J. South China Agric. Univ.* **2020**, *41*, 75–81.
17. Meng, Q.; Zhang, M.; Yang, X.; Liu, Y.; Zhang, Z. Identification of corn seedlings and weeds based on lightweight convolution combined with feature information fusion. *J. Agric. Mach.* **2020**, *51*, 238–245, 303.
18. Wang, C.; Wu, X.; Zhang, Y.; Wang, W. Weed identification in corn field scenes based on shift window Transformer network. *J. Agric. Eng.* **2022**, *38*, 133–142.
19. Takahashi, R.; Matsubara, T.; Uehara, K. Data augmentation using random image cropping and patching for deep CNNs. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 2917–2931. [CrossRef]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
21. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the Computer Vision–ECCV 2022 Workshops, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
22. Xiao, X.; Zhang, D.; Hu, G.; Jiang, Y.; Xia, S. CNN–MHSA: A Convolutional Neural Network and multi-head self-attention combined approach for detecting phishing websites. *Neural Netw.* **2020**, *125*, 303–312. [CrossRef]
23. Sheng, C.; Wang, L.; Huang, Z.; Wang, T.; Guo, Y.; Hou, W.; Xu, L.; Wang, J.; Yan, X. Transformer-Based Deep Learning Network for Tooth Segmentation on Panoramic Radiographs. *J. Syst. Sci. Complex.* **2023**, *36*, 257–272. [CrossRef]
24. Alex, K.; Ilya, S.; Geoffrey, E.H. Image net classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90.
25. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Drop Block: A regularization method for convolutional networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 2–8 December 2018; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 10750–10760.
26. Morid, M.A.; Borjali, A.; Del Fiol, G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Comput. Biol. Med.* **2021**, *128*, 104115. [CrossRef] [PubMed]
27. Hasan, A.S.M.M.; Sohel, F.; Diepeveen, D.; Laga, H.; Jones, M.G. A survey of deep learning techniques for weed detection from images. *Comput. Electron. Agric.* **2021**, *184*, 106067. [CrossRef]
28. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR: New York, NY, USA, 2021; pp. 10347–10357.
29. Wang, Z.; Qu, S. Real-time semantic segmentation network based on attention mechanism and multi-scale pooling. *Comput. Eng.* **2023**, 1–11. [CrossRef]
30. Gao, J.J.; Zhang, X.; Guo, Y.; Liu, Y.K.; Guo, A.N.Q.; Shi, M.O.M.; Wang, P.; Yuan, Y. Research on the optimization method of pest image instance segmentation by incorporating Swin Transformer. *J. Nanjing For. Univ. (Nat. Sci. Ed.)* **2023**, *47*, 1–10.
31. Garibaldi-Márquez, F.; Flores, G.; Mercado-Ravell, D.A.; Ramírez-Pedraza, A.; Valentín-Coronado, L.M. Weed Classification from Natural Corn Field-Multi-Plant Images Based on Shallow and Deep Learning. *Sensors* **2022**, *22*, 3021. [CrossRef]
32. 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]