remote sensing

MDPI

Article

Weed–Crop Segmentation in Drone Images with a Novel Encoder–Decoder Framework Enhanced via Attention Modules

Sultan Daud Khan [1,*], Saleh Basalamah [2] and Ahmed Lbath [3]

[1] Department of Computer Science, National University of Technology, Islamabad 44000, Pakistan
[2] Department of Computer Engineering, Umm Al-Qura University, Mecca 24382, Saudi Arabia; smbasalamah@uqu.edu.sa
[3] Department of Computer Science, Université Grenoble Alpes, 38400 Grenoble, France; ahmed.lbath@imag.fr
* Correspondence: sultandaud@nutech.edu.pk

Abstract: The rapid expansion of the world's population has resulted in an increased demand for agricultural products which necessitates the need to improve crop yields. To enhance crop yields, it is imperative to control weeds. Traditionally, weed control predominantly relied on the use of herbicides; however, the indiscriminate application of herbicides presents potential hazards to both crop health and productivity. Fortunately, the advent of cutting-edge technologies such as unmanned vehicle technology (UAVs) and computer vision has provided automated and efficient solutions for weed control. These approaches leverage drone images to detect and identify weeds with a certain level of accuracy. Nevertheless, the identification of weeds in drone images poses significant challenges attributed to factors like occlusion, variations in color and texture, and disparities in scale. The utilization of traditional image processing techniques and deep learning approaches, which are commonly employed in existing methods, presents difficulties in extracting features and addressing scale variations. In order to address these challenges, an innovative deep learning framework is introduced which is designed to classify every pixel in a drone image into categories such as weed, crop, and others. In general, our proposed network adopts an encoder–decoder structure. The encoder component of the network effectively combines the Dense-inception network with the Atrous spatial pyramid pooling module, enabling the extraction of multi-scale features and capturing local and global contextual information seamlessly. The decoder component of the network incorporates deconvolution layers and attention units, namely, channel and spatial attention units (CnSAUs), which contribute to the restoration of spatial information and enhance the precise localization of weeds and crops in the images. The performance of the proposed framework is assessed using a publicly available benchmark dataset known for its complexity. The effectiveness of the proposed framework is demonstrated via comprehensive experiments, showcasing its superiority by achieving a 0.81 mean Intersection over Union (mIoU) on the challenging dataset.

Keywords: smart agriculture; deep learning; semantic segmentation; remote sensing

Citation: Khan, S.D.; Basalamah, S.; Lbath, A. Weed–Crop Segmentation in Drone Images with a Novel Encoder–Decoder Framework Enhanced via Attention Modules. *Remote Sens.* **2023**, *15*, 5615. https://doi.org/10.3390/rs15235615

Academic Editor: Eufemia Tarantino

Received: 2 October 2023
Revised: 22 November 2023
Accepted: 28 November 2023
Published: 3 December 2023

1. Introduction

The population of the world is rapidly increasing and it is expected that the world population will rise to 9.1 billion approximately by 2050 [1]. This rapid increase in population will increase the demand of agriculture products. To meet the growing demands of agriculture products, it is important to enhance the crop yields. Higher yields ensure an adequate and consistent food supply, support economic growth, and address the needs of a growing population.

To increase the crop yield, it is essential to reduce and control the growth of weeds. Weeds are unwanted plants that compete with crops for essential resources such as nutrients, water, and sunlight [2]. Weeds hinder crop growth, reduce productivity, and even cause crop failure if left uncontrolled. A recent study [3] investigates the impact of soil solarization

methods on microbial count and weed growth. However, among existing methods, the use of herbicides is generally used to control the growth of weeds [4]. However, the application of herbicides carries the risk of harming the crop and potentially resulting in decreased productivity. The main reason of this issue stems from the inappropriate approach of indiscriminately spraying herbicides across the entire field [5], including areas that are already free of weeds [6]. In order to tackle this problem, it is crucial to leverage advanced technology to develop innovative approaches and methods for weed control.

Due to advancements in unmanned vehicle technology (UAVs) [7], the acquisition of numerous high-resolution images has become incredibly convenient. Consequently, computer vision technology, which involves the analysis of images and videos to identify various objects, has gained significant importance within the realm of smart agriculture [8,9]. It finds extensive application in tasks such as classifying plant diseases [10], estimating crop yields [11,12], and monitoring crop growth [13,14].

The task of weed detection in drone images is challenging due to several factors. These factors include occlusion (where weeds may be hidden by other objects), variations in color and texture caused by lighting conditions, similarities in appearance between weeds and crops, as well as differences within the same weed species. Furthermore, weed–crop segmentation in drone images suffers from a multi-scale problem, since drone images are captured from a far distant camera with varying angles and contain vegetation patterns of varying sizes, ranging from small weeds to larger crops.

Most of the existing computer vision-based weed detection methods primarily consist of traditional image processing methods and deep learning approaches. Traditional image processing methods [15–17] extract hand-crafted features including shape, texture, and color from the image and then employ statistical machine learning models, for example, Support Vector Machine (SVM), Random Forest, etc., for weed detection. Although traditional image processing methods have limited requirements for data and computation power, however, these methods are not well suited for weed detection in wild or uncontrolled environments. This is due to the reason that these methods have limited feature extraction capabilities, cannot capture wide variations in objects' scales, and are prone to illumination changes.

Deep learning models, in contrast to traditional models, are data driven and require huge amounts of data for training. These models automatically extract hierarchical features and learn discriminating features from images and have achieved remarkable success in a wide range of tasks, including classification, detection, and semantic segmentation. Recognizing this success, deep learning models are also being utilized for the weed–crop segmentation task [18], where the models classify each pixel of the input image into a weed or crop class.

Considering the weed–crop segmentation task, current deep learning-based semantic segmentation models for weed–crop detection [19–24] face challenges in achieving the desired results. This is due to the reason that the majority of current deep learning-based weed–crop segmentation models rely on Fully Convolutional Neural Networks (FCNs) [25] and U-Net [26], which inherit the following limitations: (1) These models suffer from the gradient vanishing problem, especially in the cases when the number of layers increase. In this situation, these networks lose spatial information which is crucial for accurate segmentation. (2) These networks have limited receptive fields, which can result in difficulties capturing context and integrating global information into the segmentation process. (3) These networks have fixed or limited scale range and face challenges in precisely segmenting the weed and crops in drone images.

Considering the aforementioned challenges in the weed–crop segmentation task, a framework is proposed that addresses the limitations of the FCN- [25] and U-Net [26]-based networks. Generally, the proposed framework consists of two main parts: (1) Encoder and (2) Decoder. The encoder part of the network consists of a Dense-inception network followed by Artous spatial pyramid pooling. The encoder part enables the framework to allow for dense connections among the subsequent layers and capture multi-scale

information in order to effectively incorporate global contextual information. The decoder part of the framework consists of deconvolution layers and attention units (channel and spatial attention units, denoted as, CnSAU). The deconvolution layers recover the spatial information lost during the encoding stage, while the attention units boost the localization performance of the framework by focusing on the important channel and spatial location and suppressing other less important details.

In comparison to other deep learning models utilized for weed–crop segmentation in drone images, the proposed framework offers the following contributions:

1. A novel framework is introduced to automatically distinguish between weeds and crops in high-resolution drone images.
2. The framework incorporates a Dense-inception network within its encoder component, which seamlessly integrates with the Artous spatial pyramid pooling module. This integration enables the extraction of multi-scale features and facilitates the capture of both local and global contextual information.
3. The decoder part of the framework effectively integrates deconvolution and attention units to recover spatial information and boost the localization of weeds and crops in drone images.
4. The performance of the proposed framework is evaluated on a publicly available benchmark dataset. From the experiments' results, it is demonstrated that the proposed framework precisely identifies weeds and crops in complex drone images and consequently beats other state-of-the-art methods in terms of performance standards.

The remaining sections of this paper are structured as follows: Section 2 discusses the related works, Section 3 presents the proposed framework, Section 4 provides a description and discussion of the experimental results, and finally, Section 5 concludes the paper and presents future perspectives.

2. Related Work

In recent years, there has been a growing interest among researchers in the field of automatic weed–crop segmentation from drone images. This area of study has gained significant attention due to its potential applications in agriculture and the need for efficient weed management techniques. Researchers have developed several models to tackle this problem, which can be broadly categorized into two main categories: traditional image processing models and deep learning models.

2.1. Traditional Image Processing Models

Traditional image processing models refer to approaches which rely on classical image processing techniques to identify the crops and weeds in drone images. This class of models typically extract hand-crafted features, including texture, color, shape, and other visual appearance features, to differentiate between the weeds and crops. For example, Zheng et al. [27] uses the color feature to differentiate between maize crop and weeds. The method extracts nine discriminating color feature vectors and train a support vector machine to segment maize and weeds in images. Similarly, Burks et al. [28] employed the Color Co-occurrence Method (CCM) to characterize distinct weed species. Hamuda et al. [29] introduced a novel algorithm that combines color features with morphological erosion and dilation techniques to identify cauliflower crop areas in images and distinguish them from weeds. Jafari et al. [30] presented a model that employs the RGB color components of an image to detect sugar beet plants and distinguish them from seven common types of sugar beet weeds under varying normal lighting conditions. Some studies utilize the shape feature to identify weeds. Woebbecke et al. [31] extracted the shape features and subsequent analysis was performed on binary images to classify and differentiate between various weed classes. Similarly, three different active shape models were utilized in [32] for crop and weed classification. Perez et al. [33] proposed a technique for weed and crop classification. The method utilizes color information to differentiate between the vegetation and the background while also employing shape analysis techniques to distinguish

between crops and weeds. Similarly, Lin et al. [34] proposed a method that utilizes shape and texture features and employs a decision tree to classify corn and weed species. Neto et al. [35] proposed a method that utilizes Elliptic Fourier (EF) and discriminant analyses to characterize different soybean species. The model generates Elliptic Fourier harmonic functions based on the leaf boundary and utilizes principal component analysis to identify the Fourier coefficients with the greatest discriminatory power, which ultimately helps in identifying soybean species. Chaisattapagon et al. [36] utilized color analysis, shape analysis, and texture analysis to effectively identify weeds within wheat fields. Meyer et al. [37] proposed a method that extracts four different texture features, namely, angular second moment, inertia, entropy, and local homogeneity, to identify weed regions. Bakhshipour et al. [38] proposed a method that extract 52 texture features from the image. The method employs the principal component analysis technique to extract 14 discriminating features and train a neural network that assigns labels to each pixel as a weed or crop. Wu et al. [39] generates a histogram from gray level images and extracts texture feature by utilizing the gray level Co-occurrence Matrix (GLCM). These features are then utilized via an SVM classifier that recognizes weeds and corns. Ishak et al. [40] proposed a technique that utilizes a blend of Gabor wavelet (GW) and gradient field distribution (GFD) methods to derive a novel set of feature vectors to categorize different types of weeds. Ahmed et al. [41] proposed a method for classifying weed images based on texture. This method utilizes a local binary pattern (LBP) and combines template matching with a support vector machine to effectively categorize the weed images.

The aforementioned traditional image processing methods offer advantages such as requiring a small sample size and taking shorter training times. Furthermore, these models do not require powerful graphics processing units. These models have the following limitations: (1) Traditional methods often rely on handcrafted features, including color, texture, and shape, which may not capture the full complexity and richness of the image data. (2) These methods are not robust and are usually sensitive to changes in lighting conditions, image noise, and variations in weed appearance. (3) These methods are difficult to generalize on different un-seen images and may struggle to adapt to new and unseen weed species or variations in plant growth stages.

2.2. Deep Learning Models

Deep learning models follow a unique structure and possess exceptional capability to learn rich hierarchical and discriminating features from the raw images. With the success of deep learning models across various domains, they have also been extended to precision agriculture, particularly for a task such as weed–crop segmentation. Yu et al. [42] utilized the standard VGG16 deep learning network for weed detection. You et al. [43] proposed a deep learning model for weed–crop segmentation that incorporates a Hybrid-dilated convolution and DropBlock into the backbone network. The Hybrid-dilated convolution is used to enhance the receptive field to increase the contextual awareness of the network. DropBlock regularize the weights by omitting the contiguous regions in the image. Jin et al. [44] proposed a novel method that combines the benefits of deep learning models and image processing technology for detecting weeds in an image. The method first trains the CenterNet [45] to detect vegetables by drawing bounding boxes around them. The green objects out of the boundaries of the bounding box are considered as weeds. Razfar [46] presented a comparative study to evaluate the effectiveness of various lightweight deep learning models, MobileNetV2 [47] and ResNet50 [48], for detecting weeds in soybean crops. Dos et al. [49] proposed a deep learning model for weed detection in soybean crops and also characterized weeds into different categories. Sodjinou et al. [50] presented a method that combines the deep learning-based segmentation model, U-Net, and the K-means clustering algorithm for the segmentation of crops and weeds. Osorio et al. [51] presented three different methods for estimating the presence of weed in lettuce crops. The first method employs support vector machines (SVM) with histograms of oriented gradients (HOG) as a feature descriptor. The second method utilizes YOLOV3 [52] and Mask-RCNN [53] as a

third method. To enhance the performance of these methods, the NDVI index (normalized difference vegetation index) [54] is used as a background subtractor that removes non-photosynthetic objects from the images. Kim et al. [55] introduces a novel approach, a multi-task semantic segmentation-convolutional neural network, which enables the simultaneous detection of crops and weeds. Fawakherji et al. [56] proposed a sequence of two deep learning networks for weed and crop classification in RGB images. The network follows the encoder–decoder segmentation architecture, which classifies each pixel into vegetation and soil classes. Nasiri et al. [57] employs encoder–decoder architecture based on U-Net, incorporating ResNet50 as a feature extractor, to perform weed recognition in sugar beet fields. Khan et al. [58] introduced a cascade framework called CED-Net for the segmentation of crops and weeds. The framework utilizes a sequence of small networks to achieve initial coarse-to-fine results, which are subsequently combined to obtain more refined segmentation outcomes. Xu et al. [59] proposed a novel method for weed detection in soybean crops, combining a visible color index with instance segmentation using an encoder–decoder architecture. Janneh et al. [60] proposed an improved deep convolutional neural network (DCNN) combined with a lightweight backbone, a multi-level feature re-weighted fusion module, and a convolutional weighted fusion decoder for efficient weed control in agriculture. Zhang et al. [61] introduced an improved Swin-Unet model for weed recognition in maize fields. The model addresses challenges such as color similarity and overlapping between weeds and maize, as well as varying lighting and weather conditions. Weyler et al. [62] proposed a semantic segmentation approach that aims to enhance the generalization capability of convolutional neural networks (CNNs) for effective crop and weed detection. Bertoglio et al. [63] compares two domain adaptation techniques, Fourier Transform and CycleGAN architecture, for weed segmentation. Apart from the aforementioned cutting-edge approaches, numerous survey papers [64–66] have been published that delve into the topic of crop and weed segmentation in images.

3. Proposed Methodology

This section presents the methodology proposed for weed–crop segmentation in drone images. The proposed framework follows an encoder–decoder pipeline. The encoder part of the framework consists of a dense-inception network followed by the Artous spatial pyramid pooling module. The encoder part downsamples the input feature maps after passing through subsequent transition layers and extracts multi-scale features as well as local and global context. The decoder part of the network incorporates attention modules and deconvolution layers that upsample the feature map and capture spatial information. The architecture of the proposed framework is depicted in Figure 1, providing an overview of the overall structure. The specific details and functionalities of each component are as follows:

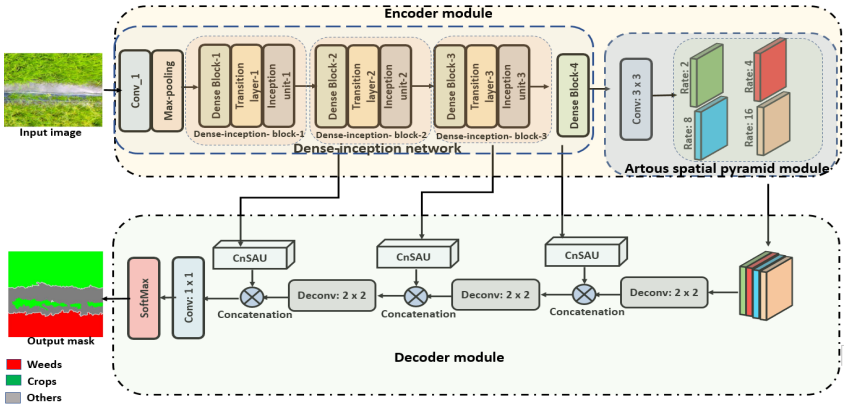

Figure 1. Detailed architecture of proposed framework for weed–crop segmentation.

3.1. Encoder Module

Considering the limitation of existing semantic segmentation networks, DenseNet is employed in our proposed framework. DenseNet, due to its unique structure, has achieved tremendous success in various image classification and object recognition tasks. However, the network faces challenges in precisely segmenting the weed and crops in drone images due to a fixed and limited scale range. To effectively address the multi-scale problem in weed–crop segmentation, it is crucial to address the multi-scale problem and process the images at different scales. For this purpose, instead of employing DenseNet alone, an inception module is embedded within the DenseNet to capture multi-scale features. Inception units perform parallel convolutions with different filter sizes and concatenate the resulting feature maps to capture information at multiple scales.

In addition to multi-scale features, it is also important to capture local and global context to enhance contextual understanding and further improve the accuracy. For this purpose, the Artous spatial pyramid pooling module is integrated that applies Atrous convolutions at multiple dilation rates, allowing the network to gather context from different spatial scales. The structure of a single Dense-inception block is illustrated in Table 1.

Table 1. Structural parameters of single Dense-inception block.

Layer Type	Filter Size
Dense block	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix}$
Transition block	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 2 \times 2 \text{ pooling, stride 2} \end{bmatrix}$
Inception module	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \\ 5 \times 5 \text{ conv} \\ 3 \times 3 \text{ Max pool} \end{bmatrix}$

From Table 1, it is shown that the Dense-inception unit consists of the Dense-block, Transition block, and inception module. The Dense-block contains a stack of two convolutional operations with the filter size 1×1 and 3×3. The transition block consists of one 1×1 convolution layer followed by a max pooling layer with the filter size 2×2. The inception module concurrently applies three convolution operations of different filter sizes, 1×1, 3×3, and 5×5. The inception module concatenates the outputs of all branches, including the pooling branch, to generate a fused feature map that contains multi-scale information from different receptive fields.

The proposed encoder module consists of three Dense-inception blocks followed by the Artous spatial pyramid pooling module. The encoder takes an input image of arbitrary size and applies a convolutional layer of size 7×7 of stride 2. With a max pooling layer with the pool size 3×3, stride 2 is then applied to the output tensor from the first convolutional layer. The output tensor from the max pooling layer is then passed through four Dense Blocks, with each consisting of a stack of two convolutional filters of size 1×1 and 3×3. The output tensor from each individual Dense Block is then passed through a Transition Layer. The transition layers consist of a batch normalization layer, a ReLU activation layer, a 1×1 convolution layer, and a 2×2 average pooling layer. The purpose of these layers is to reduce the spatial dimensions of the feature maps and reduce the number of parameters in the model. After the Transition layer, the tensor is then passed through inception units. The inception units use parallel convolutional layers with filter sizes (1×1, 3×3, and 5×5) to capture features at multiple scales. The convolutional filter of size 1×1 is used for dimensionality reduction, which helps to reduce the computational cost of the subsequent convolutional layers. The convolutional layer of size 3×3 captures small scale features, while the 5×5 filters capture features at a larger scale. Additionally, a max pooling operation is often applied to capture features across different regions of

the input image. The output feature maps from all parallel convolutional layers are then concatenated along the channel dimension and passed to the next dense block. The feature map of the 4th Dense Block is then provided as an input to the Artous spatial pyramid pooling module, which applies Atrous convolutions at multiple dilation rates (2, 4, 8, and 16) to gather local and global context from different spatial scales.

This unique combination of different networks enhances the model's ability to extract multi-scale features and enables the network to understand both fine-grained local details and broader contextual information, which leads to more effective segmentation of weed–crops in drone images.

3.2. Decoder

The proposed encoder module is designed to capture multi-scale features and integrate local and global context, which are essential for semantic segmentation tasks. Drone images contain fine-grained details about the weeds and crops patterns, as preserving their spatial distribution is crucial for accurate segmentation. However, the downsampling process in the encoder results in the loss of crucial spatial information. To recover the lost spatial details during the downsampling operations in the encoder part, the decoder part contains subsequent deconvolution layers which upsample the input feature maps by a factor of 2 and recover the spatial information. By recovering the spatial information, the decoder improves the network's ability to accurately localize the weed and crop regions.

In order to boost the network's ability to localize features, attention units are integrated into the decoder module of the proposed network. The attention module, illustrated in Figure 2, consists of two units, namely the channel attention and spatial attention unit, which is abbreviated as CnSAU. The channel attention unit enables the network to dynamically allocate attention to different channels by computing channel-wise attention weights, boosting the performance. The spatial attention unit enables the network to focus on more important spatial regions of the input feature maps while suppressing irrelevant regions. Considering the problem of segmentation in the weed–crop segmentation task, the channel attention and spatial attention unit are employed in parallel as employed in [67], which offers the following advantages over employing these attention units in series as employed in CBAM [68]: (1) By utilizing the spatial and channel attention units in parallel fashion, the network can learn both spatial and channel-wise dependencies simultaneously, which results in more comprehensive and discriminating features. (2) Utilizing attention units in parallel results in reducing the overall computation cost.

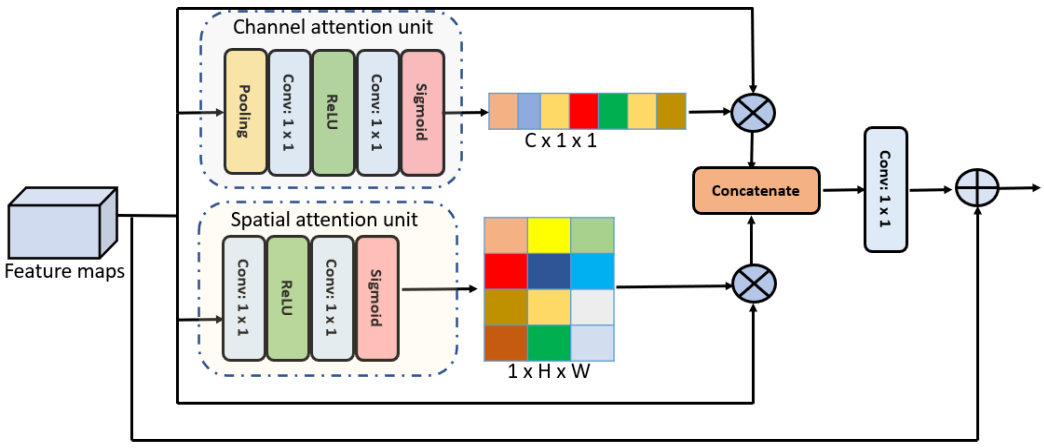

Figure 2. Structure of channel and spatial attention module.

The channel attention unit applies a global average pooling operation that squeezes the input feature map of size C × H × W to size C × 1 × 1. The units then apply a convolution operation followed by the ReLU operation to decrease the number of channels.

In the next, a second convolution operation is applied to increase the number of channels to the original count. The unit then applies element-wise product operation, represented by $\otimes$, on the original input feature map and output of the channel attention unit to generate the resultant feature map α_c. Similarly, the spatial attention unit employs a convolution operation to generate per-channel attention maps. These maps are then combined into a single map by employing the second convolution operation followed by the element-wise product with the original input feature map. Let α_s be the spatial attention map generated via the spatial attention unit. To generate α_r, α_c and α_s are concatenated followed by a 1×1 convolution operation. The final feature map is obtained by performing an element-wise summation of the original input feature map and α_r.

Within the decoder module, the input feature maps are sourced from the 4th Dense Block of the encoder module and a 2×2 deconvolutional operation is performed on these feature maps to recover the lost spatial details by effectively doubling the sizes of feature maps. Let Ω_i be the feature map obtained after performing a 2×2 deconvolution operation. After the upsampling operation in the decoder, the feature maps of the 3rd Dense Block are passed through the attention module. The channel attention module processes the feature maps to assign attention weights to each channel, emphasizing important information. The spatial attention module captures spatial dependencies and generates attention maps to highlight informative spatial regions. The outputs of both attention modules are multiplied element-wise to enhance relevant channels and informative spatial regions simultaneously. Let Ω_j be the resultant feature map. Feature maps Ω_i and Ω_j are concatenated to obtain feature maps Ω_k, followed by a 2×2 deconvolution operation on feature maps Ω_k to recover the spatial information and concatenate with feature maps of the $2nd$ Dense Block after passing through the attention unit. The decoder module repeats this procedure four times and subsequently applies a 1×1 convolutional layer followed by a log sigmoid activation function to the final feature map. By employing the log sigmoid function, the feature map is transformed into a segmentation mask, where the green pixel indicates the presence of crops, while the red pixel represents the weeds.

With the successful integration of attention units in the decoder part, our network can selectively emphasize the most relevant information, suppress noise, and capture fine-grained details, leading to more precise and accurate segmentation results. Additionally, the deconvolution layers upsample the feature maps and restore the spatial resolution that was lost during the downsampling operation in the encoder module. This enables the decoder to reconstruct a high-resolution output and capture fine-grained details that are necessary for accurate semantic segmentation.

4. Results and Discussion

In this section, the performance of the proposed framework is evaluated and compared with other state-of-the-art methods, both quantitatively and qualitatively.

The PyTorch library is employed to implement the proposed framework. For training the network, images are utilized from the training set along with their respective segmentation masks. A stochastic gradient descent is applied to optimize the network parameters, employing a consistent learning rate of 0.01.

4.1. Dataset

In this study, the performance of the proposed framework is assessed by using a publicly available dataset called the rice-weed dataset. The dataset was curated by Huang et al. [23] and was collected at the Zengcheng Teaching and Research Bases, South China Agricultural University. The data acquisition was carried out using a DJI GS PRO unmanned aerial vehicle (UAV) flying at an altitude of 6 m. The UAV's camera features a 1-inch CMOS sensor capable of capturing 12-megapixel images.

The dataset consists of 91 images, each with a pixel resolution of 4000×3000 pixels and a spatial resolution of approximately 0.3 cm. These images exhibit diverse scenes with varying densities of weeds and rice plants. Within each image, there are three distinct

categories: rice, weeds, and other elements. The rice pixels are labeled as green, the weed pixels as red, and the background pixels as gray.

Given the high resolution of images and due to a limited GPU memory, each image is sorted into twelve tiles, where the size of each tile is 1000×1000 pixels. This approach is also applied to the ground truth segmentation masks, ensuring consistency in the methodology. In this way, a total of 1092 tiles are generated, where 892 tiles are randomly selected for training and the remaining 200 are reserved for the testing phase. The number of images in the original dataset is limited; however, to address this problem, data augmentation techniques are employed by rotating, scaling, and flipping the original images. Specifically, these variations are incorporated by rotating images at 90- and 180-degree angles, adjusting their scales by enlarging or reducing the sizes of the images, as well as flipping the images horizontally and vertically. Through these diverse transformations, the size of the dataset is expanded, ensuring a more extensive and representative collection of training data for the network.

During the testing phase, a similar approach is followed where the test image is divided into 12 tiles. Each tile is individually inputted into the model for prediction. After obtaining the masks for all 12 tiles, they are stitched together to form the final segmentation mask. This final mask is then adjusted to align with the size of the original input image. A few sample frames and their corresponding ground truth masks are illustrated in Figure 3.

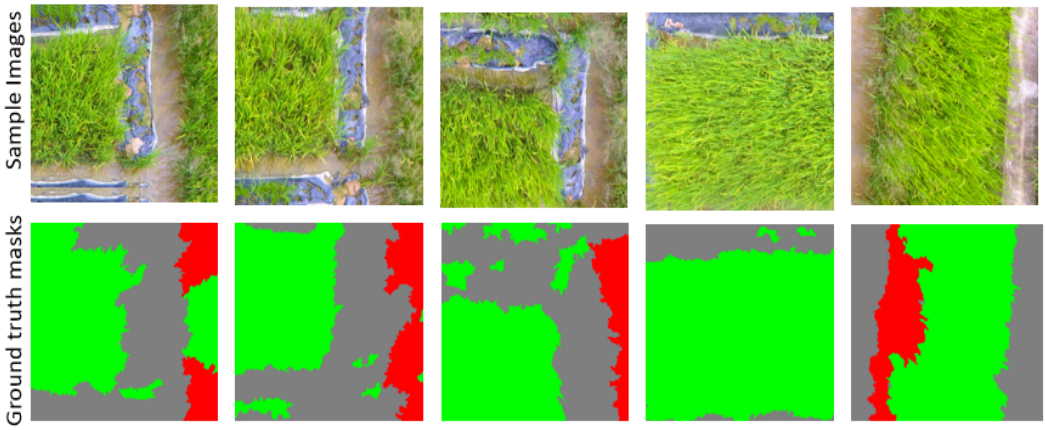

Figure 3. Illustrated sample frames randomly selected from the dataset and their corresponding ground truth masks. The red pixels in the ground truth mask represent the weed, while the green pixels represent rice crops and the gray pixels represent the others.

4.2. Evaluation Metrics

To evaluate the effectiveness of the proposed framework and compare it to state-of-the-art methods, the Jaccard similarity index, also known as Intersection over Union (IoU), is employed. The Jaccard similarity index serves as a widely adopted evaluation metric for assessing the performance of segmentation models. It is defined by the formula $\frac{\Omega_g \cap \Omega_p}{\Omega_g \cup \Omega_p}$, where Ω_g is the ground truth mask (each pixel is annotated with a class label) and Ω_p is the predicted mask. By calculating the Jaccard similarity index for each class individually, and subsequently computing the mean Intersection over Union (mIoU), a comprehensive summary of the methods' performance is provided.

Moreover, for a comprehensive assessment of the performance of various models across different categories within the dataset, Precision–Recall Curves are employed. These curves illustrate the trade-off between precision and recall at different threshold values. While the Intersection over Union (IoU) evaluates boundary accuracy and spatial overlap with the ground truth, Precision–Recall is particularly valuable for imbalanced datasets, as observed in our case. This curve offers a more thorough evaluation by considering both false positives (incorrectly labeled instances of the minority class) and false negatives (instances of the minority class that the model missed).

4.3. Comparisons and Discussion

The performance of the proposed framework is compared with other reference methods in Table 2. For comparisons, VGG-16, GoogleNet, AlexNet, U-Net, U-Net++, SegNet, DeepLabv3, and Huang et al. (FCN-8s) are selected. To enable VGG-16, GoogleNet, and AlexNet for the segmentation task, some modifications are made to their original architectures. The fully connected layers of these models are replaced with 1×1 convolutional layers with the number of filters equal to the number of classes (three classes in our case), allowing the networks to accept an input of arbitrary size and produce segmentation maps of the same size. Furthermore, an upsampling layer is employed to match the size of the input and a softmax activation layer. The pre-trained models of these networks are used which are initially trained on ImageNet. These pre-trained models served as a valuable starting point due to their ability to learn general visual representations from a diverse range of images. These models are then fine tuned on a weed–crop dataset. This strategy enables the networks to re-adjust the learned weights and parameters to better suit the requirements of the weed–crop segmentation task.

To ensure a fair comparison among the different methods, same training and testing ratios are maintained across all the evaluated methods to avoid the potential bias introduced by varying the data splits. The performance of these methods in terms of Jaccard similarity and mean IoU is reported in Table 2, while a comprehensive evaluation of these methods in terms of Precision–Recall curves is reported in Figure 4. From Table 2 and Figure 4, it is obvious that the proposed method achieved superior results compared to other reference methods.

From Table 2, it is observed that AlexNet-FCN achieves a comparatively lower performance in all three categories (Rice, Weeds, and Others) as well as the overall mean Intersection over Union (mIoU) than other reference methods. The suboptimal performance of AlexNet in the weed–crop segmentation task can be attributed to its shallow architecture, which can lead to the following challenges: (1) Owing to its limited depth, the network faces difficulties in extracting distinctive features from the input images. (2) The limited receptive fields of AlexNet hinder its understanding of the spatial relationships among various regions within the image. Consequently, the network may struggle to capture vital contextual information necessary for precise weed–crop segmentation. (3) Due to limited number of layers, the network struggles to capture features at multiple scales, which is crucial for the weed–crop segmentation task.

Table 2. Comparison of different methods with proposed framework in terms of Jaccard similarity (column 2, 3, and 4) and mean Intersection Over Union (mIoU) (column 5).

Methods	**Rice**	**Weeds**	**Others**	**mIoU**
VGG-16 (FCN)	0.72	0.71	0.75	0.73
GoogleNet (FCN)	0.7	0.69	0.74	0.71
AlexNet (FCN)	0.62	0.6	0.67	0.63
U-Net	0.7	0.68	0.76	0.71
U-Net++	0.71	0.74	0.77	0.74
SegNet	0.71	0.65	0.7	0.69
DeepLab	0.79	0.75	0.81	0.78
Huang et al. (FCN-8s) [23]	0.75	0.72	0.74	0.74
Proposed	0.81	0.79	0.84	0.81

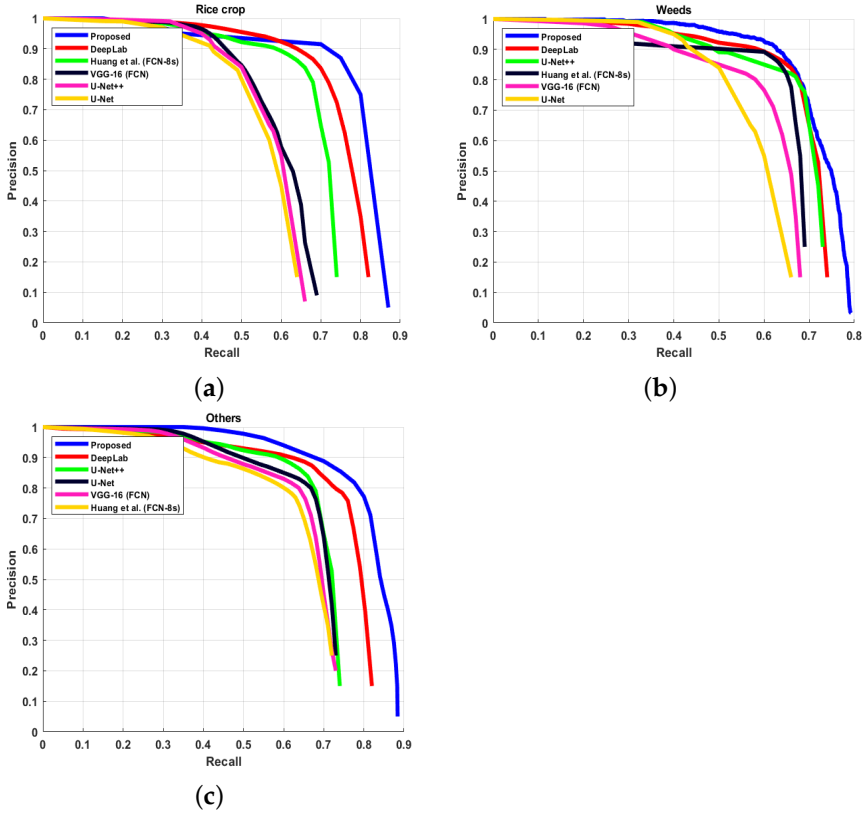

Figure 4. Performance comparison of different methods using Precision–Recall Curves. (**a**) represents the performance of different methods on "rice crop" class, (**b**) represent and compares the performance on "weed" class, while (**c**) represents the performance of different methods on "others" class.

On the other hand, DeepLab achieves a comparable performance as indicated by its higher scores in all three categories and mIoU among the evaluated methods. The superior performance can be attributed to the large receptive field of the network which is achieved via the effective incorporation of dilated convolutions. These dilated convolution operations enable the network to capture contextual information. Additionally, DeepLab incorporates the Atrous spatial pyramid pooling (ASPP) module that further enhances the model's ability to extract features at multiple scales, aiding in a better segmentation performance.

From Table 2, it is observed that all the competing methods achieve a good performance in the "Others" category, a lower performance in the "Rice" category, and the worst performance in the "Weeds" category. The superior performance of all methods in the "Others" category may be attributed to the reason that these regions have distinct visual characteristics and differ significantly from both the weed and rice regions. The methods achieve a lower performance in the "Rice" category compared to the "Others" category. This may be due to the reason that Rice plants often exhibit complex shapes, varying densities, and may have similarities in appearance to other components in the image. These problems make the segmentation task more challenging for the methods to accurately identify and segment Rice plants. The methods achieves the lowest scores in the "Weeds" category compare to the "Rice" and "Others" categories. This may be due to the reason that weeds and other vegetation may have a similar visual appearance that confuses the method to precisely identify and segments weeds, thus leading to mis-classification.

In light of the aforementioned challenges, the proposed framework demonstrates a superior performance compared to the other reference methods, achieving the highest scores in all categories. Its incorporation of the novel dense-inception units with the Artous spatial pyramid pooling module in the encoder part enables the framework to extract multi-scale and contextual information effectively. Furthermore, the incorporation of attention

units and deconvolution layers allow for the framework to recover the lost information and focus on crucial regions within the image, ultimately boosting the overall accuracy.

The evaluation and comparison of different methods across three distinct categories—rice-crop, weed, and others—were conducted using Precision–Recall curves, and the results are depicted in Figure 4a–c, respectively

In Figure 4a, which represents the rice-crop category, the Precision–Recall curve of the proposed method notably covers a larger area compared to other comparing methods, indicating a superior performance. This signifies that the proposed method achieves high precision while maintaining a high recall or vice versa, demonstrating a robust ability to correctly identify positive instances in the rice-crop category (high precision), while also capturing a substantial portion of the true positive instances in the dataset (high recall). Furthermore, the curve for DeepLab covers a smaller area than the proposed method, while Huang et al. exhibits an even smaller coverage than both the proposed method and DeepLab for the rice-crop category.

Figure 4b,c illustrates the performance of difference methods in terms of the Precision–Recall curve for the weed category and other categories, respectively. From Figure 4b,c, it is obvious that the proposed method achieves a good balance between precision and recall by covering a large area under the curve, highlighting its effectiveness in accurately segmenting weed instances. DeepLab, while still producing comparable results, falls short of the comprehensive coverage achieved via the proposed method. Notably, U-NET++ demonstrates improved results in the weed category when compared to Huang et al. This is due to the reason that Huang et al.'s model tends to generate a considerable number of false positives by mis-classifying a large number of pixels associated with the "other" category into the weed class.

The visual analysis of various model outputs is illustrated in Figure 5, where the performance of different methods is compared with the ground truth. The outputs of various models, when compared against the ground truth (second column), clearly indicate that our proposed framework consistently produces output masks closely aligned with the ground truth across all instances. In contrast, DeepLab achieves a comparable performance, accurately producing output masks in the first two samples. However, DeepLab encounters challenges in generating precise masks in other instances, leading to the mis-classification of pixels corresponding to different categories as weed. The remaining comparative models, including VGG-6, U-NET, U-NET++, and Huang et al., exhibit comparatively lower precision in their generated masks.

4.4. Ablation Study

In this section, an ablation study is conducted to understand the effect of different components of the framework on the final performance. Through this ablation study, the contribution of individual components towards the accuracy of segmentation results is assessed.

Keeping in view the proposed framework's encoder–decoder pipeline, two types of studies are performed to gain a comprehensive understanding. The first study focuses on the ablation of the encoder part, while the second study focuses on the decoder part. The details of these studies are provided as follows:

4.4.1. Ablation Study of Encoder Part

The first study focuses on evaluating the influence of the encoder's backbone network on the overall performance. Furthermore, the importance of the sophisticated Artous spatial pyramid pooling module is analyzed, taking into account its effectiveness across different pooling sizes. Through this investigation, the impact of different architectural choices for the encoder is studied as well as the size of the pooling module on the segmentation results.

For conducting this study, nine different methods are used with different settings, of which the performance (in terms of mIoU) of each method are reported in Table 3. The detail of each method is provided as follows:

1. Methods M1, M4, and M7 serve as the baseline approaches, employing the VGG-16, DenseNet, and DenseInception networks, respectively, as the encoder networks. These methods do not incorporate an ASPP module in their configurations.
2. Methods M2, M5, and M8 employ the VGG-16, DenseNet, and DenseInception networks, respectively, as the encoder network and also incorporate the ASPP module with a dilation rate of (1, 2, 3, 4).
3. Methods M3, M6, and M9 employ the VGG-16, DenseNet, and DenseInception networks, respectively, as the encoder network and also incorporate the ASPP module with a dilation rate of (2, 4, 8, 16).

Table 3. Ablation study of encoder part of the framework.

Methods	Encoder Network	ASPP Module	mIoU
M1	VGG-16	-	0.71
M2		Dilation rate (1, 2, 3, 4)	0.74
M3		Dilation rate (2, 4, 8, 16)	0.77
M4	DenseNet	-	0.73
M5		Dilation rate (1, 2, 3, 4)	0.73
M6		Dilation rate (2, 4, 8, 16)	0.76
M7	DenseInception (Proposed)	-	0.75
M8		Dilation rate (1, 2, 3, 4)	0.77
M9		Dilation rate (2, 4, 8, 16)	0.81

From Table 3, it is obvious that the choice of the encoder network affects the performance of the framework, of which it is evident from the performance variations among these models.

DenseInception, with its unique architecture, performs better than both VGG-16 and DenseNet, showcasing the importance of designing effective encoder networks tailored to the specific task of weed–crop segmentation. From Table 3, it is obvious that the methods (M1, M2, and M3) with VGG-16 as the encoder network achieve a lower performance than the methods with the DenseNet and DenseInception network. This may attribute to the reason that VGG-16 has a large number of parameters and the absence of skip connections limits the ability of the methods to capture fine-grained details and context information. On the other hand, the methods (M7, M8, and M9) with the DenseInception network as the encoder network achieve superior results than all other competing methods. This is due to the reason that the DenseInception network combines the strengths of the DenseNet and Inception architectures, which allows for dense connectivity among different layers and helps to extract multi-scale features.

From Table 3, it is observed that methods M2, M3, M5, M6, M8, and M9, which incorporate the ASPP module, demonstrate a superior performance compared to the baseline methods (M1, M4, and M7). This is due to the reason that the ASPP module enables the methods to capture multi-scale contextual information by employing Atrous convolutions at different dilation rates. It is observed that the methods utilizing the ASPP module with dilation rates of (2, 4, 8, 16) achieve a better performance than the methods with the ASPP module with dilation rates of (1, 2, 3, 4). This is due to the reason that the ASPP module with wider dilation rates results in a large receptive field, which enables the model to capture wider ranges of scale that is crucial for the semantic segmentation task in drone images.

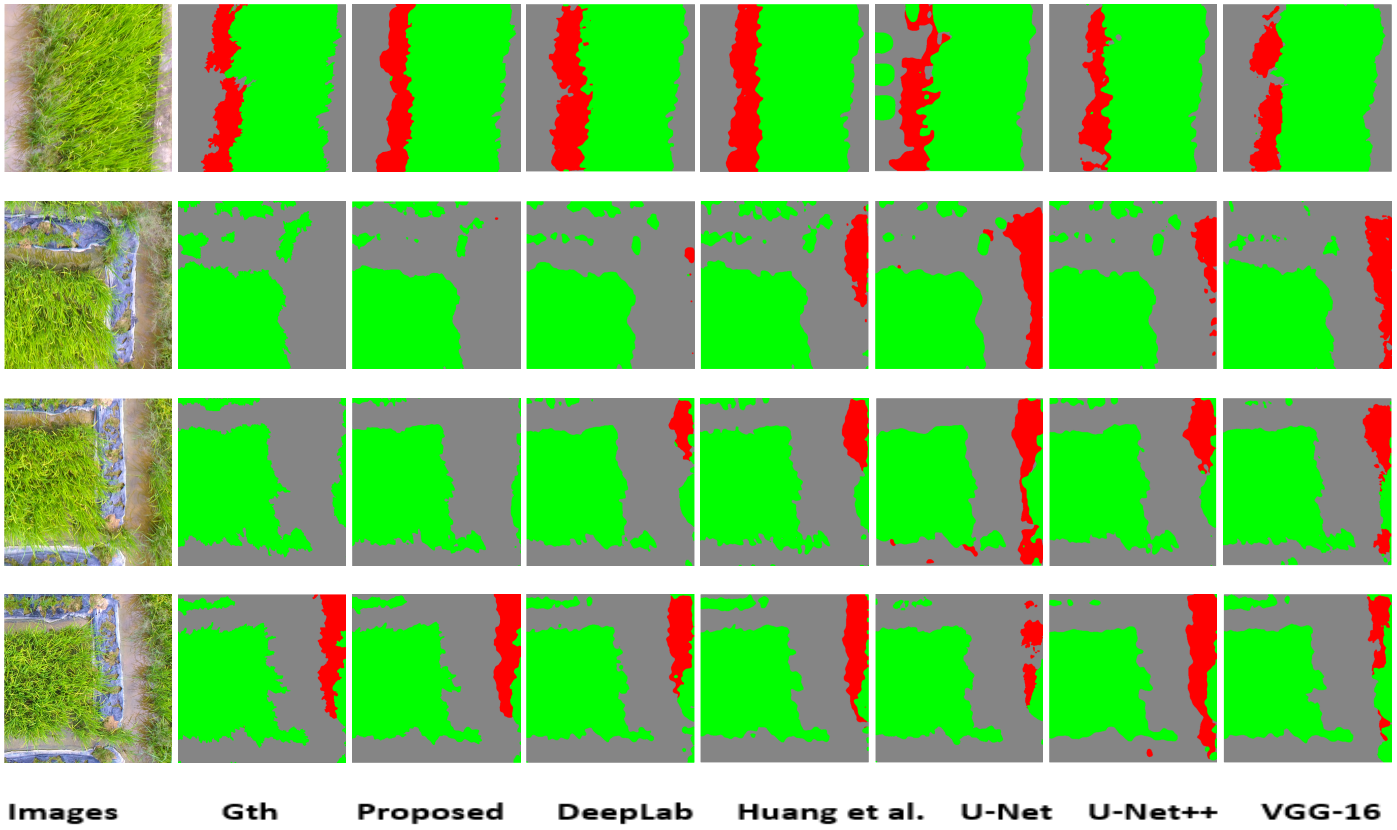

Figure 5. Comparison of predicted and ground truth segmentation masks. The first column showcases randomly selected sample frames from the dataset. In the second column, the ground truth segmentation masks are displayed, while the third column shows the predicted masks.

4.4.2. Ablation Study of Decoder Part

In this study, the influence of the attention module in the decoder component is explored. This analysis focuses on examining the effects of two distinct pooling methods: max pooling and average pooling. Furthermore, the individual impacts of the channel attention module and the spatial attention module are investigated, studying their effects in isolation. This study provides valuable insights into how different pooling techniques and attention mechanisms affect the performance of the decoder module.

To conduct this study, the performance of five distinct methods are evaluated with varying configurations and report their respective mean Intersection-over-Union (mIoU) scores in Table 4. The specific details of each method are provided below:

1. Method M1 employs the DenseInception Network and incorporates the ASPP module in the encoder module. Additionally, method M1 utilizes the channel attention unit with an average pooling operation.
2. Method M2 follows the same pipeline as method M1; however, method M2 utilizes the channel attention unit with a max pooling operation.
3. Method M3 adopts a similar encoder pipeline as method M1; however, it distinguishes itself in the decoder part by incorporating spatial attention only.
4. Method M4 adopts a similar encoder pipeline as method M1; however, its decoder part employs both the spatial attention unit and the channel attention unit with a max pooling operation.
5. Method M5 shares similarities with M4 in both the encoder and decoder components. However, in its decoder part, method M5 incorporates a channel attention unit with an average pooling operation.

Table 4. Ablation study of decoder part of the framework.

Methods		**mIoU**
M1	DenseInception + ASPP + Channel attenion (avg. pooling)	0.73
M2	DenseInception + ASPP + Channel attention (max pooling)	0.72
M3	DenseInception + ASPP + spatial attention	0.76
M4	DenseInception + ASPP + channel attention (max) + spatial attention	0.80
M5	DenseInception + ASPP + channel attention (avg) + spatial attention	0.81

From Table 4, it is obvious that Method M1 and M2 share architecture-wise similarities, yet Method M1 outperforms Method M2 in terms of segmentation results. This is due to the reason that Method M1 utilizes the channel attention unit with an average pooling operation, while Method M2 utilizes a max pooling operation. It seems that an average pooling operation is more powerful in capturing a global representation of channel-wise information, since it computes the average activation across channels and assigns weights to each channel based on its contribution, which allows the models to emphasize on meaningful features effectively. On the other hand, a max pooling operation selects the maximum activation within each channel, potentially causing information loss as it only focuses on the most dominant features. Consequently, the method using an average pooling operation captures a more comprehensive representation of channel information, resulting in better results.

Based on the experiment results in Table 4, an additional observation is made that Method M3, which solely incorporates a spatial attention unit, achieves a superior performance compared to Methods M1 and M2. This may attribute to the reason that the spatial attention unit enables the model to focus on more important regions in the image and allows the network to capture spatial contextual information which is crucial for the segmentation task. While channel attention unit enables the model to capture relevant channel-wise features. However, from the performance of methods M4 and M5, it is concluded that the combination of spatial attention and channel attention units improves the segmentation accuracy. These methods leverage the advantages of both attention units, allowing the network to capture both spatial- and channel-wise context. The integration of both channel and spatial attention units increase the ability of the network to understand complex visual patterns, resulting in a superior performance.

The proposed framework is adaptable and can be easily adapted to other crops, including sugar, rice, maize, and vegetables, by fine tuning the proposed model with the samples obtained from these crops. We believe that this adaptability will enhance the framework's versatility and utility, making it a robust solution for crop identification and characterization across a spectrum of agricultural landscapes.

5. Conclusions

This study introduces a novel weed–crop characterization framework for drone images, employing an encoder–decoder pipeline. The encoder integrates a novel DenseInception network with the Artous spatial pyramid pooling module, enabling the capture of global contextual and multi-scale information. The decoder incorporates deconvolution layers and attention units to recover spatial information and emphasize relevant regions. Extensive experiments on a challenging dataset demonstrate the effectiveness of the proposed framework. Quantitatively, the framework achieves notable success in identifying weeds and crops from drone images, showcasing its superiority. Specifically, the model achieves an impressive performance of a 0.81 mIoU score across the whole dataset (test images). These quantitative results highlight the framework's robust performance, making it a promising solution for weed–crop segmentation in precision agriculture applications.

Author Contributions: Methodology, S.D.K.; Supervision, S.B.; Funding acquisition, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data use in this work can be accessed from https://figshare.com/articles/dataset (accessed on 1 September 2023).

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

References

1. Sharma, A.; Jain, A.; Gupta, P.; Chowdary, V. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access* **2020**, *9*, 4843–4873. [CrossRef]
2. Hasan, A.M.; Sohel, F.; Diepeveen, D.; Laga, H.; Jones, M.G. A survey of deep learning techniques for weed detection from images. *Comput. Electron. Agric.* **2021**, *184*, 106067. [CrossRef]
3. Shinde, Y.A.; Jagtap, M.P.; Patil, M.G.; Khatri, N. Experimental investigation on the effect of soil solarization incorporating black, silver, and transparent polythene, and straw as mulch, on the microbial population and weed growth. *Chemosphere* **2023**, *336*, 139263. [CrossRef] [PubMed]
4. Peña, J.M.; Torres-Sánchez, J.; de Castro, A.I.; Kelly, M.; López-Granados, F. Weed mapping in early-season maize fields using object-based analysis of unmanned aerial vehicle (UAV) images. *PLoS ONE* **2013**, *8*, e77151. [CrossRef] [PubMed]
5. Bàrberi, P. Weed management in organic agriculture: are we addressing the right issues? *Weed Res.* **2002**, *42*, 177–193. [CrossRef]
6. López-Granados, F.; Torres-Sánchez, J.; Serrano-Pérez, A.; de Castro, A.I.; Mesas-Carrascosa, F.J.; Peña, J.M. Early season weed mapping in sunflower using UAV technology: Variability of herbicide treatment maps against weed thresholds. *Precis. Agric.* **2016**, *17*, 183–199. [CrossRef]
7. Urbahs, A.; Jonaite, I. Features of the use of unmanned aerial vehicles for agriculture applications. *Aviation* **2013**, *17*, 170–175. [CrossRef]
8. Lu, Y.; Young, S. A survey of public datasets for computer vision tasks in precision agriculture. *Comput. Electron. Agric.* **2020**, *178*, 105760. [CrossRef]
9. Tian, H.; Wang, T.; Liu, Y.; Qiao, X.; Li, Y. Computer vision technology in agricultural automation—A review. *Inf. Process. Agric.* **2020**, *7*, 1–19. [CrossRef]
10. Ahmad, A.; Saraswat, D.; El Gamal, A. A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. *Smart Agric. Technol.* **2022**, *3*, 100083. [CrossRef]
11. Palanivel, K.; Surianarayanan, C. An approach for prediction of crop yield using machine learning and big data techniques. *Int. J. Comput. Eng. Technol.* **2019**, *10*, 110–118. [CrossRef]
12. Kamir, E.; Waldner, F.; Hochman, Z. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *Isprs J. Photogramm. Remote. Sens.* **2020**, *160*, 124–135. [CrossRef]
13. Maimaitijiang, M.; Sagan, V.; Sidike, P.; Daloye, A.M.; Erkbol, H.; Fritschi, F.B. Crop monitoring using satellite/UAV data fusion and machine learning. *Remote Sens.* **2020**, *12*, 1357. [CrossRef]
14. Fu, Z.; Jiang, J.; Gao, Y.; Krienke, B.; Wang, M.; Zhong, K.; Cao, Q.; Tian, Y.; Zhu, Y.; Cao, W.; et al. Wheat growth monitoring and yield estimation based on multi-rotor unmanned aerial vehicle. *Remote Sens.* **2020**, *12*, 508. [CrossRef]
15. Tang, J.L.; Chen, X.Q.; Miao, R.H.; Wang, D. Weed detection using image processing under different illumination for site-specific areas spraying. *Comput. Electron. Agric.* **2016**, *122*, 103–111. [CrossRef]
16. Islam, N.; Rashid, M.M.; Wibowo, S.; Xu, C.Y.; Morshed, A.; Wasimi, S.A.; Moore, S.; Rahman, S.M. Early weed detection using image processing and machine learning techniques in an Australian chilli farm. *Agriculture* **2021**, *11*, 387. [CrossRef]
17. Burgos-Artizzu, X.P.; Ribeiro, A.; Guijarro, M.; Pajares, G. Real-time image processing for crop/weed discrimination in maize fields. *Comput. Electron. Agric.* **2011**, *75*, 337–346. [CrossRef]
18. Bah, M.D.; Hafiane, A.; Canals, R. Deep learning with unsupervised data labeling for weed detection in line crops in UAV images. *Remote Sens.* **2018**, *10*, 1690. [CrossRef]
19. Huang, H.; Deng, J.; Lan, Y.; Yang, A.; Deng, X.; Zhang, L. A fully convolutional network for weed mapping of unmanned aerial vehicle (UAV) imagery. *PLoS ONE* **2018**, *13*, e0196302. [CrossRef]
20. Mads, D.; Skov, M.H.; Krogh, M.A. Pixel-wise classification of weeds and crops in images by using a fully convolutional neural network. In Proceedings of the International Conference on Agricultural Engineering, Aarhus, Denmark, 26–29 June 2016; pp. 26–29.
21. Sahin, H.M.; Miftahushudur, T.; Grieve, B.; Yin, H. Segmentation of weeds and crops using multispectral imaging and CRF-enhanced U-Net. *Comput. Electron. Agric.* **2023**, *211*, 107956. [CrossRef]
22. Zou, K.; Chen, X.; Wang, Y.; Zhang, C.; Zhang, F. A modified U-Net with a specific data argumentation method for semantic segmentation of weed images in the field. *Comput. Electron. Agric.* **2021**, *187*, 106242. [CrossRef]
23. Huang, H.; Deng, J.; Lan, Y.; Yang, A.; Deng, X.; Wen, S.; Zhang, H.; Zhang, Y. Accurate weed mapping and prescription map generation based on fully convolutional networks using UAV imagery. *Sensors* **2018**, *18*, 3299. [CrossRef] [PubMed]

24. Lottes, P.; Behley, J.; Milioto, A.; Stachniss, C. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2870–2877. [CrossRef]
25. 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.
26. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings, Part III 18, Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
27. Zheng, Y.; Zhu, Q.; Huang, M.; Guo, Y.; Qin, J. Maize and weed classification using color indices with support vector data description in outdoor fields. *Comput. Electron. Agric.* **2017**, *141*, 215–222. [CrossRef]
28. Burks, T.; Shearer, S.; Payne, F. Classification of weed species using color texture features and discriminant analysis. *Trans. Asae* **2000**, *43*, 441–448. [CrossRef]
29. Hamuda, E.; Mc Ginley, B.; Glavin, M.; Jones, E. Automatic crop detection under field conditions using the HSV colour space and morphological operations. *Comput. Electron. Agric.* **2017**, *133*, 97–107. [CrossRef]
30. Jafari, A.; Mohtasebi, S.S.; Jahromi, H.E.; Omid, M. Weed detection in sugar beet fields using machine vision. *Int. J. Agric. Biol* **2006**, *8*, 602–605.
31. Woebbecke, D.; Meyer, G.; Von Bargen, K.; Mortensen, D. Shape features for identifying young weeds using image analysis. *Trans. ASAE* **1995**, *38*, 271–281. [CrossRef]
32. Persson, M.; Åstrand, B. Classification of crops and weeds extracted by active shape models. *Biosyst. Eng.* **2008**, *100*, 484–497. [CrossRef]
33. Perez, A.; Lopez, F.; Benlloch, J.; Christensen, S. Colour and shape analysis techniques for weed detection in cereal fields. *Comput. Electron. Agric.* **2000**, *25*, 197–212. [CrossRef]
34. Lin, F.; Zhang, D.; Huang, Y.; Wang, X.; Chen, X. Detection of corn and weed species by the combination of spectral, shape and textural features. *Sustainability* **2017**, *9*, 1335. [CrossRef]
35. Neto, J.C.; Meyer, G.E.; Jones, D.D.; Samal, A.K. Plant species identification using Elliptic Fourier leaf shape analysis. *Comput. Electron. Agric.* **2006**, *50*, 121–134. [CrossRef]
36. Chaisattapagon, N.Z.C. Effective criteria for weed identification in wheat fields using machine vision. *Trans. ASAE* **1995**, *38*, 965–974. [CrossRef]
37. Meyer, G.; Mehta, T.; Kocher, M.; Mortensen, D.; Samal, A. Textural imaging and discriminant analysis for distinguishingweeds for spot spraying. *Trans. ASAE* **1998**, *41*, 1189–1197. [CrossRef]
38. Bakhshipour, A.; Jafari, A.; Nassiri, S.M.; Zare, D. Weed segmentation using texture features extracted from wavelet sub-images. *Biosyst. Eng.* **2017**, *157*, 1–12. [CrossRef]
39. Wu, L.; Wen, Y.; et al. Weed/corn seedling recognition by support vector machine using texture features. *Afr. J. Agric. Res.* **2009**, *4*, 840–846.
40. Ishak, A.J.; Hussain, A.; Mustafa, M.M. Weed image classification using Gabor wavelet and gradient field distribution. *Comput. Electron. Agric.* **2009**, *66*, 53–61. [CrossRef]
41. Ahmed, F.; Bari, A.H.; Shihavuddin, A.; Al-Mamun, H.A.; Kwan, P. A study on local binary pattern for automated weed classification using template matching and support vector machine. In Proceedings of the 2011 IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, 21–22 November 2011; IEEE: New York, NY, USA, 2011; pp. 329–334.
42. Yu, J.; Sharpe, S.M.; Schumann, A.W.; Boyd, N.S. Deep learning for image-based weed detection in turfgrass. *Eur. J. Agron.* **2019**, *104*, 78–84. [CrossRef]
43. You, J.; Liu, W.; Lee, J. A DNN-based semantic segmentation for detecting weed and crop. *Comput. Electron. Agric.* **2020**, *178*, 105750. [CrossRef]
44. Jin, X.; Che, J.; Chen, Y. Weed identification using deep learning and image processing in vegetable plantation. *IEEE Access* **2021**, *9*, 10940–10950. [CrossRef]
45. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
46. Razfar, N.; True, J.; Bassiouny, R.; Venkatesh, V.; Kashef, R. Weed detection in soybean crops using custom lightweight deep learning models. *J. Agric. Food Res.* **2022**, *8*, 100308. [CrossRef]
47. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
49. dos Santos Ferreira, A.; Freitas, D.M.; da Silva, G.G.; Pistori, H.; Folhes, M.T. Weed detection in soybean crops using ConvNets. *Comput. Electron. Agric.* **2017**, *143*, 314–324. [CrossRef]
50. Sodjinou, S.G.; Mohammadi, V.; Mahama, A.T.S.; Gouton, P. A deep semantic segmentation-based algorithm to segment crops and weeds in agronomic color images. *Inf. Process. Agric.* **2022**, *9*, 355–364. [CrossRef]

51. Osorio, K.; Puerto, A.; Pedraza, C.; Jamaica, D.; Rodríguez, L. A deep learning approach for weed detection in lettuce crops using multispectral images. *AgriEngineering* **2020**, *2*, 471–488. [CrossRef]
52. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
53. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
54. Pettorelli, N. *The Normalized Difference Vegetation Index*; Oxford University Press: Oxford, UK, 2013.
55. Kim, Y.H.; Park, K.R. MTS-CNN: Multi-task semantic segmentation-convolutional neural network for detecting crops and weeds. *Comput. Electron. Agric.* **2022**, *199*, 107146. [CrossRef]
56. Fawakherji, M.; Youssef, A.; Bloisi, D.; Pretto, A.; Nardi, D. Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation. In Proceedings of the 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 25–27 February 2019; IEEE: New York, NY, USA, 2019; pp. 146–152.
57. Nasiri, A.; Omid, M.; Taheri-Garavand, A.; Jafari, A. Deep learning-based precision agriculture through weed recognition in sugar beet fields. *Sustain. Comput. Inf. Syst.* **2022**, *35*, 100759. [CrossRef]
58. Khan, A.; Ilyas, T.; Umraiz, M.; Mannan, Z.I.; Kim, H. Ced-net: crops and weeds segmentation for smart farming using a small cascaded encoder-decoder architecture. *Electronics* **2020**, *9*, 1602. [CrossRef]
59. Xu, B.; Fan, J.; Chao, J.; Arsenijevic, N.; Werle, R.; Zhang, Z. Instance segmentation method for weed detection using UAV imagery in soybean fields. *Comput. Electron. Agric.* **2023**, *211*, 107994. [CrossRef]
60. Janneh, L.L.; Zhang, Y.; Cui, Z.; Yang, Y. Multi-level feature re-weighted fusion for the semantic segmentation of crops and weeds. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101545. [CrossRef]
61. Zhang, J.; Gong, J.; Zhang, Y.; Mostafa, K.; Yuan, G. Weed Identification in Maize Fields Based on Improved Swin-Unet. *Agronomy* **2023**, *13*, 1846. [CrossRef]
62. Weyler, J.; Läbe, T.; Magistri, F.; Behley, J.; Stachniss, C. Towards Domain Generalization in Crop and Weed Segmentation for Precision Farming Robots. *IEEE Robot. Autom. Lett.* **2023**, *8*, 3310–3317. [CrossRef]
63. Bertoglio, R.; Mazzucchelli, A.; Catalano, N.; Matteucci, M. A comparative study of Fourier transform and CycleGAN as domain adaptation techniques for weed segmentation. *Smart Agric. Technol.* **2023**, *4*, 100188. [CrossRef]
64. Wang, A.; Zhang, W.; Wei, X. A review on weed detection using ground-based machine vision and image processing techniques. *Comput. Electron. Agric.* **2019**, *158*, 226–240. [CrossRef]
65. Wu, Z.; Chen, Y.; Zhao, B.; Kang, X.; Ding, Y. Review of weed detection methods based on computer vision. *Sensors* **2021**, *21*, 3647. [CrossRef] [PubMed]
66. Rakhmatulin, I.; Kamilaris, A.; Andreasen, C. Deep neural networks to detect weeds from crops in agricultural environments in real-time: A review. *Remote Sens.* **2021**, *13*, 4486. [CrossRef]
67. Hu, Y.; Li, J.; Huang, Y.; Gao, X. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3911–3927. [CrossRef]
68. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.