forests

MDPI

Article

Aerial and Optical Images-Based Plant Species Segmentation Using Enhancing Nested Downsampling Features

Chih-Wei Lin [1,2,3,4,5,*], Mengxiang Lin [1] and Yu Hong [2,4,5]

[1] College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China; lmx@fafu.edu.cn
[2] College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, China; 2210430001@fafu.edu.cn
[3] Forestry Post-Doctoral Station, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[4] Key Laboratory of Fujian Universities for Ecology and Resource Statistics, Fuzhou 350002, China
[5] Cross-Strait Nature Reserve Research Center, Fujian Agriculture and Forestry University, Fuzhou 350002, China
* Correspondence: cwlin@fafu.edu.cn

Abstract: Plant species, structural combination, and spatial distribution in different regions should be adapted to local conditions, and the reasonable arrangement can bring the best ecological effect. Therefore, it is essential to understand the classification and distribution of plant species. This paper proposed an end-to-end network with Enhancing Nested Downsampling features (END-Net) to solve complex and challenging plant species segmentation tasks. There are two meaningful operations in the proposed network: (1) A compact and complete encoder–decoder structure nests in the down-sampling process; it makes each downsampling block obtain the equal feature size of input and output to get more in-depth plant species information. (2) The downsampling process of the encoder–decoder framework adopts a novel pixel-based enhance module. The enhanced module adaptively enhances each pixel's features with the designed learnable variable map, which is as large as the corresponding feature map and has $n \times n$ variables; it can capture and enhance each pixel's information flexibly effectively. In the experiments, our END-Net compared with eleven state-of-the-art semantic segmentation architectures on the self-collected dataset, it has the best PA (Pixel Accuracy) score and FWloU (Frequency Weighted Intersection over Union) accuracy and achieves 84.52% and 74.96%, respectively. END-Net is a lightweight model with excellent performance; it is practical in complex vegetation distribution with aerial and optical images. END-Net has the following merits: (1) The proposed enhancing module utilizes the learnable variable map to enhance features of each pixel adaptively. (2) We nest a tiny encoder–decoder module into the downsampling block to obtain the in-depth plant species features with the same scale in- and out-features. (3) We embed the enhancing module into the nested model to enhance and extract distinct plant species features. (4) We construct a specific plant dataset that collects the optical images-based plant picture captured by drone with sixteen species.

Keywords: deep learning; plant species; semantic segmentation; features enhancing

Citation: Lin, C.-W.; Lin, M.; Hong, Y. Aerial and Optical Images-Based Plant Species Segmentation Using Enhancing Nested Downsampling Features. *Forests* **2021**, *12*, 1695. https://doi.org/10.3390/f12121695

Academic Editor: Jozef Šibík

Received: 26 September 2021
Accepted: 24 November 2021
Published: 3 December 2021

1. Introduction

Forest plays a significant multi-dimensional role in human health and life [1]. The reasonable virescence can conserve soil and water, purify the air, adjust the temperature, and execute other ecological functions vital to maintaining the earth's ecological safety [2,3]. The distribution of plant species is the basis of establishing a stable ecosystem and plant community. Understanding the plants' distribution can be extremely helpful in environmental protection and resource development from a productional and academic perspective [4]. Therefore, it is essential to realize the plant species' classification and distribution and attract considerable attention [5,6]. The traditional manual investigation

is laborious and requires many experts to conduct an adequate investigation. Recently, the studies provided more insights into automatic classification through the plants' appearance characteristics due to the increased availability of plant images [4,7,8]. Moreover, it is popular to collect plant pictures from the web or initiatively capture pictures to create vegetation datasets [9]. According to these datasets, many meaningful studies have been executed, such as the individual plant species classification, large-scale plant images classification, and the multiple plant species segmentation.

In the individual plant species' classification research, an important step is to obtain plant organ features from images; for example, studies use the global features of leaf images [10], or the contour information of leaves [11] for classification. However, the image recognition of leaves is over fine and cannot play many roles in the structural composition and spatial distribution. Collecting the real plant information aims to understand the distribution of plants in an area quickly. The plant growth is contiguous and dense in the wild, where a small-scale image of a plant is not easily accessible. Therefore, an image classification method that can segment plant species from large-scale images is needed to capture all plants in a particular region. Some studies, including parametric and non-parametric methods, have the breakthrough advancement of plant classification due to the emergence of high spatial resolution satellite images. Parametric methods are the mainstream algorithms in plant automatic classification. In early studies, K-means [12], maximum likelihood (ML) [13], linear discriminant analysis (LDA) [14], and principal components analysis (PCA) [15] can be easily implemented for plants classification. However, these methods are affected by the distribution of training data. Therefore, parts studies propose non-parametric approaches, such as Random Forest (RF) [16] and support vector machine (SVM) [17], to overcome the problem of the parametric methods. However, these methods have low efficiency, weak identifying ability, and cannot effectively understand the distribution of plant species. Therefore, studies consider the segmentation techniques and attempt to solve these problems.

In image segmentation, the convolutional neural networks (CNNs) succeed in computer vision due to the fast development of computer power and are the mainstream method in image segmentation [18]. Many studies propose various CNN variants to improve the performance of plant segmentation on various scale images. The well-trained CNNs can extract plant features and achieve plant segmentation with good performance for the small-scale images [19]. To identify a plant from a large-scale image, CNNs are used as the viable tools to remote sensing (RS) data and successfully segment vegetation with high accuracy [20]. Moreover, CNNs have also been used to identify plant crowns using aerial RS data to segment plants [21]. However, the existing CNN segment plants into families rather than species, which is the limitation of most large-scale plant segmentation studies, and this is because most studies use large-scale remote sensing, which makes some plants in remote sensing images too small and cannot be well recognized. However, the plants' size in low-aerial images is moderate, which is helpful to segment and classify plant species accurately. Therefore, we build a specific plant dataset using unmanned aerial vehicles with an onboard optical camera and develop a novel approach for plant species pixel segmentation based on the self-collected dataset. The aerial and optical images-based plant species segmentation has the following challenges:

- The optical image only has RGB channels with little information compared to multispectral and hyperspectral images.
- The aerial image misses the details of the tree species, such as the leaf's texture, edge, and shape.
- There are many tree species and staggered with each other in the aerial images.

We demonstrate some examples which are related to the plants' segmentation, as shown in Figure 1. In Figure 1, each example has two images; left-hand is the input image, and right-hand is its result. Figure 1a–c can only address a single plant; Figure 1d use the high-resolution remote sensing images to segment the object without a fine-grained tree species segment; Figure 1e take RGB remote sensing imagery for plant detection. These

approaches cannot fine-grain segment the tree species in complex backgrounds. Figure 1f is our results which can segment interlaced tree species in a complex background.

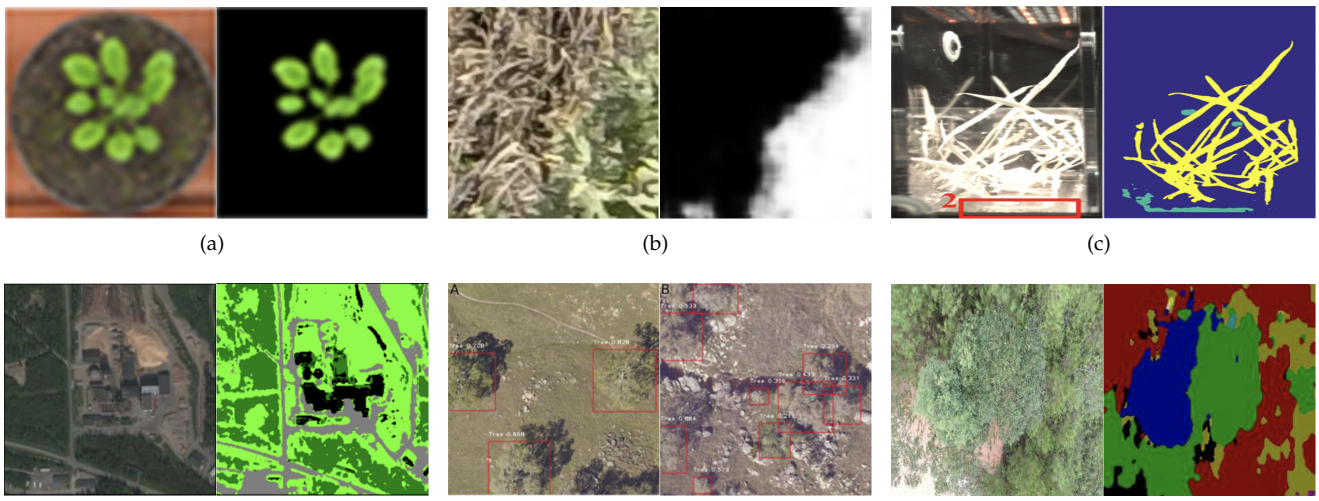


Figure 1. Some examples focus on the plants' issue. (**a**) Leaf count with optical images [22]; (**b**) Fig plant segmentation with RGB images [23]; (**c**) Rice segmentation with RGB images [24]; (**d**) Object segmentatio with high-resolution remote sensing images [20]; (**e**) Tree-crown detection with RGB imagery [21]; (**f**) Ours.

In this study, we analyze in depth the extraction of plant features with an information enhancement technique and develop the convolutional neural network using enhancing nested downsampling features, namely END-Net, which has a decent quality architecture with novel enhancing modules, for semantic segmentation of plants. The proposed END-Net nests a tiny encoder–decoder framework in each downsampling block to replace the original ordinary convolution operation and add the pixel-based enhancing module in each encoder block. The pixel-based enhancing module designs a learnable variable map with a size of n by n, the same size as the corresponding feature map, to adjust the enhancement information. Moreover, it associates with its corresponding features to obtain the enhanced features. In addition, we conduct extensive experiments with well-known semantic segmentation frameworks on self-built datasets and demonstrate the quantitative and qualitative results to prove that the proposed model is generally beneficial to the plant semantic segmentation task. In summary, our main contributions are four-fold:

- We propose a novel enhance module in this work; it is composed of a learnable variable map and can adaptively enhance each pixel's features. Moreover, the simplicity of the module makes it a plug-n-play module. In ablation analysis, the proposed variable map (plug-n-play module) improves the accuracy and is 0.45% and 0.60% higher than using a single variable in PA and FWIoU, respectively.
- We nest a tiny encoder–decoder framework in the process of downsampling to replace the original ordinary convolution operation, which can extract more in-depth information features of plant species. In ablation analysis, the accuracy of the network without the tiny encoder–decoder framework is 1.3% lower than the proposed network.
- The proposed END-Net has the advantage of the enhancing module and nested structure; it can extract much information and distinctive features of plant species to achieve the best performance on the self-collected plants' dataset (OAPI dataset) compared with other well-known methods. For example, the accuracies of END-Net are 18.49% and 20.17% higher than OCNet and ASPOCRNet in PA metric, respectively.
- We build an optical aerial plant image dataset named OAPI, which contains hundreds of optical aerial images and corresponding manual annotations. It is constructive for the study of plant species segmentation. To the best of our knowledge, the OAPI

dataset is a rare database mainly focused on collecting optical images captured by a low-altitude drone.

- This study focuses on fine-grained instance segmentation of plant species with optical aerial images, which is different from the existing studies such as single plant species segmentation [25], rough segmentation [26], plant crown segmentation [27], and single segmentation based on the current multi-functional technologies [28]. The segmentation accuracy of the proposed method is better than the newest semantic segmentation models from popular journals and conferences such as OCNet [29] and ASPOCRNet [30].

We organize the rest of this study as follows: Section 2 introduces related works, including semantic segmentation, aerial images semantic segmentation, and information fusion. Section 3 explains the network architecture of the proposed END-Net and the detail of the enhancing module. Section 4 describes the self-collected dataset, the OAPI dataset, with 16 tree species. Section 5 presents the self-built dataset, implementation details, and experimental results, including quantitative and qualitative results. Section 6 discusses the effect of various hyper-parameters settings, data expansion, external factors, and the advantages and limitations of the proposed model. We give the conclusions in Section 7.

2. Related Work

2.1. Semantic Segmentation

Semantic segmentation predicts each pixel's class in an image and conducts regional division according to each pixel's category. In semantic segmentation studies, FCN [31] is the classic deep learning approach; it proposes replacing the full connection of the network with a convolutional layer to obtain each pixel's category information and solve the segmentation problem. A growing number of followers adopt the point-by-point addition method of FCN to improve segmentation accuracy. Meanwhile, Unet [32] is another popular segmentation network different from FCN; it adopts the concatenation technique to fuse features to improve the segmentation effect. In addition, Deeplab [15] uses dilation convolution to expand the receptive field and performs multi-scale segmentation of objects based on the spatial pyramid. DeeplabV3+ [33] further applies the encoder–decoder structure with a practical decoder module to refine the segmentation boundaries and improve performance. Furthermore, DenseASPP [34] connects a set of atrous convolutional layers with the dense concept to generate multi-scale features without significantly increasing the model size. These methods provide high-quality results but are time-consuming. FC-densenet [35] extends DenseNets to deal with semantic segmentation without any further post-processing module, which has much fewer parameters. CCNet [36] adopted a novel criss-cross attention module that obtains all the pixels' contextual information to achieve high computational efficiency. These methods are well-known segmentation models and can be used for images from various sources.

2.2. Aerial Images Semantic Segmentation

The satellite and unmanned aerial vehicle(UAV) can provide high-resolution images, which contain abundant features for semantic segmentation [37]. Marmanis [38] takes several FCNs methods to deliver clutter pixel classification in high-resolution aerial images of urban areas; he discusses the network's design choices and the intricacies and demonstrates that a combination of several networks is effective. RoadNet [39] takes SegNet [40] as the backbone to execute realistic navigation over roads; it partitions the image into 'road' or 'non-road' and uses the Hough Transform as the post-process to obtain the best road curvature contour. RA-FCN [41] introduces two plug-and-play modules, the spatial relation module, and the channel relation module, to generate relation-augmented features, and utilizes relation-augmented structure to enable spatially and channel relational reasoning for aerial image segmentation. Benjdira [42] designs an algorithm using generative adversarial networks (GANs) to reduce the domain shift impact in semantic segmentation of aerial images, and the problems of the domain shift between the new city image and the source

data set are solved. Zhang [43] built the NWPU-YRCC dataset for river ice segmentation and the proposed ICENET, including fuse module and attention modules, which achieve efficient features fusion and better results. The existing semantic segmentation models of aerial images mainly focus on the segmentation of roads and buildings and lack the segmentation and classification of plants.

2.3. Information Fusion

The information in different regions or scales to perform feature enhancement is critical for achieving good pixel segmentation. AdapNet [44] proposes a convoluted mixture of deep experts (CMoDE) fusion technique that adaptively weighs class-specific features of expert networks and further learns fused representations to yield robust segmentation. PSPNet [45] explores the benefits of contextual information aggregation for segmentation and exploits the pyramid pooling modules to fuse context information of different regions. Refinenet [46] uses long-range residual connections to achieve all the information available during the downsampling process and capture high-level semantic features and rich background context effectively. Deepresunet [47] uses the rich skip connections within the network to fuse features of images that can facilitate information propagation and enhance information. HDC [48] aggregates global information by expanding the receptive field and designs dense upsampling convolution to keep more detailed features. SDN stacks multiple shallow deconvolutional networks to aggregate contextual information, and a clear segmentation boundary is obtained. The inter-unit and intra-unit connections of the designed SDN [49] unit can enhance information aggregation and the discrimination of feature representations. Recently, Fast-SCNN [50] combines high-resolution detail with lower features and efficiently performs real-time semantic segmentation on high-resolution image data. The information aggregation of these methods mainly focuses on combining global or local information and improving details of pixel segmentation. These information fusion methods are specific to their respective models but lack information enhancement modules for plant species segmentation that can be applied to various models.

To sum up, there are existing segmentation models for aerial images, but the optical aerial plants' image segmentation and plants' fusing and enhancing feature information are lost. Therefore, this study proposes a plant species segmentation model with enhancing nested downsampling features to achieve aerial image segmentation of plant species.

3. Methodology

This section introduces the details of the proposed convolution neural network with enhancing nested downsampling features (END-Net). Then, we describe the downsampling-block with an encoder–decoder structure (DS_{ED}-block) in detail. Next, we present the proposed pixel-based enhancing module and its related equations. Finally, we illustrate the multi-loss approach that is used in END-Net.

3.1. Network Architecture

The proposed convolution neural network with enhancing nested downsampling features (END-Net) has three main components, including the downsampling block with an encoder–decoder structure (DS_{ED}-block), the pixel-based enhancing module, and the multi-loss strategy. END-Net is a simple and efficient construction for plant semantic segmentation, and its illustration is demonstrated in Figure 2.

In Figure 2, the image with the size of $H \times W$ sequentially executes the procedures of three DS_{ED}-blocks, one convolution block (*Conv*. block), and three upsampling blocks (*US*-blocks). The number in the lower right corner of each operation represents the number of feature maps obtained after the operation. Each DS_{ED}-block's internal structure is similar to Unet but has the innovative improvement, which is the proposed pixel-based enhancing model, in the encoding processing; its details' structure and description as presented in Section 3.2. After processing each DS_{ED}-block, the feature size is reduced to 1/2, 1/4, and 1/8 of the original image using max-pooling operation. Moreover, we set the *Conv*.

block, which contains two convolution layers with stride size 1, between DS_{ED}-blocks and US-blocks; it connects the DS_{ED}-blocks and US-blocks using the un-shrinking feature map. Next, the upsampling blocks (US-blocks) aggregate contextual information from US-blocks and DS_{ED}-blocks, forming feature maps that contain contextual information as intermediate output. In the US-block, feature maps sequentially operate convolution, upsampling, concatenation, and convolution operations. In concatenation operation, we fuse the features of DS_{ED}-blocks and US-blocks with the same size. Notice that each US-block has one less convolutional layer compared to the upsampling module of Unet. Finally, feature maps decode into a predicted image, which has the same size as the input image.

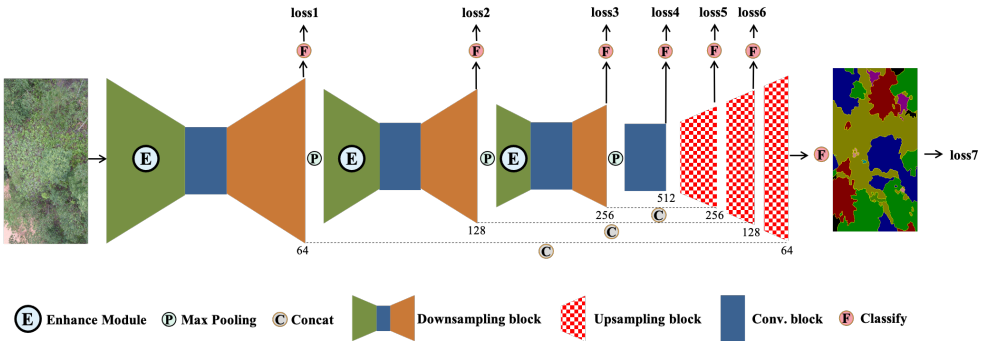


Figure 2. The architecture of the proposed framework, END-Net.

To assist the segmentation process, we add extra short-range connections between each encoding module and its corresponding decoding module at each DS_{ED}-block. Moreover, we apply the pixel-based enhancing module into each encoding module of the DS_{ED}-block. The pixel-based enhancing module emphasizes an essential position in feature maps and can extract potential features. More details of the pixel-based enhancing module are described in Section 3.3. Furthermore, we execute the deep-supervision into DS_{ED}-blocks, $Conv$. block, and US-blocks using the multi-loss approach to effectively auxiliary feature learning and introduce the details in Section 3.4.

3.2. Downsampling Block in END-Net

To obtain abundant plants' feature information during each down-sampling, we add the pixel-based enhancing module into the encoding module of each DS_{ED}-block. The enhancing module can adaptively learn the weight for each pixel and obtain the discriminative features with abstract semantic segmentation information. The DS_{ED}-block composes a tiny encode-decode framework and is similar to the overall architecture of the proposed END-Net. Hence, the complete END-Net looks like a nested structure network. The illustration of the DS_{ED}-blocks is shown in Figure 3.

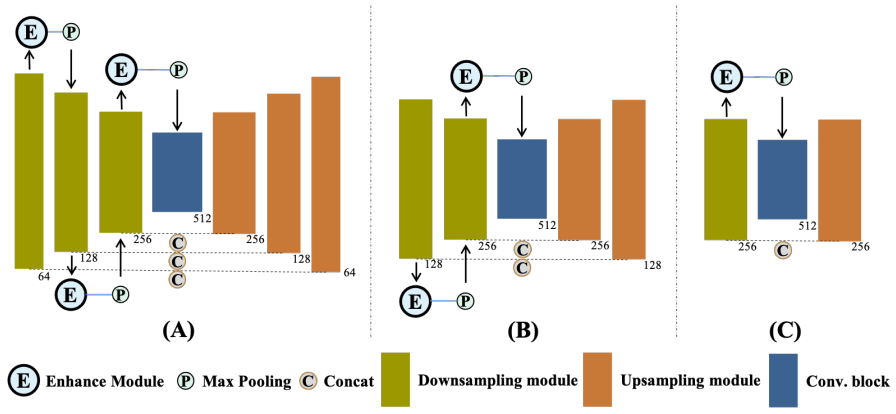


Figure 3. An illustration of the downsampling blocks in our proposed network.

In Figure 3, there are three types of DS_{ED}-blocks (A) (B) (C) with different complexity and be sequentially arranged in the process of downsampling. The number in the lower right corner of each operation represents the number of feature maps obtained after the operation. Each DS_{ED}-block has a various number of downsampling modules and sequentially operates the downsampling modules, *Conv*. block, and the upsampling modules. Moreover, each DS_{ED}-blocks operate the proposed pixel-based enhancing nodule and the max-pooling operation on each set of adjacent downsampling modules to enhance the feature of each pixel adaptively.

The feature maps sequentially go through the DS_{ED}-blocks, its size is gradually reduced to $1/2$, $1/4$, and $1/8$ of the original size, and is recovered to its original size by sequentially executing the upsampling modules. The proposed model conducts several operations in the downsampling process to extract more abstract features. Moreover, we do not enhance the sparse upsampling features but enhance the compacted downsampling features due to the sparsity of the upsampling features.

3.3. The Pixel-Based Enhancing Module

To overcome the segmentation difficulty caused by the intricate image content of the plant, we introduce the novel enhancing module, namely the pixel-based enhancing module, which applies operations based on pixels on feature maps of the downsampling to complete the enhancing operation. The pixel-based enhancing module mainly contains a learnable variable map, which considers the importance of the features' distribution based on the pixels, and its framework is shown in Figure 4. In Figure 4, the input is the logits from original feature maps and is denoted as X with size $H \times W$. The pixel-based enhancing module applies an activation function on feature map X_1 to generate a weight array with values between 0 to 1 to indicate the pixels' importance on convolution operation. Then, to create the enhancing parameters, we multiply the activation results with a learnable variable map k, which has the same size as X_1. Finally, we multiply X_1 with the enhancing parameters and add with X_1 itself.

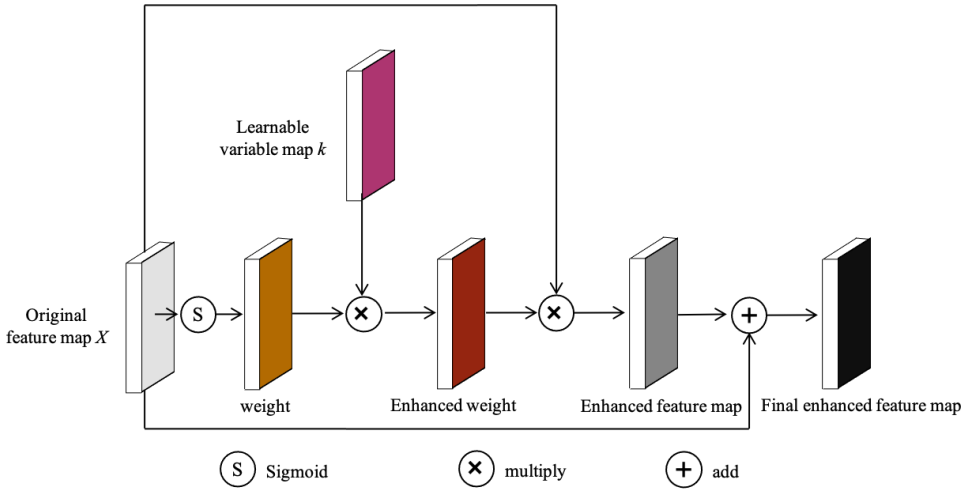


Figure 4. The computing details of the pixel-based enhancing module.

The pixel-based enhancing module can enhance each pixel with its characteristic, expand the essential pixels' effect, make the critical areas in the feature map more prominent, and make the model pay more attention to this area. The operations are sequentially defined as follows:

$$E_{X_1} = P \cdot X_1 + X_1 \tag{1}$$

$$P = S_{X_1} \cdot k \tag{2}$$

$$S_{X_1} = \delta(X_1) \tag{3}$$

where S_{X_1} is the output of executing the activation function on the feature map X_1. δ means activation function. P is the enhanced parameter generated by producing S_{X_1} with a learnable variable map k. E_{X_1} refers to the result of the proposed enhancing module, which is the sum of $P \cdot X_1$ and X_1.

Moreover, we further associate the pixel-based enhanced features with the successive convolution process and demonstrate the framework in Figure 5. In Figure 5, we construct a short connection between two neighboring convolution layers, which are in the same convolution block. We operate the proposed pixel-based enhancing module to enhance the output of the first convolution layer X_1 and generate the enhanced features E_{X_1}. Then, we add E_{X_1} into X_2, which is the output of the second convolution, to finish the process of the short connection. The detail operation in Figure 5 is defined as follows:

$$E_{DS} = E_{X_1} + X_2 = E_{X_1} + F(X_1) \tag{4}$$

where E_{DS} is the result with enhancing module in down-sampling. $F(.)$ denotes the convolution between two layers. X_2 is the convolution result of X_1.

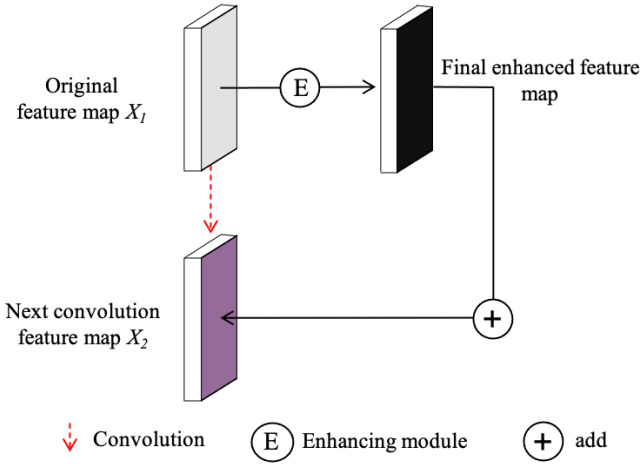


Figure 5. The illustration of the combination for the proposed model.

In addition, we demonstrate the flow chart of using the proposed pixel-based enhancing module, as shown in Figure 6. In Figure 6, the first step is to select a feature map X_1, and that is used twice; one is for the pixel-based enhancing module, and the other one is to generate a new feature map X_2 in Step 6 and be used to add with the result from the pixel-based enhancing module. Step 2 to Step 5 are operated in the pixel-based enhancing module to generate the enhanced feature map E_{X_1}. Moreover, the enhanced feature map E_{X_1} is added with X_2 in Step 7 to generate the enhanced feature map E_{DS}.

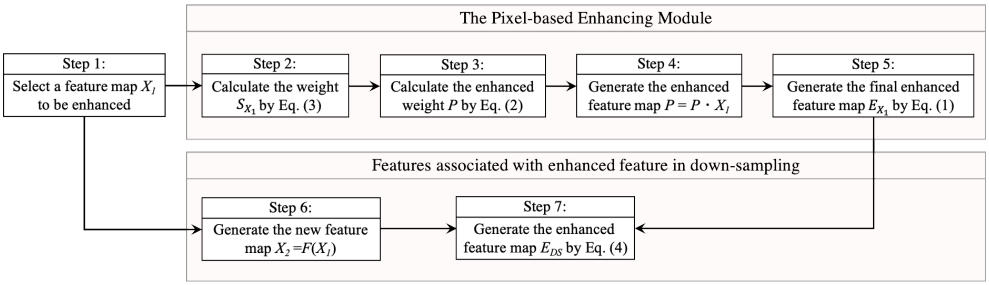


Figure 6. The flow chart of using the proposed pixel-based enhancing module.

3.4. The Multi-Loss Strategy

To effectively train the proposed network, we consider the deep-supervision into various parts of the network as shown in Figure 2. We respectively evaluate the loss of

DS_{ED}-blocks, *Conv*. block, *US*-blocks, and END-Net using the cross-entropy loss function. The equations of the cross-entropy evaluation are expressed as follows:

$$loss_j = -\sum_{i=1}^{c} l_i \cdot log(y_i) \quad j = 1, 2, 3, \ldots, n \tag{5}$$

$$loss_k = \sum_{j=1}^{n} M \cdot loss_j \quad k = 1, 2, 3, \ldots, N \tag{6}$$

$$loss = \sum_{k=1}^{N} loss_k \tag{7}$$

where $loss_j$ refers to the loss value of pixel j, n means the total number of pixels. l_i is the label of each pixel that corresponds to class i. If the pixel belongs to class i, l_i sets to 1; otherwise, it sets to 0. $loss_k$ is the total loss of each image k. N is the total number of the training image. To address the problem of some plant species having little data, we design an image mask M to ignore these species in both training and testing. The mask value is set to 1 if the species has been considered in the classification; otherwise, it is set to 0. Equation (7) represents the calculation of each loss in Figure 2, which is the summation of all training image loss. The total loss is evaluated with the following:

$$Loss = \sum_{l=1}^{7} loss_l \tag{8}$$

where *Loss* means total loss of the proposed network, and the number of losses is from 1 to 7, which represents the loss at each location in Figure 2.

In addition, we demonstrate the backpropagation of the proposed pixel-based enhancing module to present the adjustment mechanism as follows:

$$\frac{\partial Loss}{\partial \mathbf{O}} = \frac{\partial Loss}{\partial \mathbf{O}_{E_{ds}}} \cdot \frac{\partial \mathbf{O}_{E_{ds}}}{\partial \mathbf{O}} = \frac{\partial Loss}{\partial \mathbf{O}_{E_{ds}}} \cdot (P+1) = \frac{\partial Loss}{\partial \mathbf{O}_{E_{ds}}} \cdot (S_X \cdot k + 1) \tag{9}$$

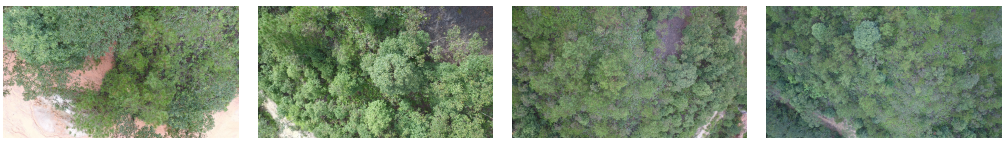
where $\mathbf{O}$ is the network's output and $\mathbf{O}_{E_{ds}}$ is the output of the enhancing module. We can see that the optimization and updates are related to the variable k, which is embedded in our proposed pixel-based enhancing module.

4. Dataset

We aim to realize the semantic plant species understanding using the aerial image. However, the existing public plant datasets set over-process plant images, which means there is only a plant or plant organ per image, and they are mainly designed for classification without the ground truth of segmentation. Therefore, the existing dataset does not satisfy our research goal, prompting us to build Optical and Aerial Plant species Images, namely the OAPI dataset.

We take Anxi and Changting counties as the study area, which suffered from severe soil erosion and vegetation restoration and is significant in research. We carefully designed the recording data to capture abundant plant information per image at high resolution and acquired thousands of aerial images from a moving UAV in the summer. We use an unmanned aerial vehicle (UAV) equipped with an optical camera to capture aerial images of the vegetation from a bird's eye view. The UAV model is the DJI inspire one raw, and its specific parameters are as follows: rotation angular velocity pitching axis is 300/s, and the heading axis is 150/s. The optical camera is a ZenmuseX5R and provides an image with a size of 4608×2592.

We consider different aerial shooting heights, set sampling height range at 20–100 m, and mainly concentrate at 20–60 m as shown in Figure 7 to verify the robustness of the proposed END-Net. In Figure 7, the leaves' appearance, such as texture, shape, and color, can be easily distinguished at low altitudes. However, the plants have a small area as the increasing of aerial height; they may be ignored and add the difficulty of segmentation. Finally, we selected 592 images to produce the training and testing sets.



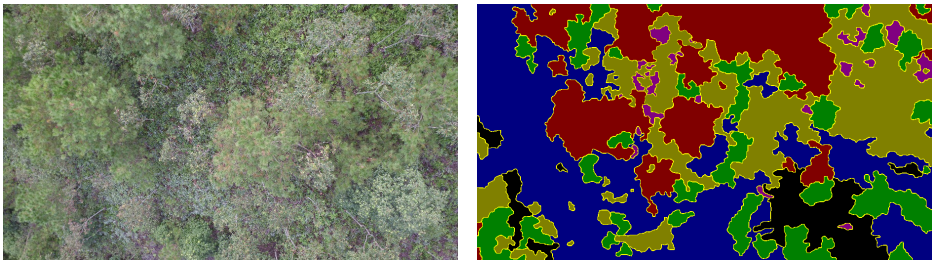
(a) 20 m (b) 40 m (c) 60 m (d) 80 m

Figure 7. The aerial images at different altitudes.

We invite relevant professionals to label the selected images by using various colors based on the distribution of plant species and ensure that each class has its unique corresponding color. The OAPI dataset has 16 classes as shown in Table 1: Background (#0), *Pinus massoniana* (#1), *Eucalyptus citriodora* (#2), *Dicranopteris dichotoma* (#3), *Photinia serrulata* (#4), *Adenanthera pavonina* (#5), *Blechnum orientale* (#6), *Miscanthus sinensis* (#7), *Withered dicranopteris dichotoma* (#8), *Withered pinus massoniana* (#9), Unkown (#10), Stone (#11), *Schima superba* (#12), *Mosla chinensis* (#13), *Carmona microphylla* (#14), *Liquidambar formosana* (#15). In Table 1, we can observe that the morphological characteristics of most plants are different. For example, the leaf of #1 is fasciculate, slender, and slightly twisted, #2 is narrow and needle-shaped, #15 is thin, leathery, and broadly ovate. However, aerial images sometimes fail to show clear differences, such as #4 and #5. In Figure 8, we demonstrate an optical and aerial image with ground truth, and we mark the same plant with the same color and outline the intersections of the different plants with yellow according to distribution.

Table 1. Plants' images and their corresponding ground truth's colors.

Class	**#0**	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**
Image								
GT								
Class	**#8**	**#9**	**#10**	**#11**	**#12**	**#13**	**#14**	**#15**
Image								
GT								



(a) Original (b) Ground truth

Figure 8. Example of the original image and its ground truth.

5. Experiments

This section firstly describes the experimental settings and the evaluation indicators of semantic segmentation. Then, we sequentially introduce the quantitative evaluation of our proposed END-Net against the eleven well-known methods, present the qualitative result with visualization, and execute the diagnostic and ablation experiments to evaluate the feasibility and robustness of END-Net. All experiments were executed on Ubuntu 16.04 using an NVIDIA 1080 graphics card, and all experimental setting parameters are consistent.

5.1. Implementation Details

We implement our model in Tensorflow and execute all experiments on a workstation with NVIDIA 1080 (11 G) under the Ubuntu16.04 system. Moreover, the hyper-parameters are: the training epoch is set to 400, but they will be terminated when the over-fitting phenomenon occurs; each batch has eight images per GPU; and the dropout rate of the proposed net is 0.5. The initial learning rate is 0.0003; the optimizer is Adam-Optimizer; images are resized into 224×224; and there are 414/178 images for training and testing, respectively. The learnable variable map k is initialized to 1.

5.2. Evaluation Indicators for Semantic Segmentation

In this study, we take PA (Pixel Accuracy), MPA (Mean Pixel Accuracy), MIoU (Mean Intersection over Union), and FWIoU (Frequency Weighted Intersection over Union), as the evaluation indicators with parameters: TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative). TP (True Positive) means that the model's prediction is a positive example, and actual observation is a positive example. FP (False Positive) means that the model's prediction is a negative example, and actual observation is a positive example. TN (True Negative) means that the model's prediction is a negative example, and actual observation is a negative example. FN (False Negative) means that the model's prediction is negative, and actual observation is positive.

5.2.1. PA (Pixel Accuracy)

PA means the ratio of correctly classified pixel points to all pixel points. PA can show the classification accuracy of the whole image. The confusion matrix can calculate the value of PA. The equation of the PA is shown as follows:

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

5.2.2. FWIoU (Frequency Weighted Intersection over Union)

FWIoU is the promotion of MIoU (Mean Intersection over Union). It means the IoU of each class is weighted and summarized according to the frequency of each class's occurrence. The IoU is the ratio of the intersection and union of the predicted results and the true values of a given class. The confusion matrix can also calculate the value of IoU and FWIoU. The equation of the IoU and FWIoU are shown as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

$$FWIoU = \frac{TP + FN}{TP + FP + TN + FN} \cdot \frac{TP}{TP + FP + FN} \tag{12}$$

5.3. Quantitative Analysis

In the quantitative analysis, we compare the proposed END-Net with eleven well-known semantic segmentation architectures on the self-collected dataset (OAPI dataset), including Unet [32], FCN [31], Refinenet [46], FC-densenet [35], FRRN [51], Deepresunet [47], BiSeNet [52], DANet [53], CFNet [54], ASPOCRNet [30], and OCNet [29], and adopt PA, FWIoU, MPA, MIoU, Parameters (*Params*), and FPS metrics to demonstrate the validity of the proposed method, as shown in Table 2.

In Table 2, the proposed END-Net has the best accuracy in both PA, FWIoU, MPA, and MIoU metrics and achieves 84.52%, 74.96%, 52.17%, and 37.49%, respectively. It is 1.98%, 3.03%, 0.56%, and 1.36% higher than the second-best approach (Unet) in PA, FWIoU, MPA, and MIoU metrics, respectively, and indicates that our improvement is noticeable compared to our backbone (Unet). Moreover, the compared methods which take the ResNet101 as the backbone have low accuracies in both metrics. It illustrates that the ResNet101 is not suitable for our dataset. In addition, we also take FCN as the backbone and use the proposed pixel-based enhancing module in the downsampling block and achieve 81.61%, 70.36%, 45.14%, and 34.14% in PA, FWIoU, MPA, and MIoU metrics. It is 4.90%, 7.15%, 3.80%, and 5.21% higher than FCN in PA, FWIoU, MPA, and MIoU metrics,

respectively. In the *Params*, our net is the third-smallest model and has the best accuracies in PA, FWIoU, MPA, and MIoU metrics. Compared to the second-best method (Unet), our model is 5.6 M smaller than Unet. In the FPS metric, the execution speed of most the methods is between 11 and 12 FPS, in which OCNet has the best FPS and achieves 14.35. The FPS of the proposed method is 11.42, which is close to most of the methods.

Table 2. Comparison with well-known networks on the OAPI dataset.

Method	Year	Source	Backbone	PA (%)	FWIoU (%)	MPA	MIoU	*Params* (M)	FPS
Unet [32]	2015	MICCAI[C]	Unet	82.54	71.93	51.61	36.13	34.51	12.40
FCN [31]	2015	CVPR[C]	VGG19	76.72	63.21	41.34	28.93	139.74	12.85
Refinenet [46]	2017	CVPR[C]	ResNet101	52.00	36.10	25.99	17.46	85.69	12.24
FC-densenet [35]	2017	CVPR[C]	DenseNet56	81.43	70.52	38.51	27.78	1.37	11.86
FRRN [51]	2017	CVPR[C]	FRRN-A	76.35	64.91	44.75	30.62	17.74	11.62
Deepresunet [47]	2018	GRSL[J]	Unet	77.90	65.71	48.65	34.34	32.60	11.64
BiSeNet [52]	2018	ECCV[C]	ResNet101	69.46	54.66	29.98	20.73	47.79	13.71
DANet [53]	2019	CVPR[C]	ResNet101	63.35	48.37	26.77	18.07	46.29	13.79
CFNet [54]	2019	CVPR[C]	ResNet101	64.78	49.48	30.00	19.65	48.70	12.40
ASPOCRNet [30]	2020	ECCV[C]	ResNet101	64.35	48.61	25.65	17.50	47.21	12.37
OCNet [29]	2021	IJCV[C]	ResNet101	66.03	49.90	29.87	18.77	43.93	14.35
FCN + $(D-k)$	2021	-	VGG19	81.62	70.36	45.14	34.14	140.81	12.05
END-Net$(D-k)$	2021	-	Unet	**84.52**	**74.96**	**52.17**	**37.49**	28.91	11.42

Additionally, we select three situations: (I) common tree species, (II) more and scattered tree species, and (III) uncommon categories to present the robustness of our model and demonstrate the fine-grained segmentation results of these situations in Tables 3–5, respectively. In Tables 3–5, we use the PA metric to present the performance of each model in classifying each plant species; ID is the number of plant species, the model is the name of the compared method, and *overall* is the PA metric of the whole image. The best accuracy is marked as red, and the second-best accuracy is marked as green.

In Table 3, it presents situation I, which is the aerial image that contains the common tree species. The accuracy of the proposed framework, including FCN + $(D-k)$ and END-Net, has almost the best accuracy on classifying tree species except #0. The proposed framework does not have the best accuracy on classifying #0 tree species but is only 0.27%, slightly lower than the best approach. In Table 4, each tree species on the image is widely distributed, and the number of tree species is more than the situation I. The proposed END-Net has good performance and the best overall accuracy except recognizing #0 (the background). Our model has the best accuracy for classifying *Blechnum orientale* (#6), which has fewer pixels on the image, and compared models do not perform well. Moreover, FCN + $(D-k)$ has the second-best overall accuracy, which embeds the proposed pixel-based enhancing module in the FCN's downsampling blocks. In Table 5, the number of tree species and that of rare tree species are increasing compared to situations I and II. None of the models perform well for most tree species, but the proposed END-Net and FCN$(D-k)$ have the best overall accuracies. More specifically, the proposed framework has performed well for most tree species compared to the compared methods.

Table 3. Performance of each model on a small number of a common classes image.

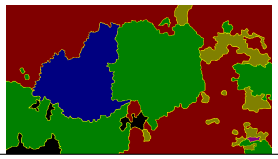
Items	**Input Image**			**Ground Truth**		
Images						
ID / Model	#0	#1	#2	#3	#4	*overall*
Uet	96.14%	**88.64%**	74.83%	83.87%	88.19%	83.37%
FCN	97.34%	75.21%	77.59%	68.83%	78.16%	76.54%
Refinenet	98.80%	83.30%	43.06%	56.13%	37.65%	61.75%
FC-densenet	**100.00%**	80.45%	52.17%	65.89%	81.84%	77.15%
FRRN	99.87%	68.17%	54.00%	69.65%	75.40%	64.65%
Deepresunet	98.27%	74.83%	82.46%	84.90%	89.03%	80.29%
BiSeNet	**100.00%**	73.22%	31.21%	75.64%	51.08%	55.67%
DANet	99.87%	73.43%	19.40%	46.70%	41.19%	48.97%
CFNet	**100.00%**	73.70%	21.02%	33.51%	50.53%	50.40%
ASPOCRNet	**100.00%**	83.87%	19.33%	30.63%	40.07%	52.79%
OCNet	**100.00%**	77.64%	29.74%	20.00%	43.89%	53.81%
FCN + $(D-k)$	99.73%	83.24%	71.61%	**93.30%**	**99.58%**	81.96%
Ours	99.73%	85.12%	**88.41%**	89.05%	97.80%	**88.40%**

Table 4. Performance of each model on an image with more classes and more dispersed class distribution.

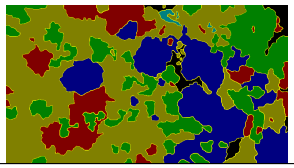
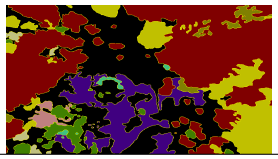
Items	**Input Image**			**Ground Truth**			
Images							
ID / Model	#0	#1	#2	#3	#4	#6	*overall*
Uet	96.31%	**97.14%**	78.98%	72.41%	85.17%	85.00%	80.12%
FCN	89.30%	85.08%	85.70%	74.21%	91.26%	91.25%	82.01%
Refinenet	73.92%	89.33%	65.18%	28.21%	21.92%	0.00%	41.97%
FC-densenet	97.54%	93.24%	31.10%	60.80%	42.87%	0.00%	55.54%
FRRN	97.79%	91.71%	36.53%	44.01%	41.76%	13.75%	48.86%
Deepresunet	83.27%	93.84%	70.84%	68.39%	70.62%	6.25%	72.61%
BiSeNet	**99.26%**	86.83%	16.99%	40.25%	35.87%	0.00%	41.58%
DANet	93.23%	90.89%	13.06%	36.83%	29.70%	0.00%	38.27%
CFNet	95.94%	91.66%	20.47%	37.71%	34.91%	15.00%	41.45%
ASPOCRNet	94.83%	92.37%	20.27%	32.43%	21.29%	0.00%	35.97%
OCNet	95.45%	92.33%	18.20%	34.68%	19.29%	0.00%	36.10%
FCN + $(D-k)$	90.41%	92.75%	86.76%	73.52%	89.67%	0.00%	82.30%
Ours	64.70%	**95.81%**	**87.59%**	**79.13%**	**95.85%**	**95.00%**	**86.36%**

Table 5. Performance of each model on a greater number of rare classes images.

Items	**Input Image**					**Ground Truth**				
Images										
ID / Model	#0	#1	#3	#7	#9	#10	#12	#14	#15	*overall*
Uet	96.68%	89.12%	64.54%	0.00%	0.39%	0.00%	15.26%	0.00%	0.00%	69.17%
FCN	98.42%	88.11%	39.01%	0.00%	0.81%	1.65%	19.38%	0.00%	5.12%	69.57%
Refinenet	94.75%	70.95%	21.51%	0.00%	0.00%	2.20%	0.68%	**13.43%**	0.93%	58.51%
FC-densenet	92.78%	82.19%	57.45%	0.00%	0.00%	4.68%	9.72%	0.00%	**18.60%**	64.39%
FRRN	89.39%	72.03%	65.25%	**6.07%**	0.00%	1.65%	30.90%	0.00%	7.44%	61.83%
Deepresunet	95.39%	89.24%	**65.48%**	1.76%	0.65%	26.45%	36.66%	0.00%	0.00%	71.86%
BiSeNet	90.44%	74.49%	33.81%	0.00%	0.00%	0.00%	3.78%	0.00%	0.00%	59.25%
DANet	75.24%	63.88%	16.55%	0.00%	0.00%	0.00%	21.41%	0.00%	8.84%	52.32%
CFNet	76.48%	77.71%	9.93%	0.00%	0.00%	0.00%	3.63%	0.00%	0.00%	56.29%
ASPOCRNet	68.28%	76.10%	24.11%	0.00%	0.00%	0.00%	1.36%	0.00%	0.00%	53.08%
OCNet	78.92%	78.27%	22.22%	0.00%	**1.07%**	0.00%	12.28%	0.00%	0.00%	58.56%
FCN + $(D-k)$	**99.51%**	84.72%	13.48%	0.00%	0.00%	**84.30%**	54.71%	0.00%	0.00%	73.23%
Ours	95.83%	**90.14%**	64.54%	0.00%	0.00%	0.00%	**88.73%**	0.00%	0.00%	**78.51%**

5.4. Qualitative Analysis

This section provides the visualization predicted results to describe the capability of each approach and shown in Table 6, respectively. In Table 6, we consider four situations: (A) few tree species with relatively concentrated distribution, (B) a piece of bare land and more tree species contain some rare classes, (C) a relatively scattered distribution of the same tree, and (D) a lot of bare land with many tree species, the number of each tree species is limited and contains most of the rare categories.

In Table 6, most of the networks have difficulty identifying the tree species on the proposed dataset, and their performance is not very ideal. However, the proposed net has the best performance in both situations, tree species with dense distribution, and tree species with scattered distribution in the images. In situation A, the proposed framework has the best visualization result compared to the other methods compared, which have broken segmentation results similar to the ground truth. In situation B, all nets do not perform well, especially the *Withered pinus massoniana* (#9). However, our model has the best overall performance and recognition performances of *Pinus massoniana* (#1), *Dicranopteris dichotoma* (#3), and *Schima superba* (#12) are very close to the Ground Truth. In situation C, the shape of the recognition result of our model for each tree is similar to Ground Truth, but it is slightly insufficient in details, and it also performs best among all the models on the whole. In situation D, the performance of all models is poor in this situation, but the proposed net has the best performance compared to the compared methods. Overall, our model has the best performance in various situations; the FCN + $(D-k)$ takes the VGG as the backbone and embeds the proposed pixel-base enhancing module into the downsampling blocks, and it has the second-best performance; the rest of the compared methods with Resnet101 have the worst performance.

5.5. Diagnostic and Ablation Experiments

In this subsection, we execute diagnostic and ablation experiments to present the feasibility and effectiveness of the proposed network.

5.5.1. Diagnostic Experiments

In the diagnostic experiments, we execute the significance testing to prove the significance of the proposed enhancing model. *Significance test:* We execute the paired-samples T-test as the significance testing to verify the significance of the pixel-based enhancing module on the self-collected datasets and demonstrate the testing result with two metrics in Table 7. In Table 7, X, Y, "Sig. (2-tailed)", and Δ refer to the proposed network with $n \times n$ variable map in enhancing module, the proposed network with 1 variable map in enhancing module, the p-value of the two-sided significance, and the difference between X and Y for each fold cross-validation. In Table 7, the performance of X is higher than that of Y at each validation. Moreover, significant p-values (Sig.) are all less than 0.05 with these two metrics. The testing proves that the proposed pixel-based enhancing module's performance is significant; and the proposed module can efficiently increase the accuracy.

Table 6. Visual schematic of all models.

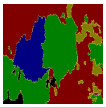
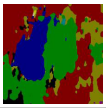
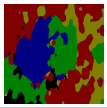
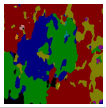
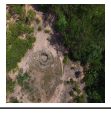
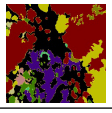
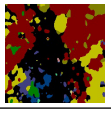
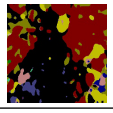
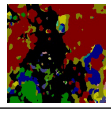
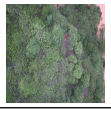
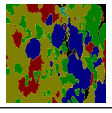
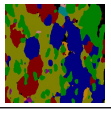
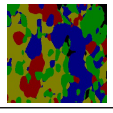
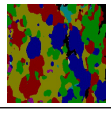
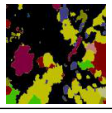
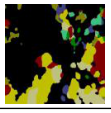
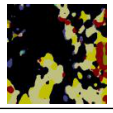
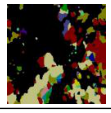
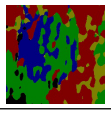
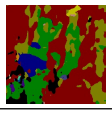
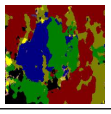
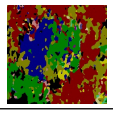
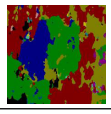
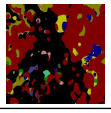
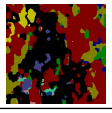
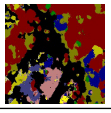
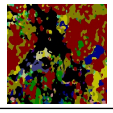
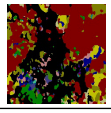
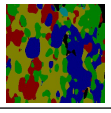
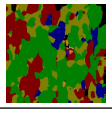
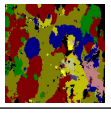
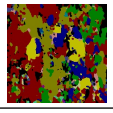
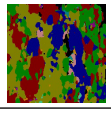
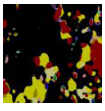
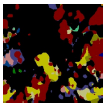
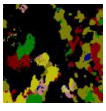
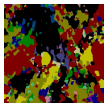
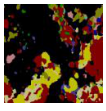
	Image	**GT**	**Ours**	**FCN + ($D - k$)**	**Unet**
1					
2					
3					
4					
	FCN	**Refinenet**	**FC-Densenet**	**FRRN**	**Deepresunet**
1					
2					
3					
4					

Table 6. *Cont.*

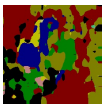
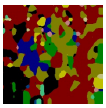
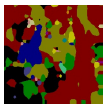
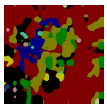
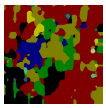
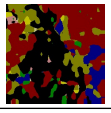
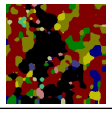
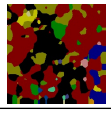
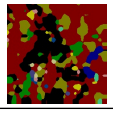
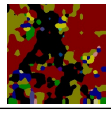
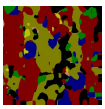
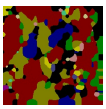
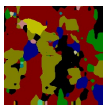
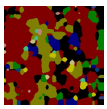
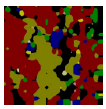
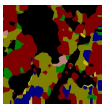
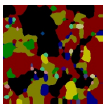
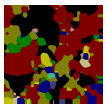
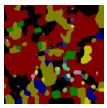
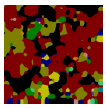
	BiSeNet	DaNet	CFNet	ASPOCRNet	OCNet
1					
2					
3					
4					

Table 7. Significance testing results of 5-fold cross-validation.

Metrics	PA (%)			FWIoU (%)		
Sample / #Fold	X	Y	Δ	X	Y	Δ
1	84.05	83.31	0.74	74.25	73.49	0.76
2	83.95	83.56	0.39	73.75	73.27	0.48
3	85.02	84.15	0.87	75.49	73.84	1.65
4	83.69	83.16	0.53	73.37	72.90	0.47
5	84.33	82.94	1.39	74.40	72.75	1.65
Sig. (2-tailed)		0.010			0.021	

5.5.2. Ablation Experiments

In this subsection, we conduct the ablation experiments on OAPI with three settings to verify the rationality and scientificity of the proposed END-Net: (1) locations of pixel-based enhancing module, (2) structures of variable map, and (3) loss strategies.

Locations of pixel-based enhancing module: We set the proposed pixel-based enhancing module at various locations of the END-Net to explore the optimal settings and demonstrate the results in Table 8. In Table 8, we consider four locations: (a) downsampling blocks (*DS*-blocks), (b) upsampling blocks (*US*-blocks), (c) $DS + US$-blocks, and (d) *NONE*. More specially, situations (a) and (b) only set the proposed pixel-based enhancing module at downsampling and upsampling blocks, respectively; situation (c) uses the enhancing module at both blocks; and situation (d) does not consider the enhancing module in the network. The network with the proposed pixel-based enhancing module at *DS*-blocks has the best accuracies and achieves 84.52% and 74.96% in PA and FWIoU metrics. Its PA accuracy is 1.64% higher than *US*-blocks, 1.75% higher than $DS + US$-blocks, 1.9% higher than *NONE*; its FWIoU is 2.32% higher than *US*-blocks 1.95% higher than $DS + US$-blocks, and 2.73% higher than *NONE*. Its accuracy is significantly higher than the rest locations and ensures that using the pixel-based enhancing module at downsampling blocks can effectively achieve the best results. The tiny encoder–decoder structure with a pixel-based enhancing module embedded in a downsample block that can efficiently improve the segmentation accuracy due to the downsampling blocks (DS-blocks) is the procedure of downsampling. The downsampling block ensembles features from the large region into a small region with convolution operation is the procedure to reduce the size of the feature map. It uses the tiny encoder–decoder structure with a pixel-based enhancing module that efficiently highlights the features and improves segmentation accuracy. In contrast, the upsampling block enlarges the feature map, and this makes the feature blur. Operating the

tiny encoder–decoder structure with the pixel-based enhancing module on a blur feature map does not get better results.

Table 8. Pixel-based enhancing module at various locations in END-Net.

Locations	PA (%)	FWIoU (%)
DS-blocks	84.52	74.96
US-blocks	82.88	72.64
DS + *US*-blocks	82.77	73.01
NONE	82.62	72.23

Structures of variable map: We consider different structures of the variable map in the enhance module, including the variable map with the size of $n \times n$ and with the size of 1×1, and demonstrate the results in Table 9. In Table 9, the item "Size" indicates the size of the variable map in the enhance module, and $n \times n$ means the size of the variable map is the same as the size of the feature map. It can seem that a variable map with size $n \times n$ is powerful than using a single variable. The variables will adjust their value (enhancing factor) adaptively as the network iterates; therefore, the variable map with the size of $n \times n$ can obtain the appropriate enhancing factor for each pixel rather than use a single enhancing factor. Notice that we set the initial value of the variable map with various sizes to one. In Table 9, the performance of using variable map with size $n \times n$ is better than with size 1×1, and it is 0.45% and 0.60% higher than using size 1×1 in PA and FWIoU, respectively. An $n \times n$ feature map with only one enhancing factor makes each pixel of the feature map have the same weight. However, some discriminative characteristics, such as edges and corners, which can highlight the differences, should assign different weights (enhancing factor) to highlight the importance of features in feature learning. Therefore, our study designs an $n \times n$ variable map that can get better results.

Table 9. Different structures of variable map.

Size	PA (%)	FWIoU (%)
$n \times n$	84.52	74.96
1×1	84.07	74.36

Loss strategies: We analyze two-loss strategies, single-loss, and multi-loss strategies to determine the best loss strategy and demonstrate the results in Table 10. In a single-loss strategy, we keep the *loss7* and abandon the rest of the loss in our network. In multi-strategy, we reserve all the losses, which are designed in our network. In Table 10, the multi-loss strategy has 84.52% and 74.96% accuracies of PA and FWIoU, which is 0.09% and 0.54% higher than using single-loss strategy. Overall, the multi-loss approach can improve the performance of the proposed model and has better accuracy than using a single-loss strategy.

Table 10. Performance of the various strategies in ENDFNet.

Loss	PA (%)	FWIoU (%)
multi-loss	84.52	74.96
single-loss	84.43	74.42

6. Discussion

This subsection discusses the effect of various hyper-parameters settings, data expansion, external factors, and the advantages and limitations of the proposed model.

6.1. Hyper-Parameters

To realize the effect of the hyper-parameters on the proposed model, we justified the choice of all used hyper-parameters, including the image size and initial value of the learnable variable map (k). We discuss two image sizes, 224 × 224 and 512 × 512, and three initial values of learnable variable map (k = 1.0, 2.0, 3.0), and demonstrate the results in Table 11. The value of variable k is learned during the training process, but its initial value could make it obtain the local or global optimal values after training. Therefore, we examine three initial values, 1.0, 2.0, and 3.0, to define the best initial value. In Table 11, the initial value set as 1.0 can obtain the best accuracies in PA and FwIoU metrics; and the initial value set as 2.0 can obtain the best accuracies in MPA and MIoU metrics. Therefore, we set the initial value of k to be 1.0 in this study.

Moreover, we take different image sizes as the input with the same external factors (NVIDIA 1080 (11 G) graphics card under the Ubuntu16.04 system) to analyze the effect of the image size. We set the batch size from 6 to 3 for image size 512 × 512 to execute on the same external factors and demonstrate the results in Table 11. In Table 11, the performance of each metric becomes worse with image size 512 × 512 due to the different sizes of the receptive field. The images with various sizes have the same content, making their respective pixels cover the various size of the field. Therefore, it shows that the proposed model is more effective for the input size of 224 × 224.

Table 11. Effects of hyper-parameters.

Size	*k*	PA (%)	FwIoU (%)	MPA (%)	MIoU (%)	*Params* (M)	FPS
224 × 224	1.0	84.52	74.96	52.17	37.49	28.91	11.42
	2.0	84.41	74.90	54.44	38.23	28.91	11.61
	3.0	82.70	72.76	51.73	36.14	28.91	11.64
512 × 512	1.0	80.21	69.32	43.59	28.60	29.27	3.99

6.2. Data Expansion and External Factors

To study the effect of DATA expansion and processing on the model complexity and energy efficiency, we expand the data set by non-overlap cropping images with size 224 × 224 from the image with size 518 × 922, which is resized from the original image with size 4608 × 2592 data, due to the difficulty of collecting data. We generate 3313/1424 images after extension for training and testing. Moreover, we execute the model on the other device to discuss the effect of the external factors. The experiment results are shown in Table 12. The specification of Environment **1** is NVIDIA 1080 (11 G) graphics card under the Ubuntu16.04 system and that of device **2** is RTX6000 (24 G) graphics card under the Ubuntu20.04 system. The Environment **1** and **2** have the same experimental parameters, including input size being 224 and k = 1 in a learnable variable map. In Table 12, the model with an expanded dataset has lower accuracies in the matrices of PA, FwIoU, MPA, and MIoU due to different sizes of input data having different receptive fields for the same model. For example, the receptive field of the image with size 224 × 224, which is cropped from the image with size 4608 × 2592, is different from the receptive field of the image with size 224 × 224, which is resized from the image with size 4608 × 2592. The FPS is increased when addressing the expanded dataset because the cropped image's content becomes simpler than the original input image. More specifically, there are fewer tree species in each cropped image compared to the original input image. In addition, the model that executes with various external factors has the same results in each metric. It proves that the proposed model is portable and can be used on various platforms.

Table 12. Effects of data expansion and external factors.

	Environment	Data (Training/Testing)	PA (%)	FwIoU (%)	MPA (%)	MIoU (%)	*Params* (M)	FPS
Data expansion	1	414 / 178	84.52	74.96	52.17	37.49	28.91	11.42
	1	3313 / 1424	80.00	68.97	46.65	31.69	28.91	18.70
External factors	2	414 / 178	84.69	74.88	52.21	37.40	28.91	14.22

6.3. Advantages and Limitations

The END-Net has the best accuracies in PA, FWIoU, MPA, and MIoU metrics. The proposed modules and strategy play critical roles in improving the performance, proved on the diagnostic and ablation experiments. Moreover, it has fewer parameters and acceptable FPS. Therefore, END-Net can be used in reality. In practical application, the user can take a mobile phone to shoot the plant species from a tall building or operate a drone with an RGB camera to capture images at low latitudes to collect the images for analysis. Moreover, the user can execute the proposed trained model at a cloud device or a personal computer with a single graphic card for analysis.

In this study, we collect sixteen tree species images, but the number of images (pixels) for each tree species is not equal. Some dominant tree species have an amount of data, and few tree species have less, which makes the data unbalanced. We have to extend the data for the tree species that has less data and add other tree species in follow-up research.

6.4. Challenges of Plant Segmentation Using UAVs

There are several challenges in the plant species segmentation with aerial images: (1) the balance between the data and cost, and (2) the balance between accuracy and efficiency. In the balance between the data and cost, the commonly used resource for plant segmentation includes optical, multispectral, and hyperspectral images. The multispectral and hyperspectral images can provide images of dozens or hundreds of channels to generate abundant features for plant segmentation, but their cost is high due to high-resolution images [55]. The Landsat satellite can provide some resource format but with a low resolution compared to the spectral images captured by the UVAs [56]. Therefore, we use UAVs with optical cameras to capture high-resolution images to balance the cost and the resource resolution.

In the balance between accuracy and efficiency, the deep learning approaches using optical images have challenges in accuracy and efficiency. Most of the existing approaches focus on improving the segmentation accuracy [25] but have the issue of executing time due to the complex model [57]. The complex models have a slow executing time and lead to inefficient and non-real-time. The proposed network nests a tiny encoder–decoder module with the proposed pixel-based enhancing module, increasing the depth of the network, and complicated operations have the challenge of executing time but overcoming the issue of segmentation accuracy. The proposed framework does not have a fast-executing time, but the difference in executing time between our network and the compared methods is slight. Moreover, the proposed framework has the best accuracy. Therefore, the proposed method has a good balance between accuracy and efficiency.

7. Conclusions

This study proposes a plant species segmentation network with enhancing nested downsampling features (END-Net) for complex and challenging plant species segmentation tasks. END-Net takes the Unet as the backbone and contains three main contributions: (1) The tiny encoder–decoder structure with a pixel-based enhancing module embedded in a downsample block can efficiently highlight the features, improving the segmentation accuracy; (2) the pixel-based enhancing module assigns different weights (enhancing factor) to adaptively highlight the importance of each pixel's features in feature learning; and (3) the multi-loss strategy is a deep-supervision strategy; it calculates and accumulates the losses for the efficient adjustment of the network. Moreover, we collect the aerial and

optical images to construct the plant dataset, namely the OAPI dataset. To the best of our knowledge, the OAPI dataset is a rare database mainly focused on collecting optical images captured by a low-altitude drone.

In the experiments, we execute the diagnostic and ablation experiments to prove the significance of the proposed pixel-based module and demonstrate the effectiveness of our network. Moreover, we provide the quantitative results with six metrics to show the performance of the proposed END-Net and give the qualitative consequence to prove the feasibility of the END-Net with visualization segmentation outcomes.

In the future, we will improve the model to increase the segmentation accuracies of rare plant species with fewer data than the dominant species. More specifically, we will consider the classes' weights into the loss strategy that makes the network pay more attention to categories with small data [58]. In addition, we also consider using adversarial networks [59] or various loss strategies [60] to improve the accuracy of categories with small samples. Furthermore, we will keep collecting aerial and optical images, infrequent ones, and balance the number of categories in the dataset. More precisely, we will expand the number of rare categories, such as *Withered dicranopteris dichotoma*, *Adenanthera pavonina*, *Blechnum orientale*, and *Miscanthus sinensis*. Moreover, we will add some new sampling points to extend the number of tree species and images.

Author Contributions: Conceptualization, C.-W.L.; methodology, C.-W.L. and Y.H.; validation, C.-W.L. and M.L.; investigation, C.-W.L., M.L. and Y.H.; writing—original draft preparation, C.-W.L. and M.L.; writing—review and editing, C.-W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Postdoctoral Science Foundation under Grant 2018M632565, the Channel Postdoctoral Exchange Funding Scheme, the Youth Program of Humanities and Social Sciences Foundation, Ministry of Education of China under Grant 18YJCZH093, and the Natural Science Foundation of Fujian Province under Grant 2021J01128.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: The authors would like to thank Yanhe Huang and Jinshi Lin from the College of Resource and Environment, Fujian Agriculture and Forestry University for the assistance with the collection of the dataset.

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

Abbreviations

The following abbreviations are used in this manuscript:

END-Net	The convolutional neural network using enhancing nested downsampling features
OAPI dataset	Optical aerial plant image dataset
PA	Pixel Accuracy
FWIoU	Frequency Weighted Intersection over Union
CNN	Convolutional Neural Network
UAV	Unmanned Aerial Vehicle
DS_{ED}-block	The downsampling block with encoder–decoder structure
Conv. block	Convolution block
US-block	The upsampling block

References

1. Folharini, S.D.O.; Melo, S.N.D.; Cameron, S.R. Effect of protected areas on forest crimes in Brazil. *J. Environ. Plan. Manag.* **2021**, *65*, 1–15.
2. Zhang, M.J.; Dong, R.; Wang, X.X. Plants with health risks undermine residents' perceived health status, evaluations and expectations of residential greenery. *Landsc. Urban Plan.* **2021**, *216*, 104236. [CrossRef]

3. Li, J.; Wang, J.; Zhang, J.; Zhang, J.; Kong, H. Dynamic changes of vegetation coverage in China-Myanmar economic corridor over the past 20 years. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102378. [CrossRef]
4. Thanh, D.N.; Quénot, G.; Goeuriot, L. *Non-Local DenseNet for Plant CLEF 2019 Contest*; CLEF (Working Notes); 2019; pp. 655–667. Available online: https://www.semanticscholar.org/paper/Non-local-DenseNet-for-Plant-CLEF-2019-Contest-Thanh-Qu%C3%A9not/5baa80aaf80ec89038f12f41bee8d2e86285e8db (accessed on 23 November 2021).
5. Choe, H.; Chi, J.; Thorne, J.H. Mapping Potential Plant Species Richness over Large Areas with Deep Learning, MODIS, and Species Distribution Models. *Remote Sens.* **2021**, *13*, 2490. [CrossRef]
6. Xi, J.; Shao, Y.; Li, Z.; Zhao, P.; Ye, Y.; Li, W.; Chen, Y.; Yuan, Z. Distribution of woody plant species among different disturbance regimes of forests in a temperate deciduous broad-leaved forest. *Front. Plant Sci.* **2021**, *12*, 618524. [CrossRef]
7. Aakif, A.; Khan, M.F. Automatic classification of plants based on their leaves. *Biosyst. Eng.* **2015**, *139*, 66–75. [CrossRef]
8. Kaur, S.; Pandey, S.; Goel, S. Plants disease identification and classification through leaf images: A survey. *Arch. Comput. Methods Eng.* **2019**, *26*, 507–530. [CrossRef]
9. Kebapci, H.; Yanikoglu, B.; Unal, G. Plant image retrieval using color, shape and texture features. *Comput. J.* **2011**, *54*, 1475–1490. [CrossRef]
10. Hsiao, J.K.; Kang, L.W.; Chang, C.L.; Lin, C.Y. Comparative study of leaf image recognition with a novel learning-based approach. In Proceedings of the 2014 Science and Information Conference, London, UK, 27–29 August 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 389–393.
11. Yang, L.W.; Wang, X.F. Leaf image recognition using fourier transform based on ordered sequence. In Proceedings of the International Conference on Intelligent Computing, Huangshan, China, 25–29 July 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 393–400.
12. Moore, M.M.; Bauer, M.E. Classification of forest vegetation in north-central Minnesota using Landsat Multispectral Scanner and Thematic Mapper data. *For. Sci.* **1990**, *36*, 330–342.
13. Carleer, A.; Wolff, E. Exploitation of very high resolution satellite data for tree species identification. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 135–140. [CrossRef]
14. Holmgren, J.; Persson, Å. Identifying species of individual trees using airborne laser scanner. *Remote Sens. Environ.* **2004**, *90*, 415–423. [CrossRef]
15. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
16. Immitzer, M.; Atzberger, C.; Koukal, T. Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data. *Remote Sens.* **2012**, *4*, 2661–2693. [CrossRef]
17. Colgan, M.S.; Baldeck, C.A.; Féret, J.B.; Asner, G.P. Mapping savanna tree species at ecosystem scales using support vector machine classification and BRDF correction on airborne hyperspectral and LiDAR data. *Remote Sens.* **2012**, *4*, 3462–3480. [CrossRef]
18. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking BiSeNet For Real-time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 9716–9725.
19. Sünderhauf, N.; McCool, C.; Upcroft, B.; Perez, T. *Fine-Grained Plant Classification Using Convolutional Neural Networks for Feature Extraction*; CLEF (Working Notes); 2014; pp. 756–762. Available online: http://ceur-ws.org/Vol-1180/CLEF2014wn-Life-SunderhaufEt2014.pdf (accessed on 23 November 2021).
20. Längkvist, M.; Kiselev, A.; Alirezaie, M.; Loutfi, A. Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sens.* **2016**, *8*, 329. [CrossRef]
21. Weinstein, B.G.; Marconi, S.; Bohlman, S.; Zare, A.; White, E. Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks. *Remote Sens.* **2019**, *11*, 1309. [CrossRef]
22. Kumar, J.P.; Domnic, S. Rosette plant segmentation with leaf count using orthogonal transform and deep convolutional neural network. *Mach. Vis. Appl.* **2020**, *31*, 1–14.
23. Fuentes-Pacheco, J.; Torres-Olivares, J.; Roman-Rangel, E.; Cervantes, S.; Juarez-Lopez, P.; Hermosillo-Valadez, J.; Rendón-Mancha, J.M. Fig plant segmentation from aerial images using a deep convolutional encoder–decoder network. *Remote Sens.* **2019**, *11*, 1157. [CrossRef]
24. Xu, L.; Li, Y.; Xu, J.; Guo, L. Two-level attention and score consistency network for plant segmentation. *Comput. Electron. Agric.* **2020**, *170*, 105281. [CrossRef]
25. Zou, K.; Chen, X.; Zhang, F.; Zhou, H.; Zhang, C. A Field Weed Density Evaluation Method Based on UAV Imaging and Modified U-Net. *Remote Sens.* **2021**, *13*, 310. [CrossRef]
26. Zhang, X.; Yang, Y.; Li, Z.; Ning, X.; Qin, Y.; Cai, W. An improved encoder–decoder network based on strip pool method applied to segmentation of farmland vacancy field. *Entropy* **2021**, *23*, 435. [CrossRef]
27. Kolhar, S.; Jagtap, J. Convolutional neural network based encoder–decoder architectures for semantic segmentation of plants. *Ecol. Inform.* **2021**, *64*, 101373. [CrossRef]
28. Mikula, K.; Šibíková, M.; Ambroz, M.; Kollár, M.; Ožvat, A.A.; Urbán, J.; Jarolímek, I.; Šibík, J. NaturaSat—A Software Tool for Identification, Monitoring and Evaluation of Habitats by Remote Sensing Techniques. *Remote Sens.* **2021**, *13*, 3381. [CrossRef]

29. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. OCNet: Object Context for Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, 129, 2375–2398.
30. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 173–190.
31. 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.
32. 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; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
33. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
34. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
35. Jégou, S.; Drozdzal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.
36. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 603–612.
37. Mou, L.; Zhu, X.X. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6699–6711. [CrossRef]
38. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic segmentation of aerial images with an ensemble of CNSS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480. [CrossRef]
39. Ayoul, T.; Buckley, T.; Crevier, F. *Uav Navigation above Roads Using Convolutional Neural Networks*; Technical Report; Stanford University: Stanford, CA, USA, 2017.
40. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder–decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
41. Mou, L.; Hua, Y.; Zhu, X.X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12416–12425.
42. Benjdira, B.; Bazi, Y.; Koubaa, A.; Ouni, K. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sens.* **2019**, *11*, 1369. [CrossRef]
43. Zhang, X.; Jin, J.; Lan, Z.; Li, C.; Fan, M.; Wang, Y.; Yu, X.; Zhang, Y. ICENET: A Semantic Segmentation Deep Network for River Ice by Fusing Positional and Channel-Wise Attentive Features. *Remote Sens.* **2020**, *12*, 221. [CrossRef]
44. Valada, A.; Vertens, J.; Dhall, A.; Burgard, W. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 4644–4651.
45. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
46. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
47. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
48. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
49. Fu, J.; Liu, J.; Wang, Y.; Zhou, J.; Wang, C.; Lu, H. Stacked deconvolutional network for semantic segmentation. *IEEE Trans. Image Process.* **2019**. [CrossRef]
50. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.
51. Pohlen, T.; Hermans, A.; Mathias, M.; Leibe, B. Full-resolution residual networks for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4151–4160.
52. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
53. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
54. Zhang, H.; Zhang, H.; Wang, C.; Xie, J. Co-occurrent features in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 548–557.

55. Khanal, S.; KC, K.; Fulton, J.P.; Shearer, S.; Ozkan, E. Remote sensing in agriculture—Accomplishments, limitations, and opportunities. *Remote Sens.* **2020**, *12*, 3783. [CrossRef]
56. Nguyen, M.-T.; Shah, D. *Improving Current Limitations of Deep Learning BASED Plant Disease Identification*; The Cooper Union: New York, NY, USA, 2019.
57. Lan, Y.; Huang, K.; Yang, C.; Lei, L.; Ye, J.; Zhang, J.; Zeng, W.; Zhang, Y.; Deng, J. Real-Time Identification of Rice Weeds by UAV Low-Altitude Remote Sensing Based on Improved Semantic Segmentation Model. *Remote Sens.* **2021**, *13*, 4370. [CrossRef]
58. Sugino, T.; Kawase, T.; Onogi, S.; Kin, T.; Saito, N.; Nakajima, Y. Loss weightings for improving imbalanced brain structure segmentation using fully convolutional networks. *Healthcare* **2021**, *9*, 938. [CrossRef] [PubMed]
59. Suh, S.; Lee, H.; Lukowicz, P.; Lee, Y.O. CEGAN: Classification Enhancement Generative Adversarial Networks for unraveling data imbalance problems. *Neural Netw.* **2021**, *133*, 69–86. [CrossRef]
60. Li, C.; Chen, M.; Zhang, J.; Liu, H. Cardiac MRI segmentation with focal loss constrained deep residual networks. *Phys. Med. Biol.* **2021**, *66*, 135012. [CrossRef] [PubMed]