agriculture

MDPI

Article

MSFCA-Net: A Multi-Scale Feature Convolutional Attention Network for Segmenting Crops and Weeds in the Field

Qiangli Yang, Yong Ye *, Lichuan Gu and Yuting Wu

School of Information and Computer, Anhui Agricultural University, Hefei 230036, China; qiangliyang@stu.ahau.edu.cn (Q.Y.); lcgu@sohu.com (L.G.); wwyt@ahau.edu.cn (Y.W.)
* Correspondence: yeyong@ahau.edu.cn; Tel.: +86-182-5602-8950

Abstract: Weed control has always been one of the most important issues in agriculture. The research based on deep learning methods for weed identification and segmentation in the field provides necessary conditions for intelligent point-to-point spraying and intelligent weeding. However, due to limited and difficult-to-obtain agricultural weed datasets, complex changes in field lighting intensity, mutual occlusion between crops and weeds, and uneven size and quantity of crops and weeds, the existing weed segmentation methods are unable to perform effectively. In order to address these issues in weed segmentation, this study proposes a multi-scale convolutional attention network for crop and weed segmentation. In this work, we designed a multi-scale feature convolutional attention network for segmenting crops and weeds in the field called MSFCA-Net using various sizes of strip convolutions. A hybrid loss designed based on the Dice loss and focal loss is used to enhance the model's sensitivity towards different classes and improve the model's ability to learn from hard samples, thereby enhancing the segmentation performance of crops and weeds. The proposed method is trained and tested on soybean, sugar beet, carrot, and rice weed datasets. Comparisons with popular semantic segmentation methods show that the proposed MSFCA-Net has higher mean intersection over union (MIoU) on these datasets, with values of 92.64%, 89.58%, 79.34%, and 78.12%, respectively. The results show that under the same experimental conditions and parameter configurations, the proposed method outperforms other methods and has strong robustness and generalization ability.

Keywords: semantic segmentation; weed segmentation; agricultural weed dataset; convolutional attention

Citation: Yang, Q.; Ye, Y.; Gu, L.; Wu, Y. MSFCA-Net: A Multi-Scale Feature Convolutional Attention Network for Segmenting Crops and Weeds in the Field. *Agriculture* **2023**, *13*, 1176. https://doi.org/10.3390/agriculture13061176

Academic Editors: Long He, Azlan Zahid and Md Sultan Mahmud

Received: 3 May 2023
Revised: 25 May 2023
Accepted: 30 May 2023
Published: 31 May 2023

1. Introduction

Agriculture is one of the fundamental human activities, which ensures the global food security. However, the weeds in farmland can cause severe damage to the growth and yield of crops, because weeds directly compete with crops for sunlight, water, and nutrients. In addition, they also become a source for spreading diseases and pests in crops [1]. Weed control helps in promoting sustainable agricultural development, thus improving the agricultural production efficiency, reducing the waste of agricultural resources, and protecting the ecological environment to achieve sustainable agricultural development [2]. Over the years, various weed control measures, such as agricultural prevention and control, plant quarantine, manual weeding, biological weed control, and chemical weed control have been explored to develop agricultural technology [3]. The conventional physical weeding operations are costly and inefficient. Currently, the chemical weed control methods are the most widely used methods [4]. However, the traditional chemical weed control methods involve spraying herbicides uniformly across the field. This not only leads to high costs but also causes environmental pollution due to the excessive herbicide use [5]. The development of information and automation technologies has opened a new era of weed control. It is of great significance to perform precise mechanical and chemical weed control measures to quickly and effectively eliminate weeds [6].

With the development of digital imaging technologies and the advancements in robotic intelligent agricultural machinery, such as field weed-removal robots, which utilize image processing technology, great results have been achieved [7,8]. In 2015, a German robotics company called Deepfield [9] launched the first generation of weeding robots that identify weeds by using cameras. The precise weed removal methods, such as selective weeding, specific point herbicide spraying, and intelligent mechanical hoeing effectively reduce the harm of pesticides and improve the quality of agricultural products [10]. Please note that weed identification is crucial for intelligent weed removal. The vision-based weed identification methods mainly use digital image processing techniques to differentiate various crops based on different features extracted from crop images [11]. Ahmed et al. [12] used support vector machines (SVM) to identify six types of weeds by using a database containing 224 images and achieved satisfactory accuracy under certain experimental conditions using a combination of optimal feature extractors. Sabzi et al. [13] used a machine vision prototype based on video processing and meta-heuristic classifiers to identify and classify potatoes and five types of weeds. Brilhador et al. [14] used edge detection techniques to detect weeds in ornamental lawns and sports turf, aiming to reduce pesticide usage. Various filters were tested, and the sharpening (I) filter with the aggregation technique and a cell size of 10 provided the best results. A threshold value of 78 yielded an optimal performance. However, slight differences in the results were observed between ornamental lawns and sports turf. Parra et al. [15] used UAVs with digital cameras to detect charlock mustard weed in alfalfa crops using RGB-based indices, which proved effective and avoided confusion with soil compared to NDVI. Combining RGB indices with NDVI reduced overestimation in weed identification. This methodology can generate weed cover maps for alfalfa and translate into herbicide treatment maps. However, these methods are unable to perform effectively in complex field environments. For instance, image acquisition in real environments may suffer from uneven exposure due to strong or weak lighting conditions, resulting in reduced recognition accuracy. Moreover, crops and weeds have different sizes and shapes, and the generalization ability of image processing systems in complex backgrounds is poor, resulting in suboptimal recognition results. Moreover, digital image processing techniques require manual feature selection, and the segmentation performance of the models is susceptible to human experience interference.

Recently, convolutional neural networks (CNN) have greatly promoted the progress of computer vision [16]. Contrary to the traditional machine learning algorithms, deep learning algorithms automatically perform feature selection and have a higher accuracy as well. These methods have been widely applied in agricultural image processing [17]. The CNNs have been used to predict and estimate the yield of mature stage rice based on remote sensing images acquired using unmanned aerial vehicles (UAVs) [18]. A deep learning-based robust detector for the real-time identification of tomato diseases and pests was proposed in [19]. Hall et al. [20] constructed a CNN for carrot and weed classification during the seedling stage. The authors used the texture information and shape features, significantly improving the accuracy of plant classification. Olsen et al. [21] constructed a large, public, multi-class deep-sea weed dataset and used ResNet50 for weed classification. Since the proposal of FCN [22], the image semantic segmentation models, such as UNet [23], DeepLabV3 [24], and DeepLabV3Plus [25], have emerged and widely applied in the agricultural weed segmentation field [26]. The semantic segmentation models quickly extract features from the crops and weeds, without requiring complex background segmentation and data model establishment during the extraction process. You et al. [27] proposed a segmentation network for segmenting the sugar beet crop. Yu et al. [28] proposed several networks for accurately detecting weeds in dogtooth grass plants. Sun et al. [29] fused near-infrared and RGB images into a four-channel image by analysing the feature distribution of a sugar beet dataset, and proposed a multi-channel depth-wise separable convolution-based segmentation and recognition method for sugar beet and weed images. The authors achieved real-time segmentation by using the MobileNet. Zou et al. [30]

proposed a simplified UNet-based semantic segmentation algorithm to separate weeds from soil and crops in images.

Please note that the aforementioned weed segmentation algorithms are typically based on popular semantic segmentation models, which use fine-tuning and conventional channel or spatial attention mechanisms to improve the performance of the networks. These attention mechanisms usually involve concatenating residuals or using 1×1 or 3×3 convolutions to implement channel- or spatial-wise attention in the network. However, these improvement methods neglect the role of multi-scale feature aggregation in network design. The previous literature shows that multi-scale feature aggregation is crucial for segmentation tasks [31,32]. Consequently, these approaches fail to effectively connect features in the spatial and channel dimensions of the upper and lower layers, and suffer from various problems, such as image illumination interference, differences in the size between crops and weeds, and mutual occlusion between crops and weeds. These problems severely affect the performance of field weed segmentation. Moreover, the existing deep learning-based weed segmentation models often require large amounts of data for training but the data for agricultural weed segmentation is scarce and difficult to obtain. It is noteworthy that the performance of these models is not efficient for small training sets.

In response to the shortcomings of previous research and existing problems, we design a weed segmentation algorithm that can be applied to various complex environments and crops, providing algorithmic support for intelligent weed control. This work proposes a field weed segmentation model based on convolutional attention mechanism. The proposed method uses large asymmetric strip convolution kernels to extract features. The proposed method achieves faster and more accurate field weed segmentation, as well as addresses multi-scale and complex background weed segmentation tasks.

The rest of the manuscript is organized as follows. In Section 2, we present the data used in this work and the proposed network. In Section 3, we present the experimental results and analysis. Section 4 discusses the research. Finally, this work is concluded in Section 5.

2. Materials and Methods

2.1. Dataset Collection

2.1.1. Soybean Dataset

The soybean dataset is acquired from a soybean field in the National High-Tech Agricultural Park of Anhui Agricultural University located in Luyang District, Hefei City, Anhui Province. We select soybean seedlings aged 15–30 days for data collection. The equipment used for image acquisition includes a DJI handheld gimbal, model Pocket 2. The acquisition device is positioned about 50 cm above the ground. The video resolution is set at 1920×1080 with a frame rate of 24 frames per second (fps). Afterwards, we extract frames from the video to obtain 553 images to construct the soybean dataset. In order to ensure faster training and convenient manual annotation, we resized the images to 1024×768. This resolution strikes a balance between computational efficiency and preserving sufficient visual details for accurate image analysis, making it a commonly used resolution in many computer vision applications and datasets. We randomly assign them to training, validation, and test sets in a ratio of 7:2:1. This allocation ratio allows for a reasonable balance between the training, validation, and testing requirements within the limited dataset. Selecting 70% of the data for the training set provides an adequate number of samples for model training. The validation set, comprising 30% of the data, is used to adjust the model's hyperparameters and fine-tuning. We reserved 10% of the data as the test set, providing a sufficient number of samples to accurately evaluate the model's performance. We manually annotated the images using the open-source tool Labelme. Each annotated image corresponds to an original image, with different colours representing different categories. The soybean seedlings are annotated in green, weeds are annotated in red, and the soil is annotated in black, as shown in Figure 1.

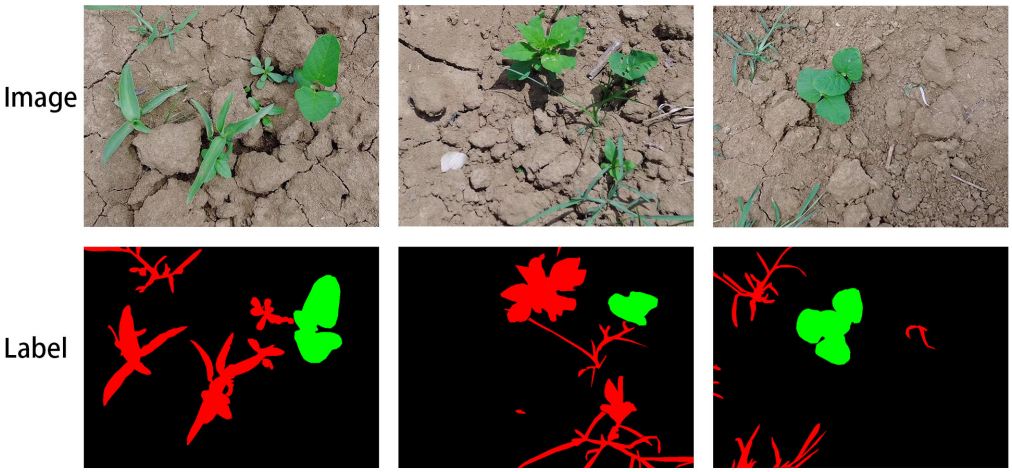

Figure 1. The original images and corresponding annotations (green: crop, red: weed) in the soybean dataset.

2.1.2. Sugar Beet Dataset

The sugar beet dataset is sourced from BoniRob [33]. The images in this dataset are captured at a sugar beet farm near Bonn, Germany. In 2016, a pre-existing agricultural robot was used to record the dataset, which focused on sugar beet plants and weeds. The robot was equipped with a JAI AD-130GE camera, with an image resolution of 1296 × 966 pixels. The camera is positioned underneath the robot's chassis, with a mounting height of approximately 85 cm above the ground. The data collection spanned over three months, with data being recorded approximately three times a week. The robot captured multiple stages of sugar beet plant during its growth. The official dataset contains tens of thousands of images. In this work, the labels are divided into three categories: sugar beet crops, all weeds, and background. For convenience, we use 2677 randomly selected images to create the sugar beet dataset. We randomly split the dataset into 70% training, 20% validation, and 10% test sets. As shown in Figure 2, green annotations represent sugar beet crop, red annotations represent weeds, and black annotations represent soil.

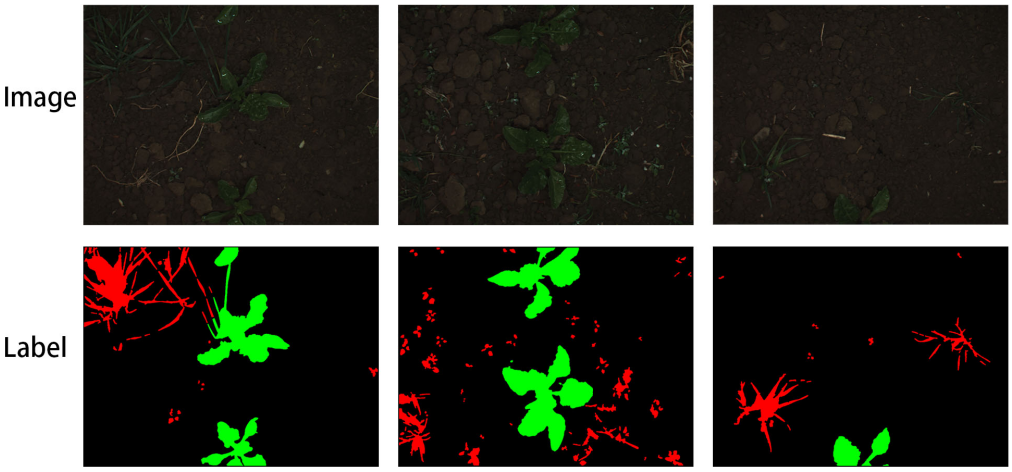

Figure 2. The original images and the corresponding annotations (green: crop, red: weed) in the sugar beet dataset.

2.1.3. Carrot Dataset

The carrot dataset is sourced from the CWFID dataset [34]. The images in this dataset are collected at a commercial organic carrot farm in northern Germany. The images are captured during the early true leaf growth stage of carrot seedlings using a JAI AD130GE multispectral camera, which can capture both visible and near-infrared light. The images

have a resolution of 1296 × 966 pixels. During the acquisition process, the camera is positioned vertically above the ground at a height of approximately 450 mm, with a focal length of 15 mm. In order to mitigate the effects of uneven lighting, artificial illumination is used in the shaded area beneath the robot to maintain consistent illumination intensity across the images. The dataset consists of 60 images, and we randomly split 70% of the samples for training, 20% for validation, and 10% for testing. As shown in Figure 3, green annotations denote the carrot seedlings, red annotations represent the weeds, and black annotations represent the soil and background.

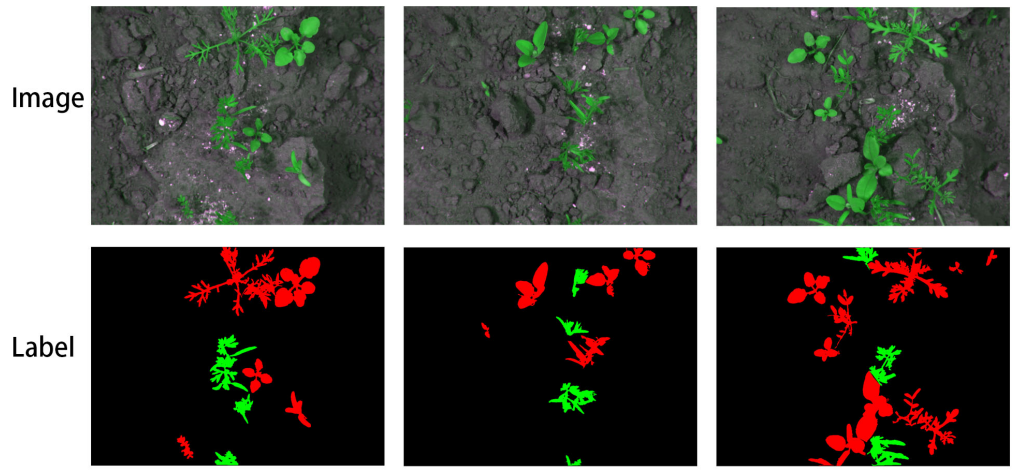

Figure 3. The original images and the corresponding annotations (green: crop, red: weed) in the carrot dataset.

2.1.4. Rice Dataset

The rice dataset is sourced from the rice seedling and weed dataset [35]. The images in this dataset have a resolution of 912 × 1024 pixels and captured using an IXUS 1000 HS camera with f-s 36–360 mm f/3.4–5.6 IS STM lens. The camera was 80–120 cm above the water surface of the fields during image capture. The dataset contains 224 images with corresponding annotations in 8-bit greyscale format. We convert the original annotations into 24-bit RGB format, and randomly split the dataset into training, validation, and test sets in a ratio of 7:2:1. As shown in Figure 4, green annotations represent the rice seedlings, red annotations represent the weeds, and black annotations represent the water or other backgrounds.

2.2. Segmentation Models

2.2.1. Model

The encoder–decoder architecture is commonly used in weed segmentation tasks. The encoder is responsible for rough classification of an image. It transfers the extracted semantic information to the decoder and maps the low-resolution features learned in the encoding stage to the high-resolution pixel space through skip connections by integrating the local and global context. The decoder gradually upsamples the feature maps to restore the resolution equalling the size of the input image, and outputs the predicted classifications for each pixel.

The attention mechanism is an adaptive selection process widely used in deep learning. It allows the network to focus on important regions of an image, thereby improving the performance and generalization ability of the model. In semantic segmentation, attention mechanisms can be categorized into channel attention and spatial attention [36]. Different types of attention serve different purposes. For instance, spatial attention focuses on important spatial regions [37–39], and channel attention aims to selectively attend to important channels or feature maps [40,41]. However, the existing weed semantic segmentation models often overlook the adaptability of channel dimension. Inspired by the visual attention

network [42], SegNext [43] re-examines the features considered by successful segmentation models and identified several key components for improving the performance of the model. SegNext proposes to use a large kernel attention mechanism to construct channel and spatial attention. The authors show that the convolutional attention is a more effective way to encode the contextual information as compared to self-attention mechanisms used in the Swin transformer [44].

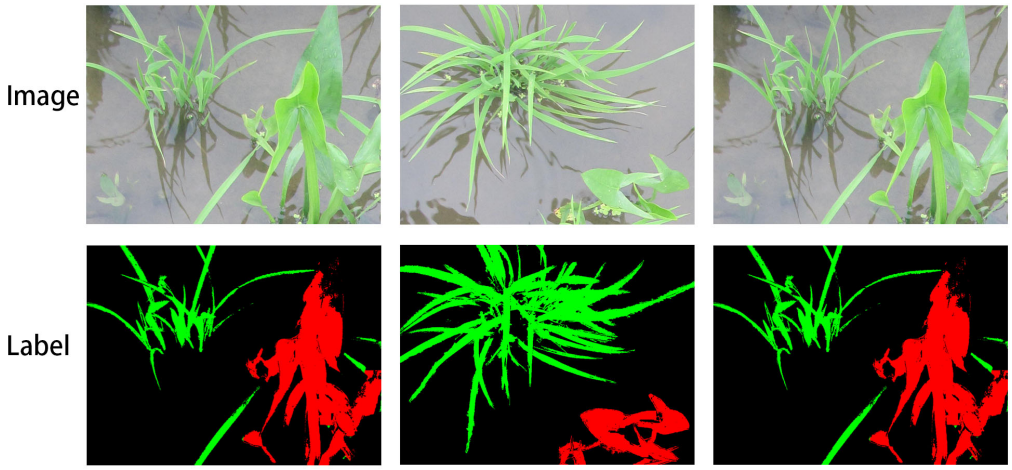

Figure 4. The original images and corresponding annotations (green: crop, red: weed) in the rice dataset.

Therefore, we use a convolutional attention mechanism to construct the proposed weed segmentation network. The convolutional attention mechanism consists of multiple sets of different convolutional kernels and deep convolutions, as shown in Figure 5. The 3 × 3 deep convolutional layer aggregates local information, while multiple sets of different depth-wise strip convolutions are used to capture multi-scale contextual information. The 1 × 1 convolutions establish the connections between different channels. In the proposed multi-scale convolutional attention (MCA), larger kernel sizes are used to capture global features. The MCA consists of three sets of strip convolutional kernels with different sizes. Each set is composed of two large convolutional kernels with relative sizes of 1 × 5 and 5 × 1, 1 × 11 and 11 × 1, and 1 × 17 and 17 × 1, combined in parallel to form multi-scale kernels. The proposed MCA can be mathematically expressed as follows:

$$F_{out} = F_{in} \otimes Conv_{1\times1}\left(\sum_{i=0}^{n} MSK_i(Conv_{3\times3}(F_{in}))\right) \quad (1)$$

where F_{in} represents the feature after passing through a 1 × 1 convolution and GELU activation. F_{out} is the output of the attention map. $\otimes$ denotes element-wise matrix multiplication. $Conv_{1\times1}$ represents a 1 × 1 convolution operation, and $Conv_{3\times3}$ represents a 3 × 3 convolution operation. $MSK_i, i \in \{0, 1, 2, 3\}$ denotes the i-th branch in Figure 5, where MSK_0 represents the direct connection used to preserve the residual information. In each branch, two depth-wise strip convolutions are used to approximate standard depth-wise convolutions with larger kernels, as the strip convolution is lightweight and serves as a complement to grid convolutions assisting in the extraction of strip-like features [45,46].

We used the convolutional attention mechanism MCA mentioned above to construct an MSFCABlock consisting of an MCA and an FFN network, as shown in Figure 6. MSFCABlock strengthens the feature association between the encoder and decoder. In the MSFCABlock, first, the contact feature is passed through a 3 × 3 convolution and batch normalization (BN). Then, it is connected with the output of the MCA module by using a residual connection. Subsequently, the feature is processed by the feed-forward network (FFN) with a residual connection. The FFN structure in the MSFCABlock maps the input

feature vectors to a high-dimensional space and then applies a non-linear transformation by using an activation function resulting in a new feature vector. This new feature vector contains more information compared to the original feature vector. The global contextual modelling multi-layer perceptron (MLP) and large convolution capture the global contextual features from long-range modelling, thus allowing the proposed MSFCABlock to effectively extract the features.

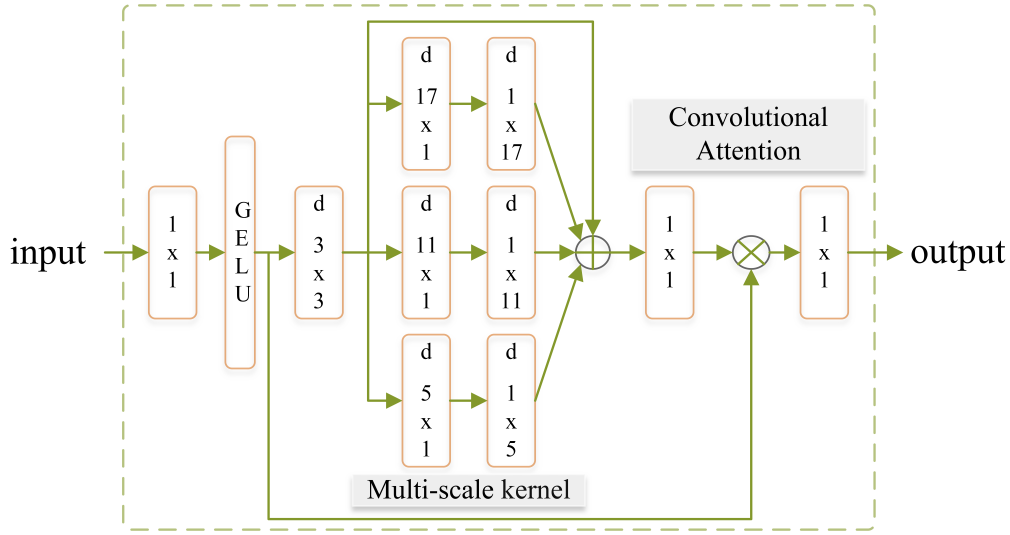

Figure 5. The architecture of the proposed MCA.

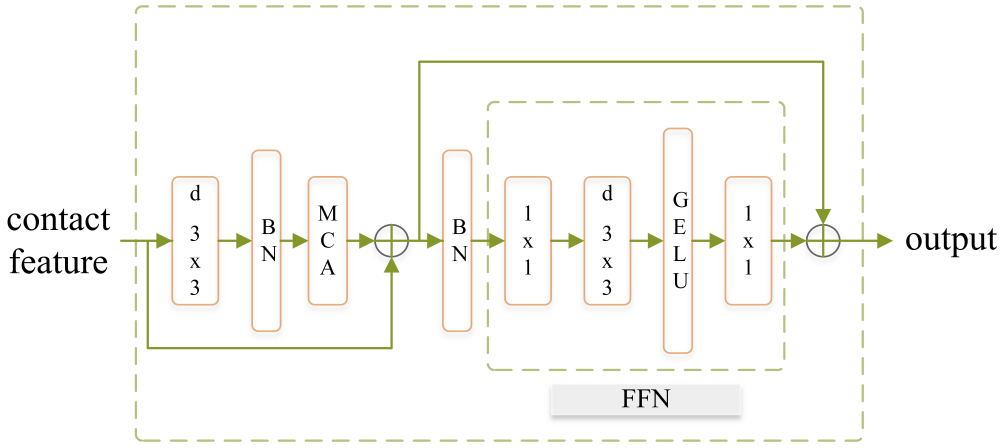

Figure 6. The architecture of the proposed multi-scale feature convolutional-attention block (MSFCA-Block).

The overall architecture of the proposed MSFCANet is shown in Figure 7. The proposed network consists of an encoder and a decoder. The encoder uses a VGG16 network as the backbone, where the blue blocks represent the convolutional layers. Since we use a VGG16-based encoder, the convolutional layers are fully convolutional. The pink blocks represent the max-pooling layers, and the green blocks represent the upsampling layers. We use the transpose convolution method for upsampling, which can learn different parameters for different tasks, thus making it more flexible compared to other methods, such as bilinear interpolation. The yellow blocks represent concatenation, and the brown blocks represent the proposed MSFCABlock. The proposed MSFCABlock combines features from different layers of the encoder during the decoding process, resulting in excellent and dense contextual information integration for weed segmentation, as well as richer scene understanding. It enhances the role of multi-scale feature aggregation in network

design. The large convolution with small parameter size also reduces the number of network parameters.

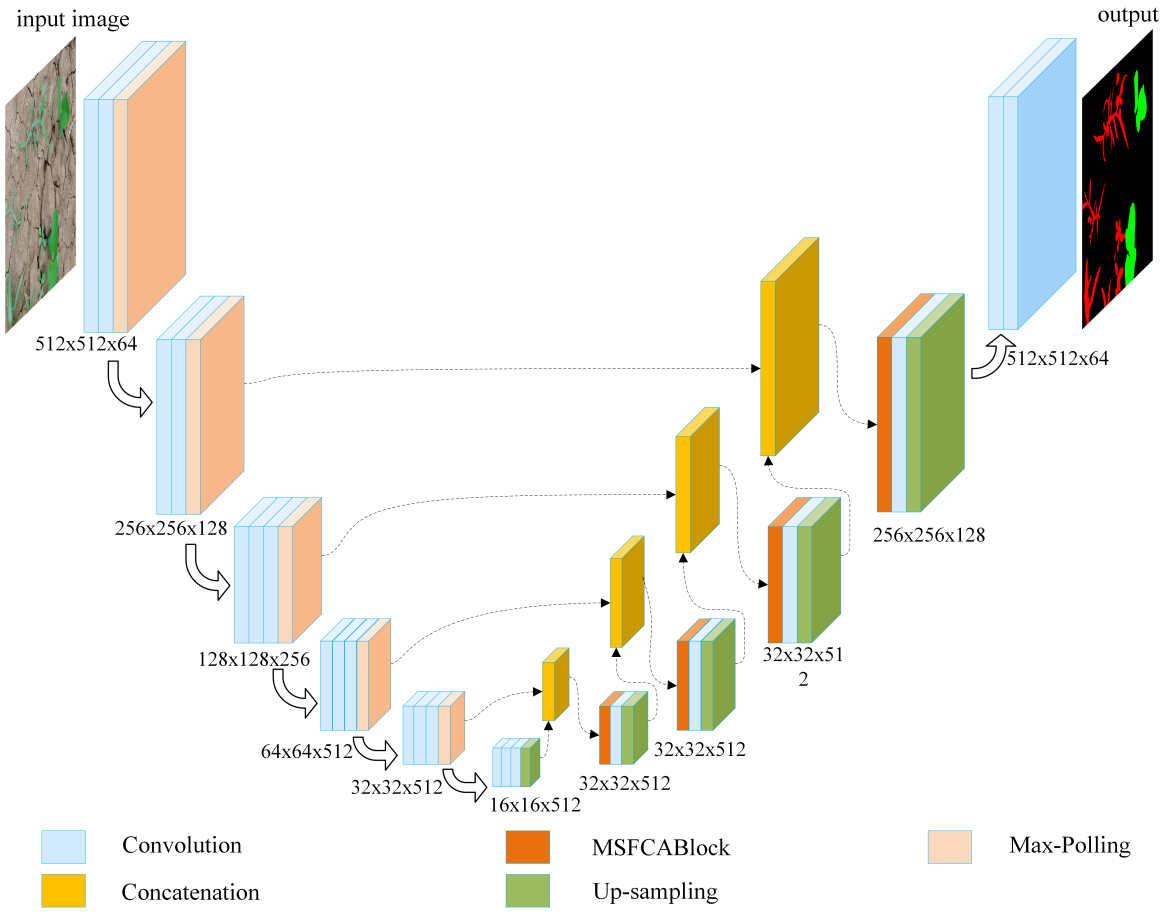

Figure 7. The architecture of the proposed MSFCA-Net.

2.2.2. Loss

In order to achieve more accurate segmentation of crops and weeds, we designed multiple losses for training the model, including plant loss, crop loss, weed loss, and crop–weed loss.

In this work, plant loss in the loss function considers crops and weeds as the same class to calculate the loss. This helps in balancing the crops and weeds in the proposed model. The cross-entropy loss, which is commonly used for semantic segmentation based on CNNs considers the high-frequency distribution of images as an important feature of the CNN. However, when the number of foreground (crops and weeds) pixels is much smaller than the number of background pixels, the background loss dominates, resulting in poor network performance. Therefore, in this experiment, cross-entropy loss is not used to calculate the plant loss. Instead, Dice loss [47] is employed, originating from the Dice coefficient, which is a measure of set similarity often used for comparing the similarity between two samples. Please note that the Dice loss is a region-based loss, i.e., the loss and gradient for a certain pixel not only depends on its label and predicted value, but also on the labels and predicted values of other pixels. The Dice loss can be used in cases of class imbalance. Considering the characteristics of Dice loss and the practical situation of this

work, plant loss in this research adopts Dice loss in order to effectively calculate the overall loss of crops and weeds. The plant loss is computed as follows:

$$Plantloss(y, \hat{y}) = 1 - \frac{2\sum_{i=0}^{B\times H\times W}(y(i)\hat{y}(i))}{\sum_{i=0}^{B\times H\times W}(y(i) + \hat{y}(i))} \quad (2)$$

where y and $\hat{y}$ represent the ground truth and predicted values of the pixels, respectively. B, H, and W denote the channel size, height, and width of the image, respectively. In this work, crop and weed loss are calculated separately for crops and weeds. Considering that the loss calculation for crops and weeds also lacks high-frequency components, the use of cross-entropy loss may result in inaccurate region detection. Therefore, Dice loss is used to calculate these losses.

For crop–weed loss, in order to efficiently optimize the severe class imbalance between the crop and weed categories, we use focal loss. The focal loss [48] addresses the issues of imbalanced training samples and different learning difficulties of samples. It is a variant of the cross-entropy loss, as shown in (3), with the addition of parameters α and β. These parameters are used to address the problems of difficult samples and imbalanced quantities.

$$CrossEntropyLoss = \sum_{i=0}^{B\times H\times W} -(p_i \ln q_i) \quad (3)$$

$$Cropweedloss = Focal(p, q) = \sum_{i=0}^{B\times H\times W} -\alpha(1 - q_i)^{\gamma}(p_i \ln q_i) \quad (4)$$

In (4), the role of α is to weight the loss of different classes of samples, where a higher weight is assigned to the class with fewer samples. On the other hand, the role of β is to handle the imbalance between easy and hard samples during the training process, where the number of easy samples is much larger than the number of hard samples. By adding a weight β, the loss of easy samples is significantly reduced, thus allowing the model to focus more on optimizing the loss of hard samples. Therefore, the total loss used for training is expressed as follows:

$$Totalloss = Plantloss + Croploss + Weedloss + Cropweedloss \quad (5)$$

where the total loss is a dimensionless metric, representing a measure of dissimilarity between the predicted and truth segmentation, with a value of 0 indicating perfect agreement and a value of 1 indicating complete dissimilarity.

2.2.3. Parameter Evaluation

In this work, we focus on semantic segmentation, which is a pixel-level prediction. Therefore, we adopt MIoU, Crop IoU, Weed IoU, Background IoU, F1-score, precision, and recall as the evaluation metrics. Please note that IoU is an important metric for measuring the accuracy of image segmentation. It is defined as the ratio of the intersection of the predicted and ground truth sets to their union and is mathematically expressed as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP + FN + FP}} \times 100\% \quad (6)$$

where TP (true positive) represents the intersection of the ground truth and predicted values and FN (false negative) + FP (false positive) represents the union of the ground truth and predicted values. MIoU is the average of Crop IoU, Weed IoU, and Background IoU, which is the intersection over union values for these three classes. It is calculated by taking the average of the IoU values for each class, which represents the ratio of intersection to union for each class.

The pixel precision refers to the ratio between the number of correctly classified pixels and the total number of pixels correctly predicted in the image. The average precision is

the mean precision calculated for each class. The pixel precision reflects the accuracy of positive predictions among the predicted positive samples, i.e., the accuracy of predictions for positive samples. It is calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{7}$$

The recall, also known as sensitivity or recall rate, reflects the probability of correctly identifying positive samples among the actual positive samples. It is calculated by using the following expression:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{8}$$

The F1-score is the harmonic mean of precision and recall. It is calculated by using the following mathematical expression:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{9}$$

3. Results

3.1. Model Training

In this work, we used an Intel Core i9-13600KF CPU (Intel, Santa Clara, CA, USA), 32 GB RAM, and an NVIDIA GeForce RTX 3090 GPU (NVIDIA, San Jose, CA, USA) to perform the experiments. The software environment included Windows 11 (Microsoft, Redmond, WA, USA), CUDA 11.3 (NVIDIA, CA, USA), Python 3.9 (Python Software Foundation, Fredericksburg, VA, USA), and TensorFlow 2.6 (Google Brain, Mountain View, CA, USA). The learning rate was set to 1×10^{-7} and Adam optimizer was used to update the weights. The batch size was set to 8 and the training was continued for 200 epochs. In order to increase the diversity of the training samples, data augmentation techniques, such as random flipping, cropping, and other operations were applied before training the model. The experiments were conducted on soybean, sugar beet, carrot, and rice weed datasets separately. For each dataset, MSFCA-Net was compared with FCN, FastFcn, OcrNet, UNet, Segformer, DeeplabV3, and DeeplabV3Plus based on experimental results. Finally, ablation experiments were conducted on the soybean dataset to validate the effectiveness of different components of the model. Six groups of ablation experiments were performed, including experiments with large convolutional kernels and hybrid losses.

3.2. Testing on the Soybean Dataset

In order to validate the performance of the proposed MSFCA-Net, experiments were conducted on the soybean weed dataset and the results were compared with other state-of-the-art methods, including FCN, FastFcn, OcrNet, UNet, Segformer, DeeplabV3, and DeeplabV3Plus. Table 1 shows the performance metrics, including MIoU, Crop IoU, Weed IoU, Bg IoU (Background IoU), recall, precision, and F1-score for the proposed MSFCA-Net and the aforementioned models based on the soybean weed test set. The quantitative analysis of results shows that the proposed MSFCA-Net performs efficiently on the soybean dataset and outperforms the other models. The proposed model achieves MIoU, Crop IoU, Weed IoU, Bg IoU, recall, precision, and F1-score of 92.64, 92.64, 95.34, 82.97, 99.62, 99.57, 99.54, and 99.55%, respectively, superior to the other models. In particular, the MIoU and Weed IoU of the proposed method are 2.6 and 6% higher compared to the second ranked OcrNet, respectively. This is due to the fact that the proposed MSFCA-Net utilizes skip connections to map the low-resolution features learned during the encoding stage to high-resolution pixel space semantically, exhibiting high performance in the presence of sample imbalance and hard-to-learn classes. Therefore, the proposed model has a strong advantage over current popular models in terms of dealing with sample imbalance and learning ability on hard-to-learn samples.

Table 1. A comparison of the proposed method with the other state-of-the-art methods based on the soybean dataset.

Model	MIoU (%)	Crop IoU (%)	Weed IoU (%)	Bg IoU * (%)	Recall (%)	Precision (%)	F1-Score (%)
FCN	86.12	90.68	68.62	99.05	92.27	91.75	92.01
FastFcn	88.12	92.79	72.35	99.22	93.61	92.94	93.28
OcrNet	89.90	93.23	77.09	99.37	94.28	94.55	94.42
UNet	87.34	9012	72.54	99.67	92.94	92.89	92.91
Segformer	86.56	89.05	71.31	99.22	92.75	92.01	92.37
DeeplabV3	88.25	92.92	72.59	99.23	93.68	93.04	93.35
DeeplabV3Plus	89.66	92.96	76.67	99.36	94.56	93.99	94.27
MSFCA-Net	92.64	95.34	82.97	99.62	99.57	99.54	99.55

* Bg IoU is the background IoU.

Figure 8 shows partial segmentation results of our method, MSFCA-Net, and other methods on the test dataset, where green represents soybean, red represents weeds, black represents background, and the labels denote manually annotated images. Analysis of the prediction results of the eight network models shows that MSFCA-Net produces more refined segmentation results and exhibits excellent noise resistance capabilities. This is because MSFCA-Net integrates multi-scale features using the multi-scale convolutional attention mechanism, effectively incorporating local information and global contextual information. The OcrNet, UNet, and Segformer tend to misclassify classes in the image and cannot accurately segment soybean seedlings and weeds. The FCN, FastFcn, DeeplabV3, and DeeplabV3Plus produced segmentation results that reflect the basic morphology of the predicted classes, but with blurred edges and lower accuracy. Our proposed method had the best segmentation results, with clear contours, complete details, smooth images, and segmentation results closest to manual annotation, indicating that the MSFCA-Net network model can effectively and accurately segment weeds, soybean, and background in the images.

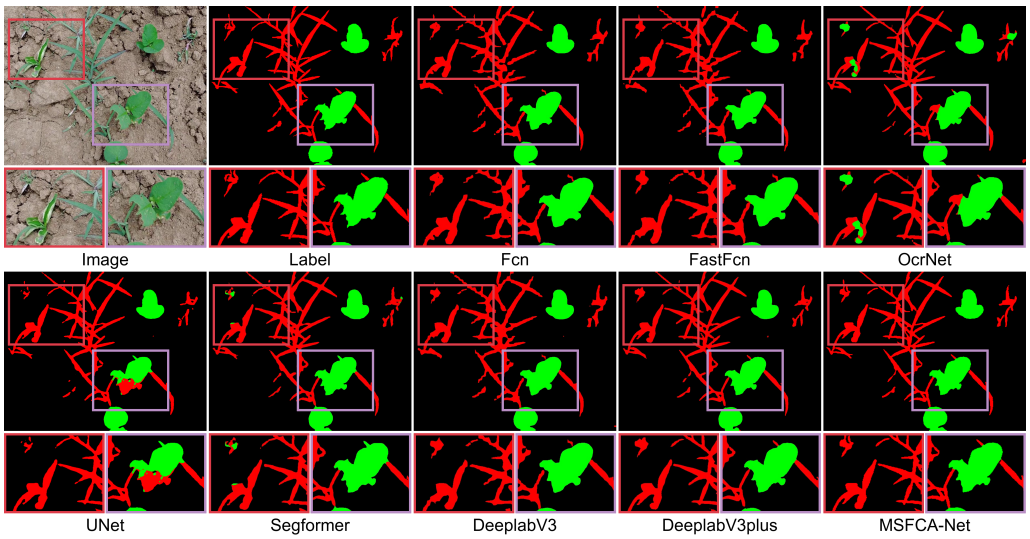

Figure 8. The segmentation results obtained using different methods based on the soybean dataset.

3.3. Testing on the Sugar Beet Dataset

Next, we conduct experiments on the sugar beet dataset, comprising a total of 2677 images, with 1874 images in the training set. As compared to other datasets used in this work, the sugar beet dataset is relatively large and used for training eight different models. The results obtained using the test set are shown in Table 2. The results show that the performance of all other models on the sugar beet dataset is significantly lower compared

to their performance on the soybean dataset. Although the sugar beet dataset has more training images, the background is more complex and the quality of data collection is relatively poor compared to the soybean dataset. As a result, training other networks becomes more challenging. This indicates that other models have higher requirements for the quality of training data and lack robustness in learning complex samples. On the other hand, the proposed MSFCA-Net still shows good performance in this challenging scenario. MSFCA-Net performs well in terms of various metrics as compared to the other models. The proposed model achieves MIoU, Crop IoU, and Weed IoU of 89.58, 9562, and 73.32%, respectively, ahead of the second ranked OcrNet by 3.5, 3.4, and 6.8%, respectively.

Table 2. A comparison of the proposed method with the other state-of-the-art methods based on the sugar beet dataset.

Model	MIoU (%)	Crop IoU (%)	Weed IoU (%)	Bg IoU * (%)	Recall (%)	Precision (%)	F1-Score (%)
FCN	81.42	90.66	54.40	99.19	86.69	90.49	88.39
FastFcn	81.38	99.12	54.68	81.38	86.32	91.14	88.40
OcrNet	86.01	92.20	66.49	99.35	90.59	93.19	91.83
UNet	82.48	90.29	55.78	99.31	87.13	90.82	88.94
Segformer	75.01	86.09	39.73	99.21	80.98	86.38	83.01
DeeplabV3	81.45	89.81	55.46	99.01	87.94	89.16	88.51
DeeplabV3Plus	84.72	91.20	63.68	99.29	90.00	91.99	90.95
MSFCA-Net	89.58	95.62	73.32	99.79	99.69	99.69	99.69

* Bg IoU is the background IoU.

Figure 9 shows the partial segmentation results of various networks on the test set, where red represents the weeds, green represents the soybean plants, and black represents the background. The "Image" refers to the original sugar beet image, and "Label" refers to the original annotated image. As presented in Figure 9, although other networks are able to recognize the categories, they show an inferior performance in terms of details and edge contours as compared to the proposed MSFCA-Net. By comparing the pink boxes in these images, it can be observed that other networks exhibit segmentation errors to varying extents, which is attributed to their poor performance in handling complex backgrounds. On the other hand, the segmentation results of the proposed MSFCA-Net are better, with more accurate classification of sugar beet, weeds, and the background.

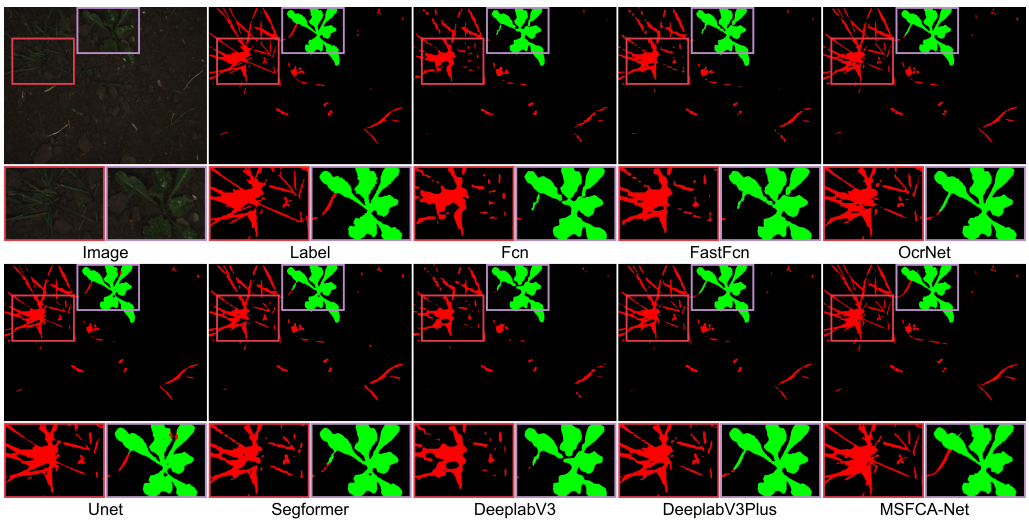

Figure 9. The segmentation results obtained using different methods based on the sugar beet dataset.

3.4. Testing on the Carrot Dataset

In the carrot dataset, there are 60 images. Based on the 70% random split, only 42 images were used to train the network. The prediction results of the eight different networks based on the test set are shown in Table 3. The results show that with few samples and high prediction density per pixel, training the model with limited samples is prone to overfitting and poor segmentation performance. Furthermore, the results show that the performance of other models is relatively low, indicating that the existing models are not effective in crop and weed segmentation for small datasets with severe sample scarcity. However, the proposed model performs significantly better compared to the existing models. The proposed model obtains MIoU, Crop IoU, and Weed IoU of 79.34, 59.84, and 79.57%, respectively, higher compared to the second ranked OcrNet by 4.2, 1.1, and 10.4%, respectively. This proves that the proposed model has a strong learning ability on small sample datasets.

Table 3. A comparison of the proposed method with other state-of-the-art methods based on the carrot dataset.

Model	MIoU (%)	Crop IoU (%)	Weed IoU (%)	Bg IoU * (%)	Recall (%)	Precision (%)	F1-Score (%)
FCN	70.40	50.83	63.49	96.89	83.85	78.81	81.16
FastFcn	70.59	50.52	64.40	96.86	82.84	80.39	81.29
OcrNet	75.19	58.73	69.12	97.64	85.38	85.42	85.38
UNet	73.89	53.46	70.02	98.18	83.84	82.64	85.08
Segformer	64.05	33.70	60.85	97.60	75.00	76.71	74.95
DeeplabV3	71.22	53.31	63.45	96.91	84.28	79.90	81.87
DeeplabV3Plus	74.79	56.29	70.44	97.63	87.10	82.24	84.50
MSFCA-Net	79.34	59.84	79.57	98.62	98.25	98.56	98.41

* Bg IoU is the background IoU.

Figure 10 shows the partial segmentation results of various networks obtained using the test set, where green represents the carrot seedlings, red represents the weeds, and black represents the background. 'Image' is the original image and 'Label' is the corresponding manually annotated image. The results show that FCN, FastFcn, and DeeplabV3 not only have inaccurate classification results on the test set, but also have blurry segmentation of weed and carrot crop contours. Although OcrNet, UNet, Segformer, and DeeplabV3Plus show some improvements in the contours of carrot seedlings and weeds compared to FCN, FastFcn, and DeeplabV3, they still have significant errors in class-wise segmentation prediction. This is because these network lack the ability to learn from small sample datasets. In contrast, the proposed network's segmentation results on the test set are almost identical to the original annotated images. This further demonstrates the strong segmentation capability of the proposed MSFCA-Net on complex and intertwined crop and weed small datasets.

3.5. Testing on the Rice Dataset

The rice dataset contains 224 images. Based on the 70% split, 157 images were used to train the model. Figure 4 shows that rice seedlings have numerous and dense leaves, and the weeds in the water often overlap with each other, thus making it difficult for the segmentation network to distinguish between rice and weeds. Moreover, the rice seedlings grow in water, and water produces reflections, thus increasing the difficulty of weed segmentation. Such data demands high feature extraction capabilities from the segmentation network due to the relatively coarse annotation provided by the official dataset.

The results presented in Table 4 show that the performances of many networks are significantly impacted. The proposed model's performance in terms of Weed IoU is only 68.70%, while FastFcn, DeeplabV3, and DeeplabV3Plus achieve 69.54, 69.96, and 70.19%, respectively. These models outperform the proposed model in terms of Weed IoU because

the proposed MSFCA-Net enhances the learning of difficult samples in the presence of class imbalance, resulting in a more balanced learning effect. Therefore, although the proposed model's performance in terms of Weed IoU is not as good as these models, it performs significantly better in terms of Crop IoU and Bg IoU, with an MIoU of 78.12%, which is 3.2% higher than the second best model.

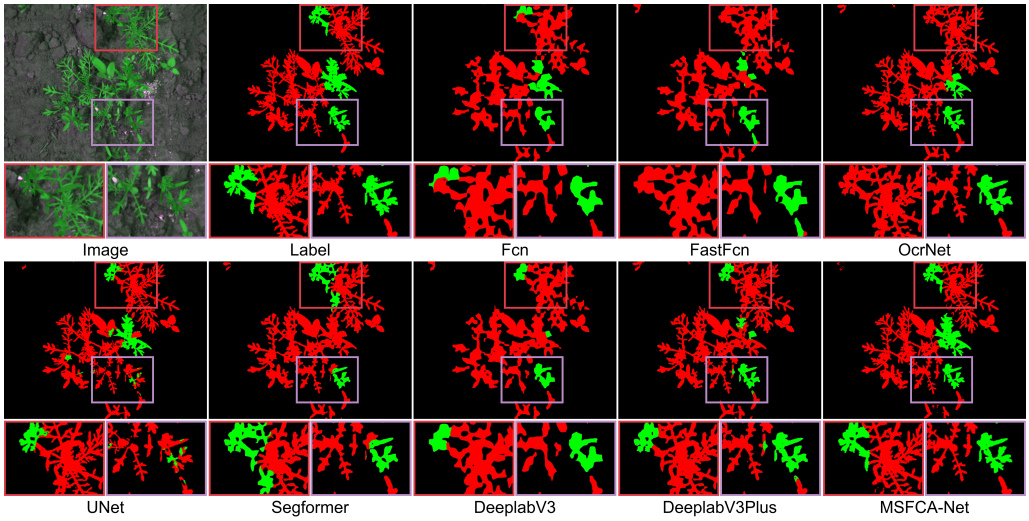

Figure 10. The segmentation results obtained using different methods based on the carrot dataset.

Table 4. A comparison of the proposed method with other state-of-the-art methods based on the rice dataset.

Model	MIoU (%)	Crop IoU (%)	Weed IoU (%)	Bg IoU * (%)	Recall/(%)	Precision (%)	F1-Score (%)
FCN	72.86	58.19	68.62	91.78	83.81	83.64	83.56
FastFcn	74.23	60.73	69.54	92.40	84.59	84.84	84.55
OcrNet	74.16	61.16	68.63	92.69	84.34	84.92	84.50
UNet	74.14	63.64	65.56	97.79	84.56	84.65	84.61
Segformer	72.51	58.59	66.92	92.04	82.93	83.93	83.31
DeeplabV3	74.07	60.01	69.96	92.24	84.83	84.30	84.43
DeeplabV3Plus	74.87	61.89	70.19	92.53	86.01	84.29	85.02
MSFCA-Net	78.12	67.56	68.70	98.12	96.41	95.47	95.93

* Bg IoU is the background IoU.

Figure 11 shows that FCN, FastFcn, and Segformer perform poorly as they fail to accurately predict the categories of rice and weeds, and many small weeds are not segmented. OcrNet, UNet, DeeplabV3, and DeeplabV3Plus are able to predict the categories of rice and weeds, but their contours are relatively rough. On the contrary, the proposed MSFCA-Net uses the multi-scale convolutional attention mechanism to effectively fuse multi-scale features, resulting in more refined segmentation results on the rice test set and more accurate classification. Therefore, the proposed model demonstrates more balanced performance when facing the complex background and class imbalance of the rice dataset, confirming the advantages of combining the convolutional attention mechanism with the hybrid loss training mode in the proposed model.

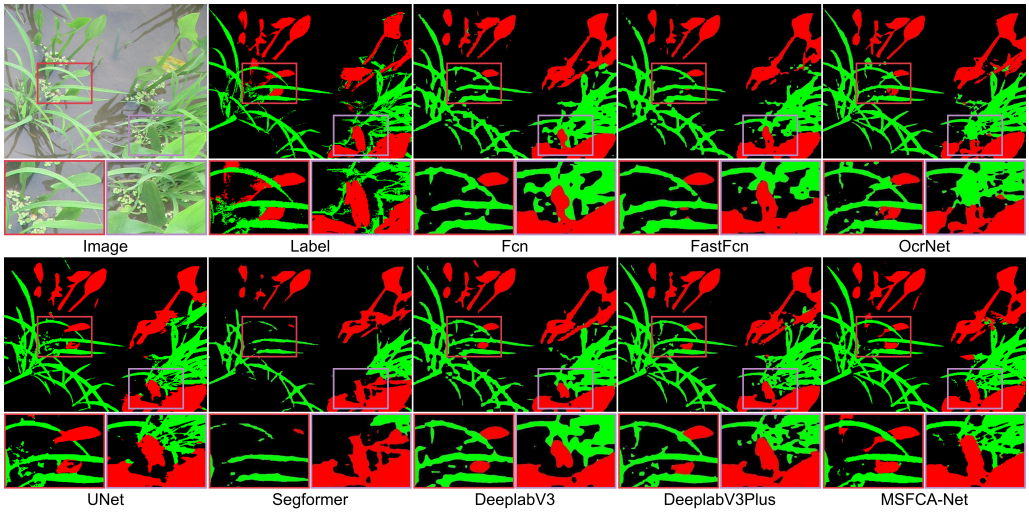

Figure 11. The segmentation results obtained using different methods based on the rice dataset.

3.6. Ablation Experiments

We conduct ablation experiments by using the soybean dataset to evaluate the contribution of different components of the proposed MSFCA-Net in the segmentation performance. We quantitatively and qualitatively compared the MSFCA-Net with existing image semantic segmentation methods. The results of the ablation experiments are shown in Table 5, where BaseNet refers to the encoder–decoder network structure based on VGG16, BABlock refers to the block using conventional 1 × 1 and 3 × 3 convolution kernels as an attention mechanism, which serves as a comparison with the MSFCABlock module in the MSFCA-Net. The hybrid loss refers to the hybrid loss proposed in this work. In total, six sets of comparative experiments were conducted: (1) BaseNet using the encoder–decoder structure based on VGG16; (2) adding the BABlock mechanism on top of the BaseNet model; (3) using the MSFCABlock module on top of the BaseNet model from experiment 1; (4) adding the hybrid loss to the BaseNet model; (5) using the BABlock module and adding the hybrid loss training mode on top of the BaseNet model; (6) using the multi-scale convolutional attention mechanism with different kernel sizes and the hybrid loss training mode on top of the BaseNet model.

Table 5. The results of ablation experiments obtained using the soybean dataset.

Model	MIoU (%)	Crop IoU (%)	Weed IoU (%)	Bg IoU * (%)	Recall (%)	Precision (%)	F1-Score (%)
BaseNet	88.33	90.89	74.65	99.44	99.28	99.33	99.30
BaseNet + BABlock	89.09	91.32	76.42	99.53	99.38	99.39	99.38
BaseNet + MSFCABlock	91.72	94.29	81.28	99.60	99.48	99.53	99.50
BaseNet + hybrid loss	90.35	93.02	78.49	99.52	99.41	99.40	99.41
BaseNet + BABlock + hybrid loss	91.33	93.62	80.79	99.58	99.50	99.47	99.48
MSFCA-Net	92.64	95.34	82.97	99.62	99.57	99.54	99.55

* Bg IoU is the background IoU.

Table 5 shows the BaseNet based on the VGG16 encoder–decoding structure as a benchmark. In the second experiment, adding the BABlock mechanism to the BaseNet model slightly improves the performance of the model. However, in the third experiment, when we use the MSFCABlock with complete multi-scale convolutional kernels, the performance of the model improves significantly, with MIoU, Crop IoU, and Weed IoU reaching 91.72, 94.29, and 81.28%, respectively. This represents an improvement of 1.37, 2.97, and 4.86%, respectively, compared to the model used in the second experiment, indicating that the proposed MSFCABlock has a strong capability in terms of multi-scale feature extraction. In the fourth experiment, we observe that the combination of Dice and focal losses in the hybrid loss training mode improves the performance of the model, showing a higher performance when dealing with class imbalance. In the fifth experiment, even with the addition of the hybrid loss training mode on top of the model used in the second experiment,

the performance improvement is still limited. This is because the BABlock has limited capability in extracting multi-scale features by using simple 3×3 and 1×1 convolutions, resulting in a lower segmentation accuracy. In the sixth experiment, we test the complete MSFCA-Net and the results showed a significant performance improvement, with MIoU, Crop IoU, and Weed IoU reaching 92.64, 95.34, and 82.97%, respectively. As compared to the BaseNet + BABlock + Hybrid Loss model in the fifth experiment, the improvements in MIoU, Crop IoU, and Weed IoU are 1.30, 1.72, and 2.18%, respectively. Compared to the BaseNet, the proposed MSFCA-Net showed even higher improvements of 4.31, 4.45, and 8.32% in terms of MIoU, Crop IoU, and Weed IoU, respectively. This is because the multi-scale convolutional kernels in MSFCA-Net focus more on multi-scale features, thus allowing better fusion of low- and high-level features, enhancing the model's ability in feature extraction.

The above ablation experiments demonstrate the effectiveness of the proposed multi-scale convolutional kernels with the convolutional attention mechanism and hybrid loss training mode for weed segmentation. It has been shown that the proposed MSFCA-Net performs well in segmenting crops, weeds, and background in agricultural images from four agricultural image datasets, with a strong performance and generalization ability, demonstrating its superiority.

4. Discussion

Currently, there are many semantic segmentation methods for crop weed segmentation based on the UNet model with simple modifications. Guo et al. [49] added a depth-wise separable convolution residual to a UNet, assigning different weights to each channel of the feature map obtained based on the convolutional operations, and using adaptive backpropagation to adjust the size of the one-dimensional convolutional kernel. This module slightly increases the number of parameters, but improves the network's feature extraction performance and enhances attention on the channels. However, the ability of this network to extract deep features for weed segmentation is insufficient and it lacks spatial attention. This method's segmentation performance is greatly influenced by imbalanced categories of crops and weeds and the generalization of the model is poor.

Brilhador et al. [50] proposed a modified UNet for crop and weed segmentation. Their training approach involved using annotated patches of images to effectively identify specific regions of crops and weeds, enabling detailed shape segmentation. The use of patch-level analysis can lead to data augmentation effects. However, it is crucial to consider that the presence of crops and weeds within the patches can vary based on their sizes. Therefore, if the dataset being used has a lower ratio of crops and weeds, training the model may pose challenges. In summary, the performance of this approach is notably influenced by the characteristics of the dataset. Zou et al. [30] simplified the neural network by removing some deep convolutional layers from the UNet to achieve a lightweight network. After fine-tuning, the performance based on their data exceeded that of the original UNet. This method reduces the computational complexity and extraction of multi-scale deep features by the network. However, when facing issues, such as complex backgrounds and overlapping crops and weeds, the network struggles to achieve a good segmentation performance, resulting in a significant decrease in the segmentation accuracy.

In order to address the issues of the existing weed segmentation methods based on semantic segmentation models, we have developed a field crop weed segmentation model using a multi-scale convolutional kernel attention mechanism based on multi-scale asymmetric convolutional kernel design. The proposed MSFCABlock enhances the network's attention in both channel and spatial dimensions by focusing on better contextual information fusion between the encoder and decoder, thus improving the multi-scale feature aggregation capability and achieving high performance in complex scenes. By comparing the results across different datasets, all models performed significantly better on the soybean and sugar beet tests compared to the carrot and rice tests. We analysed that the superior performance on our self-collected soybean dataset can be attributed to

more accurate labelling of the data and a rich variety of samples in the training set. In the sugar beet dataset, the larger quantity of data helped the network in feature extraction and learning during training. However, the carrot and rice datasets posed challenges due to their smaller size and higher complexity, which could significantly affect the model's learning capacity. Additionally, variations in data collection equipment and angles further contributed to the differences in results across different datasets. Overall, our model showed less susceptibility to these factors compared to other models. When comparing different models on the same dataset, the proposed model outperforms current popular models in almost all metrics on four different datasets, demonstrating its strong performance in handling complex scenes and imbalanced categories, as well as its strong generalization ability. From the results and segmentation graphs, it is evident that our MSFCA-Net achieved excellent performance compared to models such as FCN, Unet, and Deeplabv3 when dealing with small datasets, complex background masking, and class imbalance issues. Especially on the carrot dataset, the proposed model showed high performance even with only 42 training images, indicating its strong learning ability on small sample datasets.

Although there are many types of weeds in the field, they are usually grouped into a single category and cannot be segmented into specific types of weeds. For field crops, all weeds should be removed as targets. This work also focuses on the segmentation of crops, weeds, and background as three categories. However, with the development of smart agriculture, a single weed classification model may not meet the needs of intelligent weed segmentation system. Accurate identification and analysis of weed types are necessary for specific pesticide formulations based on statistical field information. Additionally, our model is suitable for precise image segmentation and requires a certain distance between the camera and the soil to ensure image clarity. Therefore, our weed segmentation method may not be well suited for applications in the field of UAVs. In future research, we will further deepen our study based on weed species segmentation and the application of weed segmentation on UAVs.

5. Conclusions

In this work, we proposed the MSFCA-Net, a multi-scale feature convolutional attention network for crop and weed segmentation. We used asymmetric large convolutional kernels to design an attention mechanism that aggregates multi-scale features, and employed skip connections to effectively integrate the local and global contextual information, thus significantly improving the segmentation accuracy of the proposed model and enhancing its ability to handle details and edge segmentation. We also designed a hybrid loss calculation mode combining Dice loss and focal loss. In addition, we designed separate loss functions for crops and weeds. This hybrid loss effectively improved the performance of the proposed model in handling class imbalance, enhancing its ability to learn from difficult samples. The experimental results show that our model demonstrated significantly better performance compared to other models on the soybean, sugar beet, carrot, and rice datasets, with mIoU scores of 92.64, 89.58, 79.34, and 78.12%, respectively. This confirms its strong generalization ability and ability to handle crop and weed segmentation in complex backgrounds. The ablation experiments on the network confirms the proposed model's ability to extract features using asymmetric large convolutional kernels and spatial attention. We also captured and manually annotated a dataset of soybean seedlings and weeds in a field, thus enriching the dataset of agricultural weed data and providing rich and effective data for future research. This work has important implications for the development of intelligent weed control and smart agriculture.

Our study still face challenges, such as variations in field lighting conditions, mutual occlusion between crops and weeds, and uneven sizes and quantities of crops and weeds. While research has addressed these challenges to some extent, real-world field conditions can introduce additional complexities that were not fully considered in this study. Therefore, our future research will focus on deploying weed segmentation networks

in real-world physical weed control robots and conducting relevant studies on targeted agricultural spraying.

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

Funding: This research was funded by the Major Science and Technology Projects in Anhui Province, grant number 202103b06020013; the Higher Education Quality Engineering Project of Anhui Province, grant number 2022sdxx012; Provincial Quality Engineering Project for Higher Education Institutions in Anhui Province: Virtual Simulation Experiment of Logistics Warehousing and Distribution, grant number 2021xnfzxm034; and National college logistics teaching reform teaching research project, grant number JZW2023426.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank all contributors to this study.

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

References

1. Zhu, J.; Wang, J.; DiTommaso, A.; Zhang, C.; Zheng, G.; Liang, W.; Islam, F.; Yang, C.; Chen, X.; Zhou, W. Weed research status, challenges, and opportunities in China. *Crop Prot.* **2020**, *134*, 104449. [CrossRef]
2. Tao, T.; Wei, X. A hybrid CNN-SVM classifier for weed recognition in winter rape field. *Plant Methods* **2022**, *18*, 29. [CrossRef] [PubMed]
3. Harker, K.N.; O'Donovan, J.T. Recent weed control, weed management, and integrated weed management. *Weed Technol.* **2013**, *27*, 1–11. [CrossRef]
4. Hamuda, E.; Glavin, M.; Jones, E. A survey of image processing techniques for plant extraction and segmentation in the field. *Comput. Electron. Agric.* **2016**, *125*, 184–199. [CrossRef]
5. Rodrigo, M.; Oturan, N.; Oturan, M.A. Electrochemically assisted remediation of pesticides in soils and water: A review. *Chem. Rev.* **2014**, *114*, 8720–8745. [CrossRef]
6. Gerhards, R.; Andujar Sanchez, D.; Hamouz, P.; Peteinatos, G.G.; Christensen, S.; Fernandez-Quintanilla, C. Advances in site-specific weed management in agriculture—A review. *Weed Res.* **2022**, *62*, 123–133. [CrossRef]
7. Chen, Z.; Zhang, C.; Li, N.; Sun, Z.; Li, W.; Zhang, B, Study review and analysis of high performance intra-row weeding robot. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 1–8.
8. Liu, C.; Lin, H.; Li, Y.; Gong, L.; Miao, Z. Analysis on status and development trend of intelligent control technology for agricultural equipment. *Nongye Jixie Xuebao/Trans. Chin. Soc. Agric. Mach.* **2020**, *51* .
9. Michaels, A.; Haug, S.; Albert, A. Vision-based high-speed manipulation for robotic ultra-precise weed control. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 5498–5505.
10. Quan, L.; Xiao, Y.; Wang, J. Study on pattern recognition method of intelligent weeding equipment. *J. Northeast Agric. Univ.* **2018**, *49*, 79–87.
11. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674. [CrossRef]
12. Ahmed, F.; Al-Mamun, H.A.; Bari, A.H.; Hossain, E.; Kwan, P. Classification of crops and weeds from digital images: A support vector machine approach. *Crop Prot.* **2012**, *40*, 98–104. [CrossRef]
13. Sabzi, S.; Abbaspour-Gilandeh, Y.; Arribas, J.I. An automatic visible-range video weed detection, segmentation and classification prototype in potato field. *Heliyon* **2020**, *6*, e03685. [CrossRef] [PubMed]
14. Parra, L.; Marin, J.; Yousfi, S.; Rincón, G.; Mauri, P.V.; Lloret, J. Edge detection for weed recognition in lawns. *Comput. Electron. Agric.* **2020**, *176*, 105684. [CrossRef]
15. Sánchez-Sastre, L.F.; Casterad, M.A.; Guillén, M.; Ruiz-Potosme, N.M.; Veiga, N.M.A.D.; Navas-Gracia, L.M.; Martín-Ramos, P. UAV Detection of Sinapis arvensis Infestation in Alfalfa Plots Using Simple Vegetation Indices from Conventional Digital Cameras. *AgriEngineering* **2020**, *2*, 206–212. [CrossRef]

16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
18. Yang, Q.; Shi, L.; Han, J.; Zha, Y.; Zhu, P. Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. *Field Crop Res.* **2019**, *235*, 142–153. [CrossRef]
19. Fuentes, A.; Yoon, S.; Kim, S.C.; Park, D.S. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors* **2017**, *17*, 2022. [CrossRef]
20. Hall, D.; McCool, C.; Dayoub, F.; Sunderhauf, N.; Upcroft, B. Evaluation of features for leaf classification in challenging conditions. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015; pp. 797–804.
21. Olsen, A.; Konovalov, D.A.; Philippa, B.; Ridd, P.; Wood, J.C.; Johns, J.; Banks, W.; Girgenti, B.; Kenny, O.; Whinney, J. DeepWeeds: A multiclass weed species image dataset for deep learning. *Sci. Rep.* **2019**, *9*, 2058. [CrossRef]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
23. 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; Proceedings, Part III 18; pp. 234–241.
24. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
25. Yuan, Y.; Chen, X.; Chen, X.; Wang, J. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv* **2019**, arXiv:1909.11065.
26. Lateef, F.; Ruichek, Y. Survey on semantic segmentation using deep learning techniques. *Neurocomputing* **2019**, *338*, 321–348. [CrossRef]
27. You, J.; Liu, W.; Lee, J. A DNN-based semantic segmentation for detecting weed and crop. *Comput. Electron. Agric.* **2020**, *178*, 105750. [CrossRef]
28. Yu, J.; Sharpe, S.M.; Schumann, A.W.; Boyd, N.S. Deep learning for image-based weed detection in turfgrass. *Eur. J. Agron.* **2019**, *104*, 78–84. [CrossRef]
29. Sun, J.; Tan, W.; Wu, X.; Shen, J.; Lu, B.; Dai, C. Real-time recognition of sugar beet and weeds in complex backgrounds using multi-channel depth-wise separable convolution model. *Trans. Chin. Soc. Aric. Eng* **2019**, *35*, 184–190.
30. 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]
31. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
32. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]
33. Chebrolu, N.; Lottes, P.; Schaefer, A.; Winterhalter, W.; Burgard, W.; Stachniss, C. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Robot. Res.* **2017**, *36*, 1045–1052. [CrossRef]
34. Haug, S.; Ostermann, J. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In Proceedings of the Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, 6–7 and 12 September 2014; Proceedings, Part IV 13; 2015; pp. 105–116.
35. Ma, X.; Deng, X.; Qi, L.; Jiang, Y.; Li, H.; Wang, Y.; Xing, X. Fully convolutional network for rice seedling and weed image segmentation at the seedling stage in paddy fields. *PLoS ONE* **2019**, *14*, e0215676. [CrossRef]
36. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
38. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
39. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.* **2014**, *27, 2204–2212* .
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
41. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
42. Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; Hu, S.-M. Visual attention network. *arXiv* **2022**, arXiv:2202.09741.
43. Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; Hu, S.-M. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv* **2022**, arXiv:2209.08575.

44. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
45. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters—Improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
46. Hou, Q.; Zhang, L.; Cheng, M.-M.; Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4003–4012.
47. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
48. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
49. Guo, H.; Wang, S.; Lu, Y. Real-time segmentation of weeds in cornfields based on depthwise separable convolution residual network. *Int. J. Comput. Sci. Eng.* **2020**, *23*, 307–318. [CrossRef]
50. Brilhador, A.; Gutoski, M.; Hattori, L.T.; de Souza Inácio, A.; Lazzaretti, A.E.; Lopes, H.S. Classification of weeds and crops at the pixel-level using convolutional neural networks and data augmentation. In Proceedings of the 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Guayaquil, Ecuador, 11–15 November 2019; pp. 1–6.