photonics

MDPI

Article

A Lightweight Semantic Segmentation Model of Wucai Seedlings Based on Attention Mechanism

Wen Li [1], Chao Liu [1], Minhui Chen [1], Dequan Zhu [1], Xia Chen [2] and Juan Liao [1,3,*]

[1] College of Engineering, Anhui Agricultural University, Hefei 230036, China; 20111744@stu.ahau.edu.cn (W.L.); 20111737@stu.ahau.edu.cn (C.L.); 20720712@stu.ahau.edu.cn (M.C.); zhudequan@ahau.edu.cn (D.Z.)
[2] College of Information & Computer, Anhui Agricultural University, Hefei 230036, China; xiachen@ahau.edu.cn
[3] Hefei Institute of Technology Innovation Engineering, Chinese Academy of Sciences, Hefei 230094, China
* Correspondence: liaojuan@ahau.edu.cn

Abstract: Accurate wucai seedling segmentation is of great significance for growth detection, seedling location, and phenotype detection. To segment wucai seedlings accurately in a natural environment, this paper presents a lightweight segmentation model of wucai seedlings, where U-Net is used as the backbone network. Specifically, to improve the feature extraction ability of the model for wucai seedlings of different sizes, a multi-branch convolution block based on inception structure is proposed and used to design the encoder. In addition, the expectation "maximizationexpectation" maximization attention module is added to enhance the attention of the model to the segmentation object. In addition, because of the problem that a large number of parameters easily increase the difficulty of network training and computational cost, the depth-wise separable convolution is applied to replace the original convolution in the decoding stage to lighten the model. The experimental results show that the precision, recall, MIOU, and F1-score of the proposed model on the self-built wucai seedling dataset are 0.992, 0.973, 0.961, and 0.982, respectively, and the average recognition time of single frame image is 0.0066 s. Compared with several state-of-the-art models, the proposed model achieves better segmentation performance and has the characteristics of smaller-parameter scale and higher real-time performance. Therefore, the proposed model can achieve good segmentation effect for wucai seedlings in natural environment, which can provide important basis for target spraying, growth recognition, and other applications.

Keywords: wucai seedling segmentation; multi-branch convolution block; attention mechanism; lightweight model

Citation: Li, W.; Liu, C.; Chen, M.; Zhu, D.; Chen, X.; Liao, J. A Lightweight Semantic Segmentation Model of Wucai Seedlings Based on Attention Mechanism. *Photonics* **2022**, *9*, 393. https://doi.org/10.3390/photonics9060393

Received: 29 April 2022
Accepted: 31 May 2022
Published: 2 June 2022

1. Introduction

Wucai (Brassica campestris L. ssp. Chinensis var. rosularis Tsen), as an important autumn and winter vegetable plant, belongs to a variant of nonheading Chinese cabbage in the Brassicaceae family [1]. This plant is widely cultured in most parts of China, especially in the Yangtze-Huaihe River Basin, and has become increasingly popular in other countries for its beautiful shape and significant levels of vitamins and minerals [2]. The rapid development of vegetable industry has met the needs of people's daily life, but in the process of wucai production, there are still some problems such as weeds that reduce the yield and diseases that decrease quality of vegetable. Therefore, it is increasingly demanded for targeted spraying [3], mechanical weeding [4], growth measurement [5], and a series of means to improve the wucai yield and safety quality of wucai. To enhance the feasibility and effectiveness of these technologies, it is necessary to segment wucai plant seedlings accurately.

In recent years, along with advancements in machine vision systems, several approaches using image processing for segmenting plants from the background precisely have been studied [6,7]. Among them, the use of color indices is a common method for

plant segmentation [8], where the RGB (Red Green Blue) color space is first converted into an alternative color space to highlight the green plant pixels, and then, a threshold is used to segment the plants. Liao et al. [9] constructed a color-index 2Cg-Cb-Cr in the YCrCb color space for image graying and applied an improved Otsu method to segment rice seedlings in the obtained gray image. Riehle et al. [10] presented a combination of color-index-based method for pre-segmentation and used the pre-segmentation results to calculate the threshold value for a final segmentation in HSV color space and CieLab color space. Color-index-based methods are sensitive to color variations introduced by the capture conditions, such as outdoor illumination and weed, although color is the most direct information to distinguish plants and the background of the images. Moreover, the color-index-based methods with manual threshold optimization cannot achieve the desired segmentation result.

In order to overcome the disadvantage of color-index-based methods, researchers have applied other cues, including shape and texture, to accurately separate plants from the background. Bakhshipour et al. [11] extracted four shape factors, namely area, perimeter, and major and minor axis length values of each plant from binary images, integrated these shape features to establish a pattern for each variety of the plants, and then detected crops and weeds based on their pattern and the pattern recognition methods, such as support vector machine and artificial neural networks. Zou et al. [12] proposed a segmentation algorithm of broccoli seedlings based on color-texture features where gray-level co-occurrence matrix (GLCM) was used to extract five texture features, including contrast, dissimilarity, homogeneity, energy, and correlation. Although plant segmentation by means of texture and shape analysis can obtain good segmentation results in a complex environment, they require additional computational costs, and those features such as texture and shape are not very stable and significant if there is occlusion or leaves overlapping. Furthermore, these previous studies on plant segmentation utilized handcrafted, low-level features such as color, shape, and texture to quantify the pixel character of plants, and their success depended on some theoretical knowledge of botany and crucial parameters. In practice, these methods were of limited value and may therefore not be suitable for real applications.

With the improvements in computer calculation speed, deep learning has developed rapidly in recent years, and the results of plants segmentation are better than those of the above-mentioned traditional methods. Gong et al. [13] used the U-Net codec network [14] as the backbone network and constructed a semantic segmentation model of early-stage in-bag rice root under strong noise, in which the ResNet module [15] and squeeze-and-excitation (SE) block [16] were embedded into the skip connection layer to make the gradient more convenient to spread layer by layer. The model can achieve better segmentation performance compared with the classical Otsu method and the U-Net. To study the plant leaf phenotype of overlapping poplar seedling leaves under heavy metal stress, Liu et al. [17] proposed an accurate automatic segmentation method, which combined mask R-CNN [18] with density-based spatial clustering of applications and noise (DBSCAN) clustering algorithm [19] based on depth information, and the proposed method can automatically detect leaves with high accuracy. Wu et al. [20] used an unmanned aerial vehicle equipped with red-green-blue camera to acquire field images of rice seedlings and constructed a recursive segmentation network based on a deep fully convolutional neural network to regress the density map and count rice seedlings. The application of deep learning technology to plant seedling images can undoubtedly achieve better results. However, because of different amounts of water, light, nutrients, etc., absorbed by different wucai seedling plants, the size and shape of plants is likely to be inconsistent, and the feature information of small-sized seedling plants is not easy to be learned by network compared with large size plants, especially edge information. In addition, for most semantic segmentation networks, the accurate segmentation requires a large amount of training data. Therefore, it is necessary to apply a network that can perform well on a small dataset for the accurate segmentation of wucai seedlings and not only to extract multi-scale features

of wucai seedlings but also to pay more attention to the segmentation target from a large amount of information.

To develop a segmentation method based on deep learning for segmenting wucai seedlings, this study chose the U-Net as the backbone because the U-Net has good performance in semantic segmentation tasks, and the number of training samples of U-Net is relatively small [21]. In view of the wucai seedlings with different sizes and shapes, the encoder of the U-Net is improved to expand the model width and enhance the ability of extracting features with different scales. Besides, the attention mechanism is added before the decoder to make the segmentation model pay more attention to wucai seedlings, and a lightweight decoder unit is built to reduce the computational cost and number of parameters. Thus, the results show that the proposed segmentation model could accurately segment wucai seedlings from images.

2. Materials and Methods

2.1. Image Collection

Images of wucai seedling used in this study were captured using a Canon EOS 850 camera in September, in the Agricultural Research Park of Anhui Agricultural University, which is located in Anhui province, China (30.57 N, 117.01 E). The camera was 500–900 mm above the field surface during image capture. To improve the adaptability of the segmentation model, the image-taking process was randomly arranged on the farmland of the experiment sites and under different lighting conditions.

A total of 436 images were captured. The acquisition format was RGB images. The resolution of original image is 3648 $\times$ 2736, which is much larger than input image size of the U-Net. The high resolution of the original image could increase the difficulty of network training and lead to excessive use of GPU memory and training failure. To avoid this situation, the acquired wucai seedling images were scaled to reduce the resolution to 512 $\times$ 512.

2.2. Image Augmentation and Annnotation

The ground truth (GT) images of semantic segmentation model are pixel-to-pixel labeled. To obtain the labeled samples, the LabelMe software is used to manually annotate the semantic labels of each pixel in wucai seedling images with two separate categories: wucai seedlings and background, as shown in Figure 1. The areas containing wucai seedling is marked as white, and the background area is marked as black. The obtained GT images were used to train the model with the training dataset and calculate the performance with the validation dataset and test dataset. The training dataset and validation dataset were randomly allocated in a ratio of 7:2 for RGB seedling images, and the remaining images were used as a test set to evaluate the segmentation performance and robustness of the proposed segmentation model. The data sample included GT images and RGB images.

The training of a deep neural network needs a large number of images. To increase the amount of image data to better improve the robustness of the model and obtain stronger generalization capabilities, the common methods of data augmentation, including brightness enhancement, flip, and rotation, were applied to the training dataset, validation dataset, and test dataset to generate additional new images. After data augmentation, it is expanded from the original 436 images to 2180. Figure 2 is the effect image after data augmentation. Figure 2b,c have different brightnesses from the original image, which are obtained by brightness enhancement and brightness reduction, respectively. Figure 2d,e shows new images using flipping vertically and horizontally, respectively.

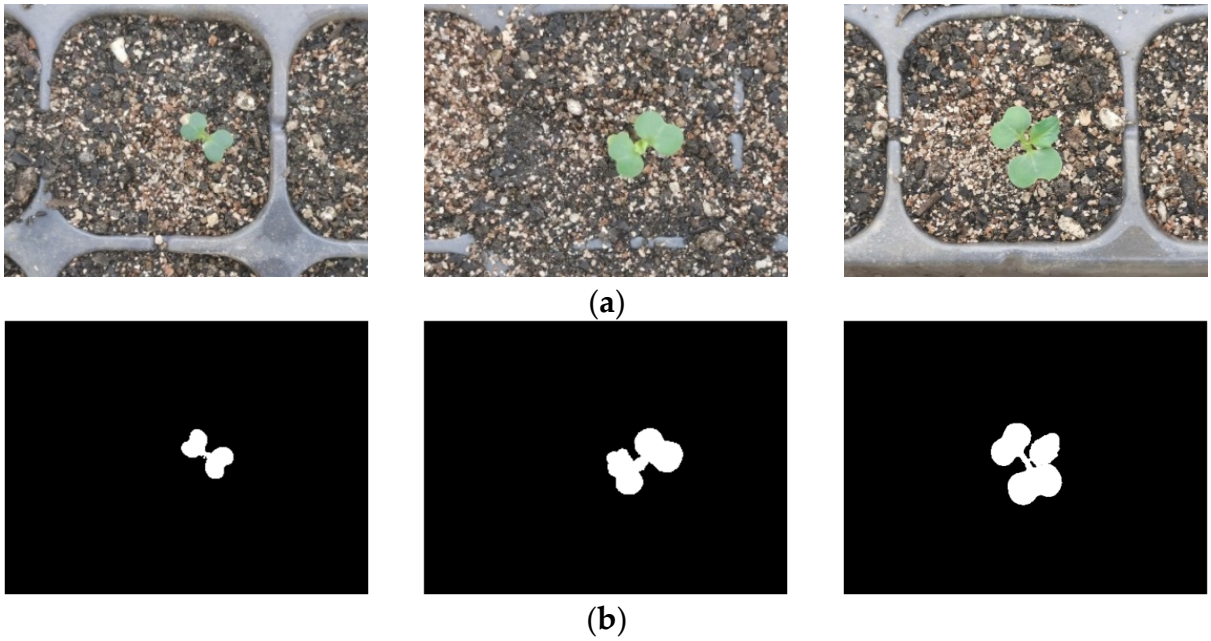

Figure 1. Original wucai seedling image and annotated image. (**a**) Original wucai seedling image samples with different scales; (**b**) annotated image.

Figure 2. Wucai seedling images after data augmentation. (**a**) Original image; (**b**) brightness enhancement; (**c**) brightness reduction; (**d**) horizontal flip; (**e**) vertical flip.

2.3. Wucai Seedling Segmentation Model Design

Convolutional neural networks (CNNs) have been widely developed in pixel-level images and many famous networks are based on CNNs. The U-Net used as the backbone of our segmentation model in this paper is a kind of CNN with a U-shaped structure based on encoder–decoder, which extracts features of the input image through convolution, pooling, and other operations and restores the image resolution by deconvolution operation. The U-Net has achieved good performance on semantic segmentation of different applications, and the number of train samples for U-Net is relatively small, but the U-Net model can only predict on a single scale [22], which cannot solve the problem of the size change of wucai seedlings, making the edge of image segmentation not smooth. On the other hand, previous research has shown that a small number of convolutional layers could extract simple image features, while using serial convolution block structures with different numbers of filters can obtain higher-level image features. [23,24]. Applying a large number of serial convolutional blocks to deepen the convolutional neural network easily makes model complexity increase and leads to data overfitting significantly although it can increase the dimension of feature information.

Based on this, to improve the performance and multi-scale adaptability of the network for wucai seedling segmentation, we embedded the multi-branch convolution module, expectation maximization attention (EMA) module [25], and depth-wise separable convolution block [26] into the backbone U-Net network. Figure 3 gives the architecture of

our segmentation model for the wucai seedlings. As shown in the figure, the multi-branch convolution module is built and introduced into the encoder to train a deep-wide network, and the EMA module is put between the encoder and decoder to make the network pay more attention to the wucai seedling region through the attention mechanism. In view of improving the optimization speed of the network, depth-wise separable convolution is used to replace the original convolution block in the decoding stage.

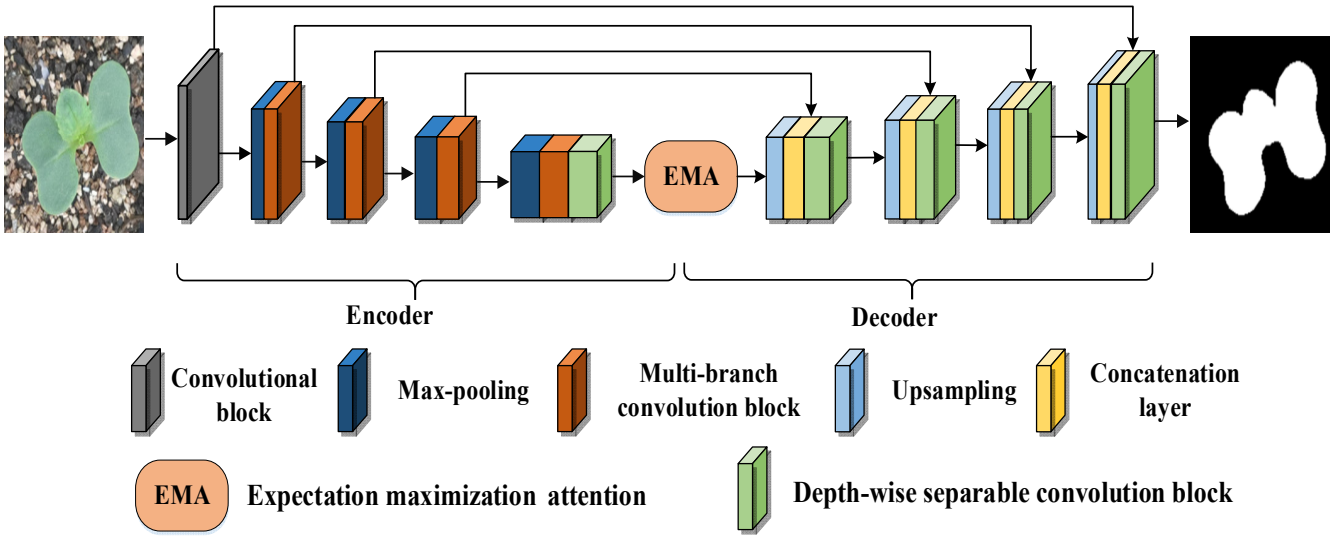

Figure 3. The architecture of the proposed segmentation model.

2.3.1. The Encoder Structure

In this study, the backbone of the network is the U-Net codec network, and the encoder is mainly used to extract image features, which consist of the repeated two convolution layers followed by a max-pooling layer for downsampling. To enhance the robustness of segmentation model to the size diversity of wucai seedlings, multiscale feature extraction is the key concern in the encoder design. It is well-known that the width and depth of convolutional neural networks are important indexes that affect the performance of convolutional neural networks, and multi-branch convolution block can widen the neural network and be beneficial to multiscale feature extraction. To increase the depth and width of the network, Szegedy et al. [27] proposed the inception module, which is a multi-branch learning structure and contains multiple convolutions with different kernel sizes. It learns the feature information of different scales through concatenating the outputs of different-sized filters, including 3×3, 5×5, and 7×7. Compared with the 3×3 convolutional layer, the convolution with a larger spatial filter (e.g., 5×5 or 7×7) can obtain the dependence between signals further away in the earlier layers, but it could result in the increase of the computational cost [28]; e.g., a 7×7 convolution with n filters is $49/9 = 5.44$ times more computationally expensive than a 3×3 convolution with the same number of filters.

On this basis, we build a multi-branch convolution module (named MBC) based on the inception structure, as shown in Figure 4, where two layers of 3×3 convolution layers are used to replace the 5×5 convolution in the original inception module, and the 7×7 convolution is replaced by three layers of 3×3 convolution layers. The multi-branch convolution module is applied to build the encoder part, which has the same receptive field as in the original inception module and fewer parameters. Table 1 gives detailed information about the encoder construction. Considering that a 1×1 convolution in MBC is used for dimension reduction, the first stage of the encoder is the same as that of the original U-Net encoder, and the convolution layers at other stages of the encoder are replaced by the MBC. In addition, we put the depth-wise separable convolution block after the MBC output of the encoder's fifth stage to enhance the feature extraction ability of the encoder.

Compared with the convolutional blocks of the original U-Net encoder, the improved encoder in this study can deepen and widen the network structure without increasing the computation parameters. In addition, the multi-branch convolution architecture implied by the MBC can extract and fuse the features of different receptive fields, achieving feature extraction of wucai seedlings of different sizes. Therefore, the network encoder constructed by multi-branch convolution modules can effectively improve the robustness of multiscale feature extraction and the segmentation performance.

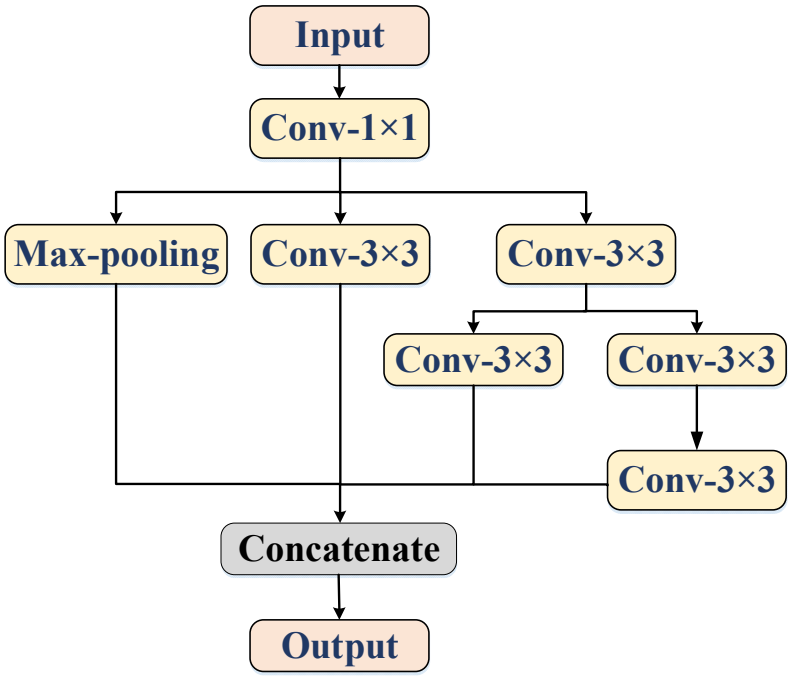

Figure 4. The structure of multi-branch convolution module.

Table 1. Detailed information of the encoder.

Stage	Layer	Filter Size	Stride	Output Size
Stage 1	Conv + BN + ReLu	3 × 3	1	512 × 512 × 32
	Conv + BN + ReLu	3 × 3	1	512 × 512 × 32
Stage 2	Maxpooling	2 × 2	2	256 × 256 × 32
	MBC	——	1	256 × 256 × 64
Stage 3	Maxpooling	2 × 2	2	128 × 128 × 64
	MBC	——	1	128 × 128 × 128
Stage 4	Maxpooling	2 × 2	2	64 × 64 × 128
	MBC	——	1	64 × 64 × 256
Stage 5	Maxpooling	2 × 2	2	32 × 32 × 256
	MBC	——	1	32 × 32 × 512
	Depth-wise_Sep_Conv	3 × 3 / 1 × 1	1	32 × 32 × 512
	Depth-wise_Sep_Conv	3 × 3 / 1 × 1	1	32 × 32 × 512

2.3.2. Expectation Maximization Attention Module

Attention mechanism, widely used for segmentation tasks, is essential to make the segmentation model pay different attention to different parts of input images and select feature information that is more critical to the current task, thus significantly improving the performance of the network. To make the segmentation model pay more attention to wucai seedlings, we add the EMA module after the last depth-wise separable convolution and before the input of the decoder, which is beneficial to the extraction of wucai seedling features.

The structure of the EMA module is shown in Figure 5, which consists of three operations, including responsibility estimation, likelihood maximization, and data re-estimation. The EMA computes attention maps from the perspective of expectation maximization, and iteratively estimates a much more compact set of bases upon the attention maps [29]. The EMA input is a feature map denoted as $\mathbf{X} \in \mathbb{R}^{N \times C}$, where N is the product of the length and width of the feature map $\mathbf{X}$, and C is the number of channels. Given the initial bases $\boldsymbol{\mu} \in \mathbb{R}^{K \times C}$, the responsibility estimation operation is used to estimate the responsibility $\mathbf{Z} \in \mathbb{R}^{N \times K}$, which is applied to update the base $\boldsymbol{\mu}$ through the likelihood maximization operation. The converged $\mathbf{Z}$ and $\boldsymbol{\mu}$ are obtained by executing the responsibility estimation and likelihood maximization operations alternately for a pre-specified number of iterations. Based on the EM algorithm, the estimated responsibility of the t-th iteration is defined as:

$$Z^{(t)} = softmax(\lambda \mathbf{X} (\boldsymbol{\mu}^{t-1})^T) \tag{1}$$

where λ is a hyper-parameter to control the distribution of $\mathbf{Z}$.

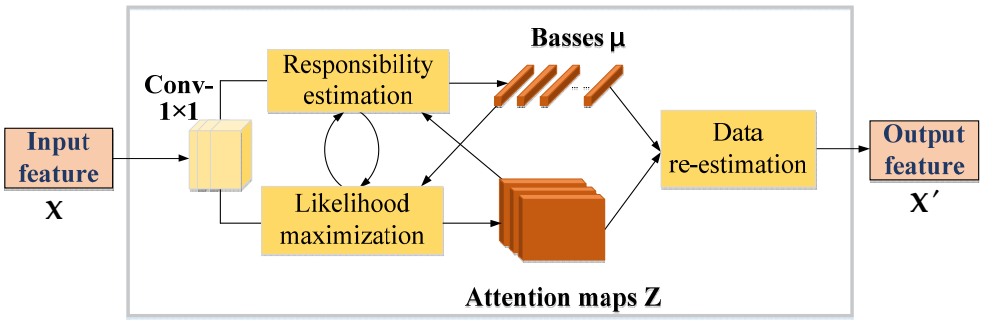

Figure 5. EMA structure diagram.

Let the responsibility of the k-th basis μ to x_n be z_{nk}, where $1 < k < K$ and $1 < n < N$. Hence, in the t-th iteration, μ_k^t is calculated as:

$$\mu_k^{(t)} = \frac{z_{nk}^{(t)} x_n}{\sum_{m=1}^{N} z_{mk}^{(t)}} \tag{2}$$

After obtaining the converged $\mathbf{Z}$ and $\boldsymbol{\mu}$, the data re-estimation reconstructs the original $\mathbf{X}$ as $\mathbf{X}'$ and outputs it, where the output feature map $\mathbf{X}'$ is formulated as:

$$\mathbf{X}' = Z^T \boldsymbol{\mu}^T \tag{3}$$

According to the above description, the feature map $\mathbf{X}'$ is the enhanced $\mathbf{X}$, and it has more detailed feature information in the feature space. Hence, the EMA module is applied before the decoder input, which is helpful for the encoder to obtain more detailed features, especially the edge characteristics of wucai seedlings of small size, thus improving the performance of the segmentation model.

2.3.3. Lightweight Decoder Structure

The U-Net consists of an encoder that extracts features and a decoder that gradually recovers the spatial information and generates pixel-level probability distributions. To obtain more feature information, the concatenation operation is applied in both the encoder and decoder to fuse high-level and low-level image features. Nevertheless, previous work shows that the layer-by-layer upsampling can increase the complexity of model and generate a large number of parameters. Therefore, in order to reduce the computation cost and number of parameters while maintaining similar performance, we replace traditional convolutions of the decoder with the depth-wise separable convolutions.

Figure 6 shows a comparison between a standard convolution and the depth-wise separable convolution. Different from traditional convolution, the depth-wise separable convolution factorizes a standard convolution into a depth-wise convolution and a point-wise convolution, and batch normalization (BN) and a rectified linear unit (Relu) are employed after each depth-wise and point-wise convolution. The basic idea behind depth-wise separable convolution is that a spatial convolution independently for each input channel through the depth-wise convolution, while the point-wise convolution is used to combine the output from the depth-wise convolution. Based on the structures of standard convolution and the depth-wise separable convolution, comparative calculation of the number of parameters is defined as:

$$\frac{K_1 K_2 C_I + C_I C_o}{K_1 K_2 C_I C_o} = \frac{1}{C_o} + \frac{1}{K_1 K_2} \tag{4}$$

where K_1 and K_2 are the width and height of convolution kernels, respectively; C_I and C_o are the channel numbers of the input and output feature maps, respectively. As shown in Equation (4), it is noted that the depth-wise separable convolution can significantly reduce the computation complexity.

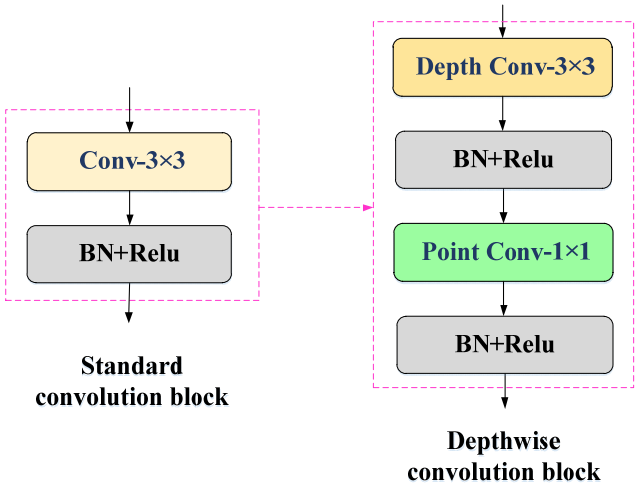

Figure 6. The details of a standard convolution and depth-wise separable convolutional block.

2.3.4. Loss Function

The design of loss function can impact on the performance of the network. For the training loss of the proposed segmentation model, we design a composite loss function formulated as follows:

$$L_{seg} = L_{miou} + L_{dice} \tag{5}$$

where L_{miou} is the MIOU loss function, and L_{dice} is the Dice loss function based on Dice coefficient. The MIOU loss function, constructed based on the principle of mean intersection

ratios, is used to compare the similarity between the ground truth and the predicted results. The calculation is as follows:

$$L_{miou} = 1 - \sum_{m \in N} \frac{y_m g_m}{y_m + g_m - y_m g_m} \tag{6}$$

where N is the set of all the pixels of the image, and m is the pixel index. $y_m$is the predicted value of pixel "m", and g_m is the ground truth value of pixel "m".

The Dice loss function L_{dice} is adopt to balance the imbalance between the wucai seedling area and the background in the image, which is calculated based on Dice coefficient as follows:

$$L_{dice} = -\sum_{m \in N} [g_m \log(y_m) + (1 - g_m) \log(1 - y_m)] \tag{7}$$

3. Results

3.1. Experimental Setup and Evaluation Metrics

The server environment for network training uses Windows 10 and python 3.6 to train and test the model under TensorFlow 2.0.0. The server is equipped with NVIDIA Quadro P2000 graphics cards with 5 GB of video memory for acceleration.

In the experiment, the adaptive authorization mechanism was used for model optimization. The initial learning rate was set to 0.001, the batch size was set to 2, and the segmentation model was trained with 200 iterations. The wucai seedling dataset was used to train and test the model, and the wucai seedling area were segmented by the segmentation model.

To evaluate the segmentation results, we adopt the precision, recall, MIOU, and F1-score as evaluation indicators, which are defines as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{MIOU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{10}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{11}$$

where the TP, FP, and FN denote the true-positive, false-positive, and false-negative measurements.

3.2. Influence on MBC and EMA

We investigated the performance of different modules in the proposed model using the U-Net as the backbone network design. Table 2 shows the precision, recall, MIOU, and F1-score indicators obtained with different modules, including MBC and EMA, described in the manuscript. From the table, the combination of MBC and EMA modules achieves the best segmentation performance for wucai seedlings in terms of the recall and MIOU as well as F1-score although the precision is relatively lower than that of U-Net with MBC and U-Net with EMA. By contrast, the single U-Net performs the worst in terms of all the evaluation measures, i.e., lower than U-Net with MBC by approximately 2.32% and 3.15%, respectively, in terms of precision and MIOU. This may be because the MBC module can widen the network and learn discriminative features for multi-size segmentation targets. Besides, by adding the EMA module to the connecting part of the encoder and decoder of the U-Net, the edges and more detail features of the wucai seedlings can be better transmitted. Using the design of combining MBC and EMA yields much better performance than the U-Net with MBC in recall, MIOU, and F1-score indexes, indicating the power of the EMA attention module.

Table 2. Comparison of evaluated indicators of the combination of different modules.

	Precision	Recall	MIOU	F1
U-Net	0.970	0.951	0.922	0.960
U-Net + MBC	0.993	0.955	0.952	0.974
U-Net + EMA	0.995	0.954	0.954	0.969
Ours	0.992	0.973	0.961	0.982

3.3. Comparison of Different Segmentation Models

In order to verify the segmentation performance of the proposed segmentation model in this study, we compared the segmentation results of the proposed model with the other four segmentation models, namely U-Net, SegNet, PSPNet, and DeepLabV3. In the comparison experiment, the same dataset and loss function mentioned above were used for five segmentation models, and the evaluation indexes introduced in Section 3.1 were applied to examine the segmentation models. The segmentation performances of the different models are shown in Table 3.

Table 3. Segmentation performance of different segmentation models.

Model	Precision	Recall	MIOU	F1
U-Net	0.970	0.951	0.922	0.960
PSPNet	0.983	0.861	0.859	0.917
SegNet	0.993	0.954	0.951	0.973
DeepLabV3	0.996	0.938	0.942	0.965
Ours	0.992	0.973	0.961	0.982

From Table 3, it can be seen that the performance of the proposed model is significantly better than that of the other four models for the segmentation of wucai seedling areas. Comparatively, the precision, recall, MIOU, and F1-score of the proposed model are 0.992, 0.973, 0.961, and 0.982, which are higher than U-Net by approximately 2.2%, 2.2%, 3.9%, and 2.2%, respectively. Compared with the U-Net, the precision of PSPNet is slightly better, but the other three evaluation indicators are lower. For SegNet and DeepLabV3, they perform better than U-Net and PSPNet, but they are still not comparable to the proposed model, with only precision slightly higher than the proposed model. The biggest difference is recall value, where the proposed model is 3.5% higher than DeepLabV3 and 1.9% higher than SegNet. The experiment results indicate that the design of encoder and attention mechanism in the model can promote the model to focus on some regions of interest or significant regions and learn more details in the training process, thus being superior to the other four models in the evaluation for the task of wucai seedlings segmentation

In addition, several images were randomly selected from the test set for detailed comparison, and the visual comparisons of segmentation results using these five models are presented in Figure 7. All the images in the first and second rows are the original images to be detected and the annotated images, respectively, and the following are the final prediction results of each model. It is observed that the proposed model can accurately segment wucai seedlings of different sizes and shapes under different illumination. This is because the proposed model used the multi-branch convolution module as a convolution block in the encoder and can learn features of wucai seedlings of different sizes, and the EMA module allows the model to learn more detail features, thereby filtering out the complex interference information of background and obtaining a good segmentation performance in the wucai seedling segmentation tasks. However, compared with the proposed model, the U-Net, SegNet, PSPNet, and DeepLabV3 models are too coarse for edge segmentation of wucai seedings, with especially poor adaptability to different sizes and shapes as well as different illumination. As shown in Figure 7e, the U-Net, SegNet, PSPNet, and DeepLabV3 models all incorrectly predict many of wucai seedling pixels as

background; thus, it cannot meet segmentation requirements, while the proposed model is closer to the ground truth.

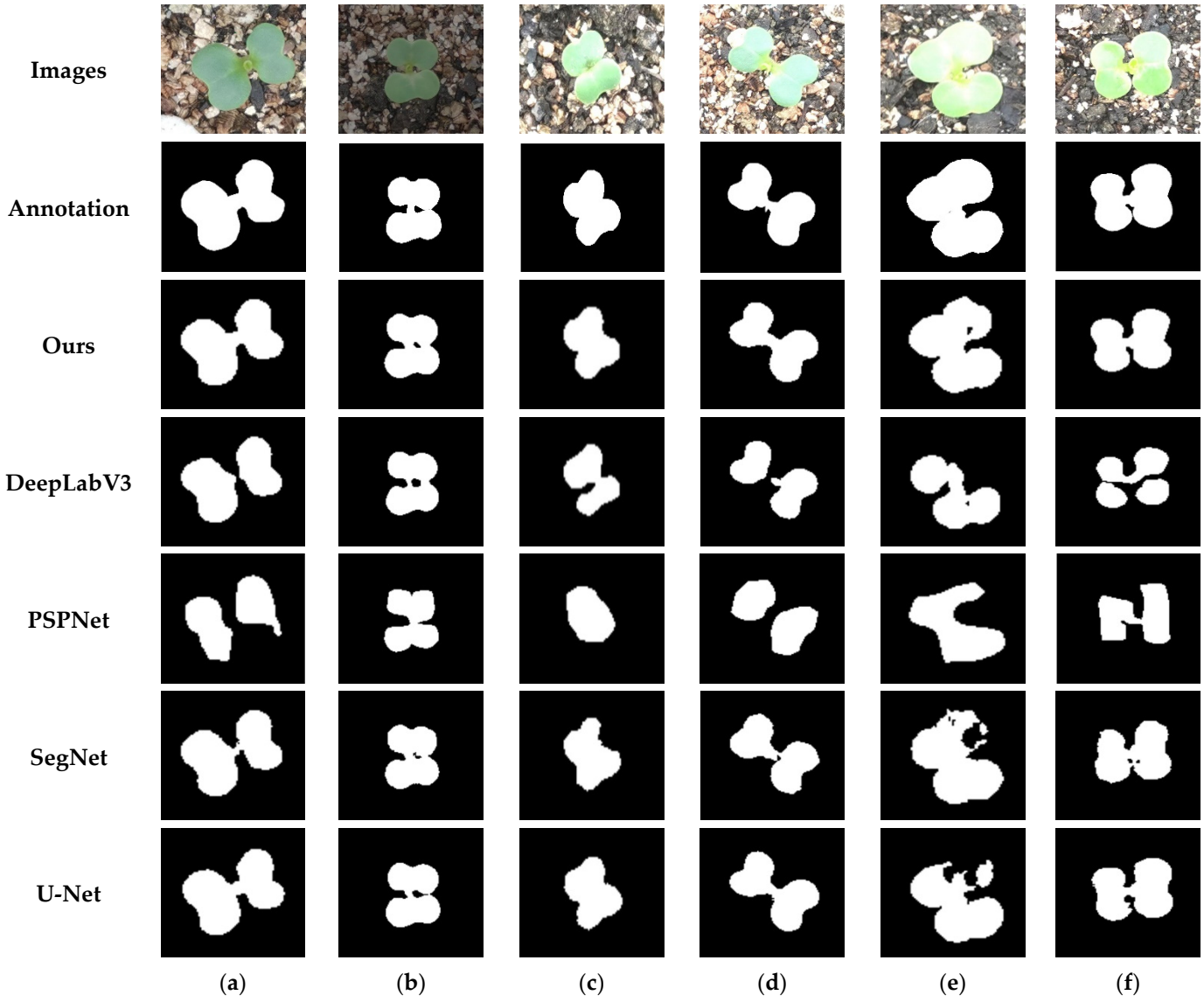

Figure 7. Segmentation results of different models; the images (**a**–**f**) are randomly selected from the test set.

Moreover, we also evaluated the complexity requirements of the segmentation models and the time required to successfully complete their execution, where the params, flpos, and macc are used as complexity evaluation indicators. Params indicates the number of parameters of the network. Flops refers to the computation of floating-point operation, and macc is multiply–accumulate operations, which are applied to measure the computational complexity of the network. The complexity requirements of the five segmentation models the proposed model and U-Net, SegNet, PSPNet, and DeepLabV3 are shown in Table 4. From the table, the proposed model has fewer parameters and the lowest computational complexity compared with the other four models. A total of 2.465394×10^6 network parameters are for the proposed model, while that of the original U-Net is 8.027783×10^6. Further, the average recognition time of a single-frame image of the proposed model is 0.0066 s, which is lower than that of other segmentation models. It is indicated that the proposed model has the characteristics of smaller-parameter scale and higher real-time

performance. Compared with other models, the proposed model is a lightweight model that is suitable for intelligent agricultural equipment with limited hardware system resources.

Table 4. Comparison of complexity requirements of different models.

	Params	Flops	Macc	Time (s)
U-Net	8,027,783	51.03	101.83	0.0196
PSPNet	46,706,626	59.11	810.69	0.0804
SegNet	29,444,162	160.68	321.0	0.0570
DeepLabV3	5,813,266	39.87	79.54	0.0158
Ours	2,465,394	14.6	29.08	0.0066

4. Conclusions

In this study, we have proposed a pixel-level segmentation model based on deep learning for wucai seedlings. Firstly, the proposed model is the backbone design of U-Net, and the encoder was equipped with a multi-branch convolutional block, thus improving the feature extraction ability of the model for wucai seedling of different sizes. The introduction of the expectation maximization attention module after the encoder can pay more attention to the segmentation target and reduce the extraction of background information, which can improve the precision of the model. Besides, a lightweight decoder based on the depth-wise separable convolution was designed to reduce the network parameters and increase the training speed. Experiments delivered on the wucai seedling image dataset have shown that the proposed model has a good segmentation outcome for wucai seedlings. The precision, recall, MIOU, as well as F1-score index were chosen to evaluate the model performance. The final test set scores were 0.992, 0.973, 0.961, and 0.982, respectively. Moreover, the average recognition time of a single-frame image of the proposed model is 0.0066 s. Furthermore, compared with U-Net, SegNet, PSPNet, and DeepLabV3, the proposed model is a lightweight model and has better segmentation results for wucai seedlings of different sizes and shapes, which provides object support toward the target spraying, growth recognition, and other applications and is suitable for intelligent agricultural equipment with limited hardware system resources.

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

Funding: This work was funded by the Natural Science Foundation of Anhui (No. 2108085MC96), the Key R&D Program of Anhui (No. 202004a06020016), the University Natural Science Research Project of Anhui (No. KJ2021A0180), the University Postgraduate Science Research Project of Anhui (No. YJS20210232), and the Natural Science Foundation of Anhui Agricultural University (No. K2148001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Zou, M.Q.; Yuan, L.Y.; Zhu, S.D.; Liu, S.; Ge, J.D.; Wang, C.G. Response of osmotic adjustment and ascorbate-glutathione cycle to heat stress in a heat-sensitive and a heat-tolerant genotype of wucai (*Brassica campestris* L.). *Sci. Hortic.* **2016**, *211*, 87–94. [CrossRef]
2. Chen, G.H.; Ye, X.Y.; Zhang, S.Y.; Zhu, S.D.; Yuan, L.Y.; Hou, J.F.; Wang, C.G. Comparative transcriptome analysis between fertile and CMS flower buds in Wucai (*Brassica campestris* L.). *BMC Genom.* **2018**, *19*, 908. [CrossRef] [PubMed]
3. Liu, L.; Mei, T.; Niu, R.X.; Wang, J.; Liu, Y.B.; Chu, S. RBF-based monocular vision navigation for small vehicles in narrow space below maize canopy. *Appl. Sci.* **2016**, *6*, 182. [CrossRef]

4. Li, Y.; Guo, Z.Q.; Shuang, F.; Zhang, M.; Li, X.H. Key technologies of machine vision for weeding robots: A review and benchmark. *Comput. Electron. Agric.* **2022**, *196*, 106880. [CrossRef]
5. González-Barbosa, J.J.; Ramírez-Pedraza, A.; Ornelas-Rodríguez, F.J.; Cordova-Esparza, D.M.; Gonzalea-Barbosa, E.A. Dynamic measurement of portos tomato seedling growth using the Kinect 2.0 sensor. *Agriculture* **2022**, *12*, 449. [CrossRef]
6. Hou, W.H.; Zhang, D.S.; Wei, Y.; Gao, J.; Zhang, X.L. Review on computer aided weld defect detection from radiography images. *Appl. Sci.* **2020**, *10*, 1878. [CrossRef]
7. Liao, J.; Wang, Y.; Zhu, D.Q.; Zou, Y.; Zhang, S.; Zhou, H.Y. Automatic segmentation of crop/background based on luminance partition correction and adaptive threshold. *IEEE Access* **2020**, *8*, 202611–202622. [CrossRef]
8. 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]
9. Liao, J.; Wang, Y.; Yin, J.N.; Liu, L.; Zhang, S.; Zhu, D.Q. Segmentation of rice seedlings using the YCrCb color space and an improved Otsu method. *Agronomy* **2018**, *8*, 269. [CrossRef]
10. Riehle, D.; Reiser, D.; Griepentrog, H.W. Robust index-based semantic plant/background segmentation for RGB-images. *Comput. Electron. Agric.* **2020**, *169*, 105201. [CrossRef]
11. Bakhshipour, A.; Jafari, A. Evaluation of support vector machine and artificial neural networks in weed detection using shape features. *Comput. Electron. Agric.* **2018**, *145*, 153–160.
12. Zou, K.L.; Ge, L.Z.; Zhang, C.L.; Yuan, T.; Li, W. Broccoli seedling segmentation based on support vector machine combined with color texture features. *IEEE Access* **2019**, *7*, 168565–168574.
13. Gong, L.; Du, X.F.; Zhu, K.; Lin, C.H.; Lin, K.; Wang, T.; Lou, Q.J.; Yuan, Z.; Huang, G.Q.; Liu, C.L. Pixel level segmentation of early-stage in-bag rice root for its architecture analysis. *Comput. Electron. Agric.* **2021**, *186*, 106197. [CrossRef]
14. 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; pp. 234–241.
15. He, K.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
16. 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, 18–23 June 2018; pp. 7132–7141.
17. Liu, X.; Hu, C.H.; Li, P.P. Automatic segmentation of overlapped poplar seedling leaves combining mask R-CNN and DBSCAN. *Comput. Electron. Agric.* **2020**, *178*, 105753. [CrossRef]
18. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
19. Elnashef, B.; Filin, S.; Lati, R.N. Tensor-based classification and segmentation of three-dimensional point clouds for organ-level plant phenotyping and growth analysis. *Comput. Electron. Agric.* **2019**, *156*, 51–61.
20. Wu, J.T.; Yang, G.J.; Yang, X.D.; Xu, B.; Han, L. Automatic counting of in situ rice seedlings from UAV images based on a deep fully convolutional neural network. *Remote Sens.* **2019**, *11*, 691.
21. Zou, K.L.; Chen, X.; Wang, Y.L.; Zhang, C.L.; 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]
22. Smith, A.G.; Petersen, J.; Selvan, R.; Rasmussen, C.R. Segmentation of roots in soil with U-Net. *Plant Methods* **2020**, *16*, 13.
23. Zhou, D.Y.; Li, M.; Li, Y.; Qi, J.T.; Liu, K. Detection of ground straw coverage under conservation tillage based on deep learning. *Comput. Electron. Agric.* **2020**, *172*, 105369. [CrossRef]
24. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N. Image segmentation using deep learning: A survey. *arXiv* **2020**, arXiv:2001.05566.
25. Qiao, W.T.; Ma, B.; Liu, Q.W.; Wu, X.G.; Li, G. Computer vision-based bridge damage detection using deep convolutional networks with expectation maximum attention module. *Sensors* **2021**, *21*, 824. [CrossRef] [PubMed]
26. Kamal, K.C.; Yin, Z.D.; Wu, M.Y.; Wu, Z.L. Depth-wise separable convolution architectures for plant disease classification. *Comput. Electron. Agric.* **2019**, *165*, 104948.
27. Szegedy, C.; Liu, W.; Jia, Y.Q.; Sermanet, P.; Redd, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
28. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
29. Li, X.; Zhong, Z.S.; Wu, J.L.; Yang, Y.B.; Lin, Z.C.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9167–9176.