agronomy

MDPI

Article

A Study of Kale Recognition Based on Semantic Segmentation

Huarui Wu [1,2,3], Wang Guo [1,2,3,*], Chang Liu [1,2,3] and Xiang Sun [1,2,3,*]

[1] National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China; wuhr@nercita.org.cn (H.W.); liuchang@nercita.org.cn (C.L.)
[2] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
[3] Key Laboratory of Digital Village Technology, Ministry of Agriculture and Rural Affairs, Beijing 100125, China
* Correspondence: guow@nercita.org.cn (W.G.); sunx@nercita.org.cn (X.S.)

Abstract: The kale crop is an important bulk vegetable, and automatic segmentation to recognize kale is fundamental for effective field management. However, complex backgrounds and texture-rich edge details make fine segmentation of kale difficult. To this end, we constructed a kale dataset in a real field scenario and proposed an UperNet semantic segmentation model with a Swin transformer as the backbone network and improved the model according to the growth characteristics of kale. Firstly, a channel attention module (CAM) is introduced into the Swin transformer module to improve the representation ability of the network and enhance the extraction of kale outer leaf and leaf bulb information; secondly, the extraction accuracy of kale target edges is improved in the decoding part by designing an attention refinement module (ARM); lastly, the uneven distribution of classes is solved by modifying the optimizer and loss function to solve the class distribution problem. The experimental results show that the improved model in this paper has excellent performance in feature extraction, and the average intersection and merger ratio (mIOU) of the improved kale segmentation can be up to 91.2%, and the average pixel accuracy (mPA) can be up to 95.2%, which is 2.1 percentage points and 4.7 percentage points higher than the original UperNet model, respectively, and it effectively improves the segmentation recognition of kale.

Keywords: kale; semantic segmentation; Swin transformer; UperNet

check for updates

Citation: Wu, H.; Guo, W.; Liu, C.; Sun, X. A Study of Kale Recognition Based on Semantic Segmentation. *Agronomy* **2024**, *14*, 894. https://doi.org/10.3390/agronomy14050894

Academic Editor: Francis Drummond

Received: 2 April 2024
Revised: 18 April 2024
Accepted: 18 April 2024
Published: 25 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Kale is an important bulk vegetable species in China and is well liked by consumers because it is rich in various nutrients required by the human body, and it is very beneficial to human health. The main growth characteristics of kale are its outer leaves and leaf bulb, whose growth traits affect the field water and fertilizer management measures, and ultimately affect the overall yield of kale. Therefore, it is important to keep the growth status of the outer leaves and leaf bulbs of kale abreast to reduce the risk of damage to kale.

In recent years, with the rapid development of deep learning, semantic segmentation [1], target detection [2], and image classification [3] have also made significant progress. Among them, in computer vision, semantic segmentation is a very important direction, and the main method used is to judge the category that this image belongs to by the pixels that have been labelled in the image. Field kale images have color, texture, and spatial structure information. Traditional image processing algorithms, such as pixel-level cluster-based segmentation, pixel-level threshold-based segmentation, and pixel-level decision tree-based classification, usually use the underlying features of the image for segmentation. These traditional algorithms may face the problem of poor accuracy in image segmentation tasks because they mainly focus on the underlying features and ignore higher-level semantic information. With the rise of deep learning techniques, modern image segmentation methods are increasingly favoring the use of deep learning models, such as convolutional neural networks, which are capable of learning higher level abstract features to achieve better

performance in image segmentation tasks. Deep learning methods are able to improve segmentation accuracy by automatically learning feature representations in images through end-to-end learning. Since the introduction of fully convolutional neural networks, many classical semantic segmentation networks have emerged, including FCN [4], Unet [5], PSPnet [6], DeepLab [7], DeepLabv3+ [8], and so on. These classical networks have had a profound impact on later semantic segmentation research, and their applications on the agricultural field have gradually increased. Junrui Xue et al. [9] proposed an image segmentation method based on improved FCN-8s and obtained a date class intersection and merger ratio of 93.50% and a segmentation speed of 16.20 frames/s. Song et al. [10] used DeepLabV3 + based on Resnet-101 to enable the best segmentation of fruit calyx, branches, and filaments in kiwifruit canopy images. Hongjie et al. [11] proposed a method to segment crops in UAV remote sensing images by replacing the ordinary convolution in the ASPP module with a depth-separable convolution and by adding a double attention mechanism to improve the DeepLabv3+ model; the results show that the pixel accuracy of this method reached 93.9%, and the average intersection and merger ratio was 83.3%. Compared with convolutional neural networks, deep neural networks based on the transformer architecture [12] with self-attention mechanism have achieved excellent results in natural language processing [13] problems facing time-series data in recent years. The self-attention mechanism is able to increase the feature weights in the input linear data so that it can better extract the feature information; thus, scholars have begun to use it in the field of image processing. However, the Swin transformer [14] is a deep neural network based on the self-attention mechanism. Swin transformer [14] architecture is thus proposed; the window self-attention mechanism of a Swin transformer enables the transformer architecture to be well applied to the task of processing two-dimensional and above images and videos with large input data. Liu et al. [14] proposed the Swin transformer model by restricting windows to control the interaction between different slices, and the method also greatly reduced the computational effort of the network.

Semantic segmentation of vegetable crops, such as kale, is a critical task for efficient field management and yield optimization. However, current methodologies face significant challenges when segmenting the leaf bulb and outer leaves of kale against complex backgrounds, often resulting in low accuracy and compromised real-time performance. Traditional machine learning techniques have struggled to capture the intricate details and variations present in kale images, leading to suboptimal segmentation results.

In this context, the following limitations of existing research are summarized by this study: (1) insufficient datasets: the lack of comprehensive datasets capable of capturing kale diversity under real-world conditions hinders the training and validation of robust segmentation models [15]; (2) limited feature representation: existing methods often cannot effectively represent and extract the unique features of kale bulbs and outer leaves, especially in scenes with complex backgrounds [11]; (3) edge detection accuracy: the accuracy of edge detection of kale leaves often falls short of the requirements, which are crucial for accurate segmentation [10]; (4) class imbalance: the class imbalance problem during training has not been fully resolved, which may lead to biased model performance [14].

To overcome these limitations, the main contributions of this work are summarized as follows [16]:

- Building upon the UperNet architecture, a tailored semantic segmentation framework is developed that considers the specific growth characteristics of kale leaves.
- By adopting the Swin transformer as our backbone network, the channel attention module (CAM) is introduced to significantly improve the network's ability to represent and extract information about the kale's outer leaves and leaf bulb.
- An ARM within the decoding part of the framework is designed to refine the edge detection accuracy of kale's target areas.
- The issue of class imbalance is addressed by modifying the optimizer and loss function during the network training phase, ensuring a more balanced class distribution and improved learning dynamics.

The remainder of this article is organized as follows. Section 2 briefly summarizes related research. Section 3 provides technical details of model training. Section 4 provides a comprehensive discussion of the experiments. Finally, Section 5 outlines some concluding remarks and suggestions for future research.

2. Materials and Methods

2.1. Experimental Data

The image data in this study were obtained from the Precision Agriculture Experimental Base of the National Agricultural Informatization Engineering and Technology Research Centre in Xiaotangshan, Beijing, and the selected kale variety was Zhonggan-21, which was annotated according to the Pascal Voc dataset format using the open-source image annotation tool Labelme (v4.5.6), which was then preprocessed synchronously with the annotated images and their original images. In order to further expand the data samples, the randomly cropped images were subjected to data enhancement [17] operations, such as horizontal flipping, vertical flipping, brightness adjustment, and adding Gaussian noise. Finally, a dataset with a size of 5000 images was obtained. The images were randomly divided into training and test sets in an 8:2 ratio. The combination of these image enhancement methods simulates the changes in the shooting angle and light intensity during image acquisition, increases the diversity of training samples, and improves the robustness and generalization ability of the model. An example of using the LabelMe annotation is shown in Figure 1, where the black part is the background, the red part is the outer leaf, and the green part is the leaf bulb.

Figure 1. Comparison of kale images before and after labelling.

2.2. Experimental Methods

2.2.1. An Improved Semantic Segmentation Model

In this paper, UperNet is used as an implementation framework for semantic segmentation. The feature extractor of this semantic segmentation architecture is set as a Feature Pyramid Network (FPN) based on the Swin transformer backbone. It utilizes the multi-level feature representation obtained by the Swin transformer to represent the corresponding pyramid hierarchies, using a top-down FPN architecture with horizontal connectivity and downsampling ratios that are consistent with those of the Swin transformer. The Pyramid Pooling Module [18] (PPM) is located before the top-down branches of the FPN and is connected to stage 4 of the Swin transformer network, and the PPM is capable of delivering effective global a priori feature representations that are highly compatible with the FPN architecture. This form of architecture can effectively cooperate with the hierarchical feature expression obtained by the Swin transformer to achieve a better semantic segmentation effect based on the fusion of high- and low-level semantic information. The feature extraction capability of the network is further improved by incorporating EAM on the Swin transformer module. In the feature fusion module, all the hierarchical features output from the FPN are adjusted to the same size by bilinear interpolation, and then a convolutional layer is applied to fuse the features from different levels. The target segmentation header is appended to the fused feature map with a separate convolutional layer in front of each

classifier. All additional non-classifier convolutional layers have a batch normalization of 512 channel outputs [19], and ReLU [20] activation functions are applied. STM is added to the encoder part to avoid omission of the extracted information. The model output is a category mask generated from the pixel classification prediction labels, which in turn yields a segmentation map. This results in early recognition of kale while obtaining a fine segmentation of its target region. The improved UperNet semantic segmentation framework is shown in Figure 2.

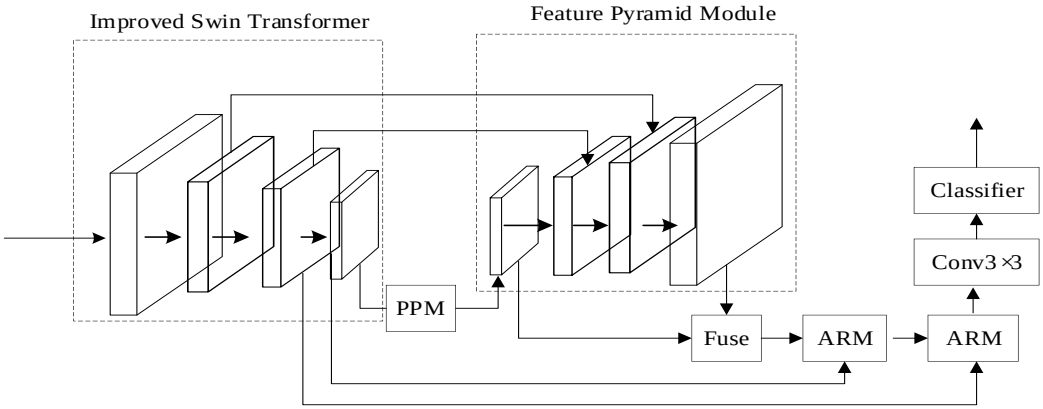

Figure 2. Improved UperNet model structure.

2.2.2. Swin Transformer

The Swin transformer serves as the backbone network for our visual task, leveraging a fully self-attentive mechanism. It surpasses traditional convolutional neural network architectures in semantic segmentation tasks. Unlike standard transformer architectures, the Swin transformer excels in constructing hierarchical feature representations essential for pixel-dense prediction. Its modeling capability is notably enhanced by the shift-window-based self-attention mechanism.

Key features of a Swin transformer: (1) hierarchical feature representation: a Swin transformer constructs hierarchical feature representations crucial for precise pixel-level prediction, contributing to superior segmentation outcomes. (2) Shift-window-based self-attention: this mechanism computes self-attention locally within non-overlapping windows of the segmented image, enabling efficient cross-window connectivity and expediting model inference. The structure of the Swin transformer block is shown in Figure 3. The first Swin transformer block maintains a constant number of input and output tokens at [H/4, W/4], which is designated as Stage 1 alongside the linear embedding layer. As data flows into the Swin transformer block, multi-head self-attention computation for the window commences.

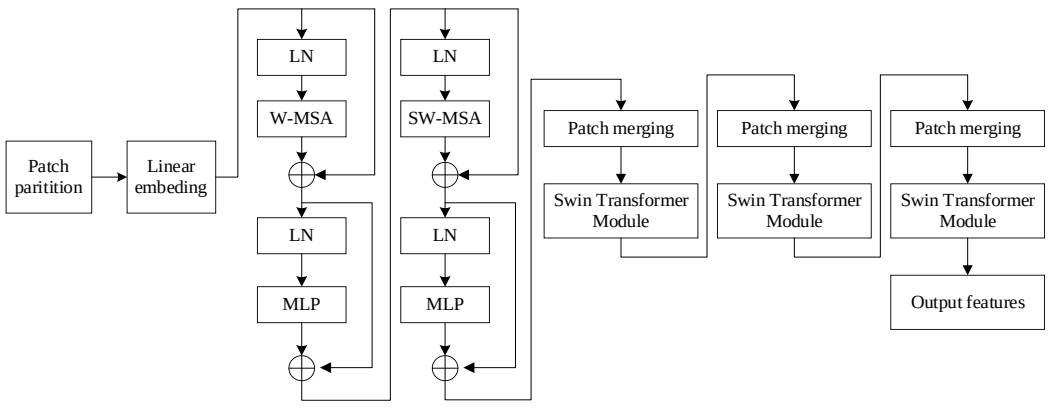

Figure 3. Structure of Swin transformer.

2.2.3. Channel Attention Module (CAM)

Attentional mechanisms play a crucial role in semantic segmentation tasks by enhancing the focus on important features while suppressing distracting ones. However, in the case of kale, which grows in complex outdoor environments influenced by various factors such as soil, light changes, weeds, and leaf texture similarity, accurately capturing the intricate relationship between kale and its surroundings is challenging. While a Swin transformer excels in integrating spatial context information, it may not accurately capture the complex interactions and dependencies between channels.

To address this limitation, we propose the channel attention module (CAM), which is inserted after the W-MSA and SW-MSA stages of the Swin transformer. CAM is used to assign different weights to each channel of the input feature map based on their importance in representing relevant information about the kale's morphology. This mechanism allows the network to emphasize key features associated with the outer leaves and leaf bulb, such as texture, shape, and color variations, while attenuating distractions from the background or other non-essential elements in the image. By adaptively adjusting the weights of different channels, the CAM module effectively enhances the discriminative power of the network, enabling it to capture and highlight crucial details about the kale's structure and appearance, ultimately improving segmentation performance, particularly in delineating the boundaries and contours of the outer leaves and leaf bulb with greater accuracy and precision. The structure of the CAM module is illustrated in Figure 4. Initially, CAM performs a squeeze operation on the feature map obtained after self-attention to extract global features at the channel level. Subsequently, an excitation operation is performed on these global features to learn the relationship between each channel and obtain the weights of different channels. Finally, the original feature map is multiplied by these weights to generate the final features. This module preserves the channel dimension while compressing the spatial dimension, thereby enhancing the interaction of information between input channels.

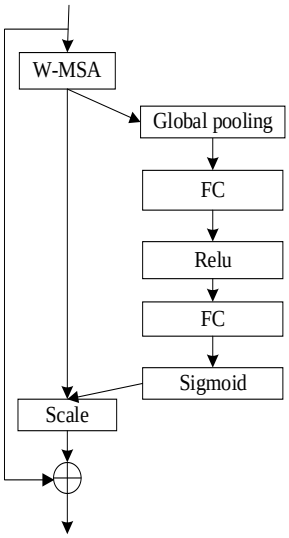

Figure 4. Structure of the channel attention module.

By incorporating CAM into both W-MSA and SW-MSA stages, the correlation between channels can be better explored and utilized at different levels of the model. Moreover, since W-MSA mainly focuses on global spatial relations and SW-MSA on local spatial relations, adding channel attention to both stages allow for better integration of global and local spatial information with channel correlation. The Swin transformer block enhanced with CAM is depicted in Figure 5.

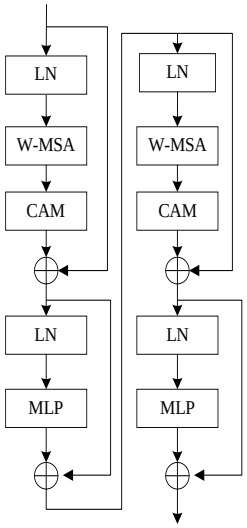

Figure 5. Structure of the improved Swin transformer block.

2.2.4. Attention Refinement Module (ARM)

To extract detailed information effectively from kale images, we introduce an attention refinement module (ARM) [21,22], as depicted in Figure 6. This module plays a crucial role in improving the model's ability to capture intricate details and semantic context relevant to kale leaf segmentation. The ARM takes high-order features extracted from earlier layers of the network and refines them using channel attention mechanisms. This step allows the model to focus on important features specific to kale leaves, enhancing their representation in the feature space. In parallel, the ARM processes low-order features, which typically capture fine details and edges in the image. Through spatial attention mechanisms, these features are recalibrated to better align with the semantic information of high-order features. This refinement process ensures that important details related to kale leaves are preserved and accurately represented in the final feature maps. The refined high-order and low-order features are then fused together to create a comprehensive representation of the kale leaves. This fusion process leverages both the semantic context captured by high-order features and the fine details captured by low-order features, resulting in a more robust and accurate representation of kale leaves in the dataset. By refining both high-order and low-order features, the ARM enhances the model's ability to segment kale leaves accurately. It ensures that important details and semantic context are effectively captured, leading to improved segmentation performance, especially in challenging scenarios with complex backgrounds or occlusions.

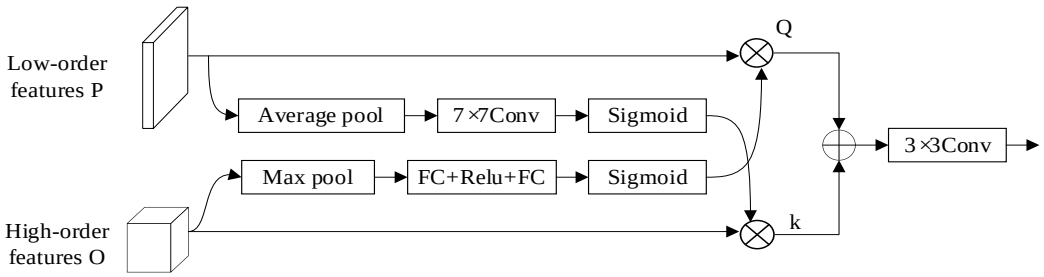

Figure 6. Structure of the attention refinement module.

Firstly, the lower-order features are averaged and pooled across the entire channel axis to compute the average value of each pixel along the channel axis. Subsequently, a 7×7 convolutional operation followed by a sigmoid function is applied to capture local spatial dependencies. On the other hand, the high-order features are maximally pooled along a single channel to generate a feature vector, ensuring that each channel of the high-order features contributes its representative response to the feature vector. The outputs from

the spatial and channel attention branches are then fused, and the number of channels is adjusted using a 3×3 convolution operation to obtain the final output of the module.

This multi-step process enables the ARM to refine feature representations effectively, capturing both spatial and channel-wise dependencies to recover fine details while preserving segmentation accuracy.

3. Model Training and Performance Evaluation

3.1. Test Platform Configuration

The experimental environment of this study is based on a computer with a Windows 10 operating system, an AVX2 central processing unit (CPU), and an NVIDIA GeForce RTX 2080 Ti graphics processing unit (GPU). The programming language is Python 3.8, the general-purpose parallel computing architecture is CUDA 10.2, the deep neural network (GPU) acceleration library is cuDNN 7.6.5, the computer vision library is OpenCV 4.7.0, and the Pytorch 1.7.1 deep learning framework is used to build and adjust the parameters of the segmentation model in this paper. All the training and testing processes of the models were completed on the test platform built in this study to ensure the consistency of the comparison conditions.

3.2. Model Training Strategies

The model training is conducted end-to-end, with the inputs being the original images and the outputs being the corresponding recognition segmentation maps, with no human intervention in the process. The Swin transformer backbone network and the semantic segmentation framework are combined into a holistic model through a decoder-encoder structure and are trained at the same time. The backbone network acts as a decoder responsible for feature transformation and extraction, while the semantic segmentation framework acts as an encoder to reconstruct and fuse the output features of the backbone network and generate classification predictions based on them. During training, the optimizer is used to update and compute the network parameters that affect the model's training and model's output, so that they approach or reach the optimal values, thus minimizing the loss function. In the improved UperNet model, the AdamWarmup, Adam and Adadelta optimizers are used to fit the data, as shown in Figure 7, which shows a comparison of the change curve of loss during the training process of the model using different optimizers; it can be seen that the Adam and Adadelta optimizers have a general effect, the fluctuation of the value of the loss function is relatively large, and it is obvious that the AdamWarmup optimizer is more effective. It can be seen that the Adam and Adadelta optimizers are generally effective, and the value of the loss function fluctuates during the training process.

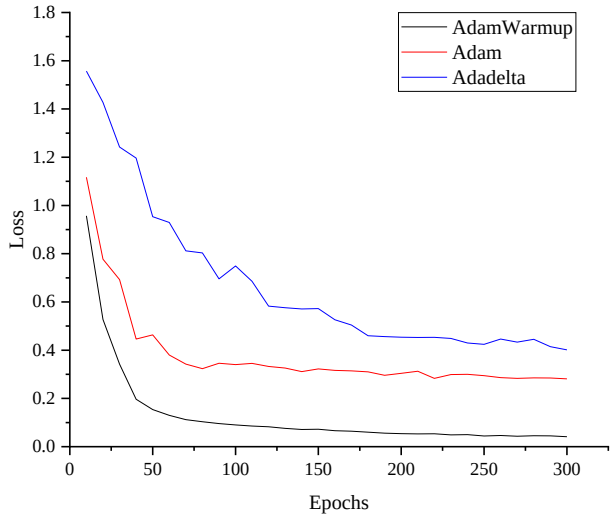

Figure 7. Loss function during training of different optimizers.

In this paper, we choose the exponential decay learning rate strategy with Warmup [23] using AdamWarmup as the model optimizer, with an initial learning rate of 1×10^{-3} and the minimum value of the learning rate set to 1×10^{-5}. During the training process, this learning rate strategy can automatically adjust the learning rate of each Epoch; during the first training Epoch, the model can quickly correct the data distribution; after the first training Epoch, the learning rate is set to a smaller value than the initial learning rate to ensure the model has good convergence. After the first Epoch, the learning rate is set to a smaller value than the initial learning rate to ensure good convergence of the model. After the model is relatively stable, the learning rate of the model is gradually increased to the preset initial learning rate to accelerate the convergence speed of the model, so that the model training effect is better in the late stage of model training, when the model can learn less new knowledge, and a larger learning rate will destroy the stability of the existing one. Therefore, in the subsequent training Epoch, the learning rate of the model is gradually reduced to approach the minimum value. The cross-entropy loss function is used to measure the distance between the predicted probability distribution of pixel categories and the probability distribution of real label categories during the training process, which is calculated as in Equation (1):

$$Loss = \frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{N} h(b_i)\log(p_{ic}) \tag{1}$$

where M is the number of pixels; N is the number of categories; i is the current pixel; c is the current category; b_i is the true labelling category of pixel I; h is the 0–1 probability distribution function, and it is 1 if $b_i = c$ and 0 otherwise; pic is the predicted probability of pixel i belonging to the class c, which is obtained from the computation of the predicted category score by the Sigmoid function. The training performance of the model is measured by the calculation of the loss function during the iteration process, and the weights are adjusted by back propagation so that the error distance represented by the loss values is gradually reduced to achieve the training goal.

3.3. Indicators for Model Evaluation

The performance evaluation metrics of the model in this paper mainly use the mean pixel accuracy (*mPA*), mean intersection, and merger ratio (*mIOU*) to assess image segmentation model performance. Among them, *mPA* measures the average accuracy of the model in correctly predicting the pixels of each category—the higher the *mPA* value indicates, the better the pixel prediction accuracy of the model—which is calculated as shown in Equation (2). *mIOU* measures the segmentation accuracy of the model by calculating the ratio of the intersection and concatenation of the predicted segmentation results to the real segmentation results, with higher metrics indicating that the predicted results overlap with the real results and the model's segmentation effect is better. Its definition is shown in Equation (3).

$$mPA = \frac{1}{N+1}\sum_{i=0}^{N}\frac{n_{ii}}{t_i} \tag{2}$$

$$mIoU = \frac{1}{N+1}\sum_{i=0}^{N}\frac{n_{ii}}{t_i + \sum_{j=0}^{N} n_{ji} - n_{ii}} \tag{3}$$

where N is the number of target categories segmented (in the case of no background), n_{ii} denotes the number of correctly categorized pixels, denotes the number of pixels in target category i, n_{ji} denotes the number of pixels in target category i predicted to be category j, and n_{ji} denotes the number of pixels in target category j predicted to be category i.

4. Results and Analysis

4.1. Ablation Experiment

In order to verify the effectiveness of the improved semantic segmentation model for the kale segmentation recognition algorithm, this paper designs ablation experiments for

the improved UperNet model based on the original model UperNet. Table 1 shows a data comparison of kale segmentation recognition performance. In order to do so, the following steps were followed: (1) use the original model UperNet-Resnet network structure; (2) replace the backbone network of the UperNet model with the Swin transformer; (3) add the CAM on the basis of UperNet-Swin transformer; (4) add ARM to the UperNet-Swin transformer; (5) add CAM and ARM to the UperNet-Swin transformer. The comparison of the data in Table 1 illustrates that the network model in this paper will pay more attention to the detail information after the improvement of the network model, and mPA and mIOU are both improved, which proves that the modules are effective for network improvement.

Table 1. Ablation experiments.

Programs				mPA (%)	mIOU (%)
ResNet	Swin Transformer	CAM	ARM		
√	×	×	×	87.8	85.9
×	√	×	×	90.5	89.1
×	√	√	×	93.3	90.7
×	√	×	√	91.7	90.0
×	√	√	√	95.2	91.2

4.2. Comparison of Segmentation Recognition Effect of Different Models

In order to further validate the effectiveness of the model in this paper, a comparative analysis of the segmentation performance of the model in this paper with FCN, Unet, PSPNet, Deeplabv3+, UperNet (Resnet), and UperNet (Swin transformer) was performed on a homemade kale dataset, and the results of the experiments are shown in Table 2. The improved UperNet-Swin transformer model achieves 91.2% mIOU for kale, which is higher than the FCN, Unet, PSPNet, Deeplabv3+, UperNet (Resnet), and UperNet (Swin transformer) models in the following order of mIOU for kale: 11%, 6.9%, 12.1%, 4.7%, 5.3%, and 2.1%, respectively. In this paper, mPA reached 95.2%, as above, and improved by 13.6%, 9.3%, 12.9%, 6.8%, 7.4%, and 4.7%, respectively. It can be concluded that the model segmentation effect of this paper has shown a significant improvement, and the main reason for the better segmentation effect of the model in this paper is that this model introduces the attention module and displays improvement in the encoder site, which strengthens the judgement of the target features at each stage and obtains the effective global contextual information.

Table 2. Comparative experiments with different models.

Module	Backbone	mPA/%	mIoU/%
FCN	ResNet	81.6	80.2
UNet	ResNet	85.9	84.3
PSPNet	ResNet	82.3	79.1
DeepLabv3+	ResNet	88.4	86.5
UperNet	ResNet	87.8	85.9
UperNet	Swin transformer	90.5	89.1
proposed	Swin transformer + ECA	95.2	91.2

4.3. Visualization and Analysis

To verify the effectiveness and interpretability of this paper's method in the field kale image segmentation task, this paper combines Table 2 to select the three models of UNet, DeepLabv3+, and UperNet (Swin transformer)—which have better segmentation and recognition effects—for visual comparative analysis (shown in Figure 8).

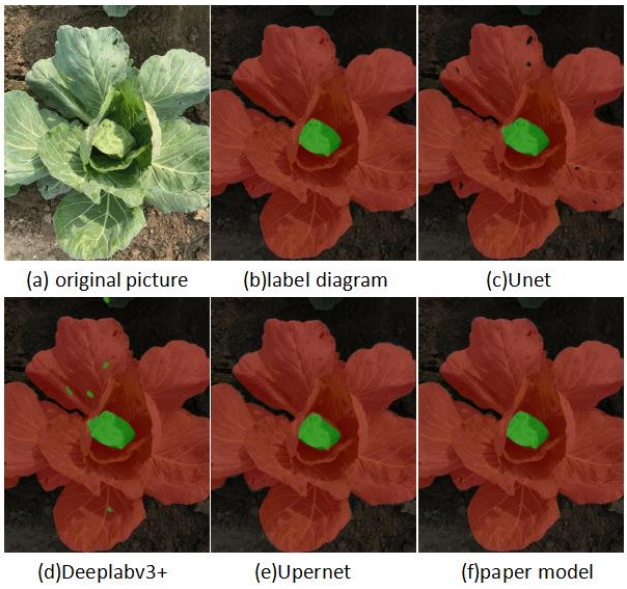

Figure 8. Visual analysis of the model.

According to Figure 8, the following conclusions can be drawn. Segmentation details:

(1) The proposed model demonstrates superior segmentation details compared to the other models. Specifically, UNet exhibits a noticeable pixel area missing in the outer leaf region, indicating some information loss. DeepLabv3+ misidentifies the leaf sphere, resulting in information misjudgment. Although UperNet (Swin transformer) also displays misidentification in the leaf sphere area with slight information loss, our model achieves segmentation results that closely resemble the ground truth labels.

(2) Precision in leaf sphere identification: While the UperNet (Swin transformer) model may not be precise enough in identifying the leaf sphere area, our model's segmentation effect is remarkably similar to the ground truth when observed by the human eye. We note that mis-segmentation typically occurs near the edges of the target, where only a few pixels are misclassified. However, these small discrepancies have minimal impact on the overall segmentation quality and are challenging to detect without close inspection.

(3) Generalization ability: Combining the visual analysis with the quantitative results presented in Table 2, the proposed method exhibits stronger practical generalization ability in terms of pixel recognition accuracy across different growth states of kale. The model's ability to accurately segment kale leaves under varying conditions highlights its robustness and effectiveness for real-world applications.

In summary, the detailed visual analysis of the model confirms the efficacy of the proposed method for kale image segmentation. The close resemblance of the segmentation results to the ground truth labels underscores the model's accuracy and reliability. Additionally, the model demonstrates strong generalization ability across diverse kale growth states, further validating its practical utility.

5. Conclusions

This paper presents an enhanced semantic segmentation model aimed at accurately segmenting kale leaves in complex environments. Leveraging the UperNet semantic segmentation framework, our approach integrates a Swin transformer as the backbone network and introduces the channel attention module (CAM) to enhance feature extraction from kale leaves. Additionally, an attention refinement module (ARM) is designed to refine target features and improve edge extraction. Finally, optimization techniques are employed to mitigate the issue of an uneven class distribution. Through extensive comparative experiments, the proposed method outperforms other semantic segmentation models, achieving mIOU and MPA scores of 91.2% and 95.2%, respectively. The successful implementation of our method enables accurate segmentation recognition of kale in diverse field conditions,

with significant implications for field management practices, facilitating timely decision making and risk reduction in the kale industry.

In the future, we plan to further optimize the model and enhance recognition accuracy through technologies such as reinforcement learning and advanced machine learning.

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

Funding: This study was supported by the National Key Research and Development Program of China under Grant 2023YFD2001205, the China Agriculture Research System of MOF and MARA under Grant CARS-23-D07, and the Beijing Academy of Agriculture and Forestry Sciences Key Technology Innovation Capacity Building Special Project: Key Technologies for Soil Conservation of Cultivated Land in the Beijing-Tianjin-Hebei Region under Grant KJCX20230219.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to the privacy policy of the authors' institution.

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

References

1. Huang, L.; He, M.; Tan, C.; Jiang, D.; Li, G.; Yu, H. Jointly network image processing: Multi-task image semantic segmentation of indoor scene based on CNN. *IET Image Process.* **2020**, *14*, 3689. [CrossRef]
2. Zhao, E.; Dong, L.; Dai, H. Infrared maritime target detection based on edge dilation segmentation and multiscale local saliency of image details. *Infrared Phys. Technol.* **2023**, *133*, 104852. [CrossRef]
3. Machado, G.R.; Silva, E.; Goldschmidt, R.R. Adversarial Machine Learning in Image Classification: A Survey Toward the Defender's Perspective. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
4. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; IEEE: Piscataway, NJ, USA, 2015.
5. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
6. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017.
7. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
8. 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 Computer Vision—ECCV 2018, Berlin, Germany, 8–14 September 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018.
9. Xue, J.; Wang, Y.; Qu, A.; Zhang, J.; Sun, H. Image segmentation method for Lingwu long jujube based on improved FCN-8s. *J. Agric. Eng.* **2021**, *37*, 191–197. [CrossRef]
10. Song, Z.; Zhou, Z.; Wang, W.; Gao, F.; Fu, L.; Li, R.; Cui, Y. Canopy segmentation and wire reconstruction for kiwifruit robotic harvesting. *Comput. Electron. Agric.* **2021**, *181*, 105933. [CrossRef]
11. Ren, H.J.; Liu, P.; Dai, C.; Shi, J.C. Crop Segmentation Method of Remote Sensing Image Based on Improved DeepLabv3+ Network. *Comput. Eng. Appl.* **2022**, *58*, 215.
12. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. [CrossRef] [PubMed]
13. Trappey, A.J.; Chang, A.C.; Trappey, C.V.; Chien, J.Y.C. Intelligent RFQ Summarization Using Natural Language Processing, Text Mining, and Machine Learning Techniques. *J. Glob. Inf. Manag.* **2022**, *30*, 1–26. [CrossRef]
14. Liu, Z.; Lin, Y.T.; Cao, Y.; Hu, H.; Wei, Y.X.; Zhang, Z.; Lin, S.; Guo, B.N. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
15. Hahn, C.; Howard, N.P.; Albach, D.C. Different Shades of Kale—Approaches to Analyze Kale Variety Interrelations. *Genes* **2022**, *13*, 232. [CrossRef] [PubMed]
16. Khan, H.; Hussain, T.; Khan, S.U.; Khan, Z.A.; Baik, S.W. Deep multi-scale pyramidal features network for supervised video summarization. *Expert Syst. Appl.* **2024**, *237*, 121288. [CrossRef]

17. Qian, H.; Huang, Z.; Xu, Y.; Zhou, Q.; Wang, J.; Shen, J.; Shen, Z. Very high cycle fatigue life prediction of Ti60 alloy based on machine learning with data enhancement. *Eng. Fract. Mech.* **2023**, *289*, 109431. [CrossRef]
18. Ke, X.; Li, J. U-FPNDet: A one-shot traffic object detector based on U-shaped feature pyramid module. *IET Image Process.* **2021**, *15*, 2146–2156. [CrossRef]
19. Kumar, P.; Hati, A.S. Convolutional neural network with batch normalisation for fault detection in squirrel cage induction motor. *IET Electr. Power Appl.* **2021**, *15*, 39–50. [CrossRef]
20. Shen, Z.; Yang, H.; Zhang, S. Optimal approximation rate of ReLU networks in terms of width and depth. *J. Math. Pures Appl.* **2022**, *157*, 101–135. [CrossRef]
21. Sekharamantry, P.K.; Melgani, F.; Malacarne, J. Deep learning-based apple detection with attention module and improved loss function in YOLO. *Remote Sens.* **2023**, *15*, 1516. [CrossRef]
22. Sekharamantry, P.K.; Melgani, F.; Malacarne, J.; Ricci, R.; de Almeida Silva, R.; Marcato Junior, J. A Seamless Deep Learning Approach for Apple Detection, Depth Estimation, and Tracking Using YOLO Models Enhanced by Multi-Head Attention Mechanism. *Computers* **2024**, *13*, 83. [CrossRef]
23. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.