Computers and Electronics in Agriculture 178 (2020) 105750

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

A DNN-based semantic segmentation for detecting weed and crop

Jie You[a], Wei Liu[a], Joonwhoan Lee[b,*]

[a] Artificial Intelligence Lab, Department of Computer Engineering, Jeonbuk National University, Jeonju-si, South Korea
[b] RCAIT, Jeonbuk National University, South Korea

ARTICLE INFO

Keywords:
Weed detection
Semantic segmentation
Precision agriculture
Image processing
Computer vision

ABSTRACT

Weed control is a global issue, and has attracted great attention in recent years. Deploying autonomous robots for weed removal has great potential in terms of constructing environment-friendly agriculture, and saving manpower. In this paper, we propose a weed/crop segmentation network that provides better performance for precisely recognizing the weed with arbitrary shape in complex environment condition, and offers great support for autonomous robots to successfully reduce the density of weed. Our deep neural network (DNN)-based segmentation model obtains persistent improvements by integrating four additional components. i) Hybrid dilated convolution and DropBlock are introduced into the classification backbone network, where the hybrid dilated convolution enlarges the receptive field, while DropBlock regularizes the weight parameters to learn robust features by random drops contiguous regions. ii) A universal function approximation block is added to the front-end of the backbone network, which adaptively converts the existing RGB-NIR bands into optimized (RGB + NIR)-based indices to increase the classification performance. iii) The bridge attention block is exploited, in order to make the network "globally" refer to the correlated region, regardless of the distance for capturing the rich long-range contextual information. iv) The spatial pyramid refinement block is inserted to fuse multi-scale feature maps with different size of receptive fields to provide the precise localization of segmentation result, by maintaining the consistency of feature maps. We evaluate our network performance on two challenging Stuttgart and Bonn datasets. The state-of-the-art performance on the two datasets shows that each added component has notable potential to boost the segmentation accuracy.

1. Introduction

Weed control has become one of the big challenges to agricultural productivity. In biology, the weed includes a wide variety of species, which is defined as any wild plant that grows in an unwanted place, such as Dandelion, Creeping Charlie, and Pigweed. Because of the ambiguous weed definition, they have no biological significance, and are hard to describe by their biometric characteristics. But weeds usually have short growth cycles, and strongly compete with crops in acquiring limited resources, such as water, sunshine, and nutrition. As a result of the competition, weed can reproduce very rapidly, and makes a serious negative impact on the crop yield.

Conventional weed control can be divided into two directions of treatment: one is mechanical, while the other is chemical. Mechanical weeding usually applies in broader areas, where the growing weed is relatively concentrated in a certain zone, and might be convenient for removal by proper instruments. But the drawback of mechanical weeding is the increasing expenditure along with the larger operation zone. The farmer usually needs additional budget to conduct instrument maintenance, and hire more skilled workers to control the eradicator. In contrast, the chemical technology of weeding focuses on using a chemical agent, in which the farmers need to spray a considerable amount of weedicide to kill target plants, and reserve desired crops. Agrochemicals often bring side-effects on the environment, while some weed becomes resistant to chemical agent, and hard to remove thoroughly.

Recently, with the improvement of precision farming, many researchers (Slaughter et al., 2008) have announced automatic weed detection systems for weed control, to reduce the cost, and alleviate the usage of chemical medicament. In order to support the automatic robot in precisely finding the weed location and economically using the chemical weedicide, pixel-wise classification, so called semantic segmentation, is inevitable.

Semantic segmentation refers to pixel-level image classification, which is an important area of digital image analysis to locate the region of interest of objects. Previous works (Felzenszwalb and Huttenlocher, 2004; Pal and Pal, 1993) usually apply preprocessing to convert the image into a graph-based representation, and define a certain energy

* Corresponding author.
E-mail address: chlee@chonbuk.ac.kr (J. Lee).

https://doi.org/10.1016/j.compag.2020.105750
Received 6 March 2020; Received in revised form 10 July 2020; Accepted 23 August 2020
Available online 12 September 2020

function (Jiang et al., 2019) to leverage the small object region by optimize the target function. In recent years, the deep neural network (Krizhevsky et al., 2012; LeCun et al., 1998) has made a huge breakthrough in semantic segmentation, and achieved significant progress. (Long et al., 2015) builds a "fully convolutional" network to feed the image with arbitrary resolution as input, and produce successful segmentation results at any size of input images. (Chen et al., 2016a; Couprie et al., 2013; Farabet et al., 2012; Lin et al., 2017; Zhao et al., 2017) make full use of the Laplacian Pyramid to generate multi-scale structured features for effectively aggregating the coarse to fine contextual information. Another notable improvement for DNN-based segmentation is the technique of extracting discriminative features, in which dilated convolutions are adopted (Chen et al., 2017a; Chen et al., 2017b; Yu and Koltun, 2015) and extend the receptive field, without adding extra network parameters. Moreover, (Liu et al., 2015; Shuai et al., 2017; Zhao et al., 2017) build special components, such as a pyramid pooling module, to enforce the network not to barely learn the discriminative features from local regions, but to integrate the correlated features globally.

In the paper, we propose an improved DNN-based semantic segmentation network to identify weed and crop, without any domain knowledge. Our model makes great progress in precisely segmenting the weed objects by introducing the following components:

- *Hybrid dilated convolution and DropBlock.* We alter the backbone network with two technical methods, hybrid dilated convolution (HDC), and DropBlock. By stacking the two tweaks, we observe significant improvement by as much as around 5.6% in mIOU accuracy, compared with the vanilla backbone network.
- *Color-based indices.* The handcrafted color-based indices play an important role in accentuating the desired color region, and precisely separating the plant from the soil. Each index has its own characteristic, and it is not trivial to choose the proper index to meet the purpose. Following the idea in the literature (Huang et al., 2017; Lin et al., 2013; Pei et al., 2008), we propose a universal function approximation block to make a rich pool of nonlinear color-based indices. Moreover, our UFAB is fully integrated with DNN, and can automatically find the desired indices to best agree with the purpose by end-to-end learning. This idea of exploring the best machine-learned color-based indices can be expanded to multispectral images.
- *Attention mechanisms in deep network.* We adapt the attention mechanism (Fu et al., 2019; Oktay et al., 2018) to build a bridge attention block (BAB), which enforces the model to see "globally" on the feature map, and assign high attention to related regions across a vast area. Our BAB can refine the boundaries, and strongly supervise the model to learn robust features on different scale feature maps. To the best of our knowledge, our paper is the first report to incorporate the attention mechanism in weed detection. The experiment shows the proper introduction of attention operation can accentuate the interesting region, as well as attenuating undesired regions, to highlight the essential features, and increase the segmentation accuracy.
- *Spatial pyramid refinement block.* To make the model effectively decode rich spatial contextual clues at large-scale dimension, we present the spatial pyramid refinement block at the top of the architecture. The SPRB combines different scale feature maps through the diverse kernel sizes of convolution filters, while keeping the local consistency, and contributes to 0.6% improvement.

Our benchmark use two different crop/weed datasets, Bonn, and Stuttgart. The experiment on Bonn dataset shows that our network achieves 89.01% mIOU, and outperforms the second place by over 2% under stacking the above components. In addition, our architecture can produce desirable segmentation results for the different crop growth stages, even though there are some mis-annotated ground truth.

2. Related works

2.1. Semantic segmentation for crop and weeds

In the farm field, crop/weed segmentation is a challenging task for real applications. Crop and weed share similar visual characteristics, making them hard to distinguish from one another. In addition, various illumination conditions and capture angles of camera can directly influence the description of crop/weed features. The conventional method (Lottes et al., 2018b) applies different normalized color-based indices to compute the vegetation mask and eliminate the irrelevant background, then extracts the handcrafted features (i.e. biological morphology, spectral, visual features, etc.) from the generated vegetation mask to feed into the classifier to distinguish weed and crop. However, convolutional neural work (CNN) (Milioto et al., 2018) provides great interest in an end-to-end crop/weed segmentation system by using multi-spectral images. (Lottes et al., 2018a) extracted important visual features in the encoder network and fully exploit them in the decoder process with two distinct tasks for stem detection, and pixel-wise semantic segmentation. (Lottes and Stachniss, 2017) learns the plant arrangement information from the image sequences to improve the model generalization capabilities in different fields. (Lottes et al., 2020) incorporates spatial information to precisely learn the stem and features for segmentation at the same time. Our approach does not use such domain knowledge for plant location in a farm field. We solely depend on the multi-band image itself, and try to improve the performance by the additional network components.

2.2. Additional components to improve the segmentation performance

2.2.1. Hybrid dilated convolutions

Dilated convolution gains a significant improvement in semantic segmentation (Chen et al., 2014; Chen et al., 2017a; Chen et al., 2017b; Chen et al., 2016b; Wang et al., 2017a) by efficiently enlarging the receptive fields, without increasing the number of parameters. The key idea is to insert "holes" in convolution kernels to include more large-area contextual information. For example, the kernel size of the convolution filter is $k{\cdot}l$ in the standard convolution operation, while in 2-D dilated convolution, it is $k_d{\cdot}l_d$, where $k_d = k + (k-1)(r-1)$, $l_d = l + (l-1)(r-1)$, and r is the dilation rate. But, there is the problem called "gridding": when the subsequent layer has an equal dilation rate to the previous one, the convolution can only see the feature in checkerboard fashion, and loses a large portion of local information. To alleviate the adverse effect, (Wang et al., 2017b; Yu and Koltun, 2015) proposed the hybrid dilated convolution (HDC), which aggregates the convolutional layers with different dilation rates to cover the entire information, without any holes or missing edges. Others (Chen et al., 2017b; Yang et al., 2018; Yu et al., 2017) have similar ideas to concatenate multiple atrous-convolved features into final feature representation.

2.2.2. DropBlock

Regularization technique, such as dropout, is widely used to regularize the deep neural network. In most cases, dropout "ignores" part of the neurons, and enforces the remaining units to learn more discriminative features, and prevent overfitting. While the convolution operation densely scans the continuous regions and the nearby activation unit contains close semantic information, it is not "wise" for dropout to break the close relation of local semantic information. To address this problem, (Ghiasi et al., 2018) proposed the DropBlock, which discards strong association between the contiguous areas, and consequently learns spatially independent information from the remaining regions, to produce better representation learning, with less chance of overfitting.

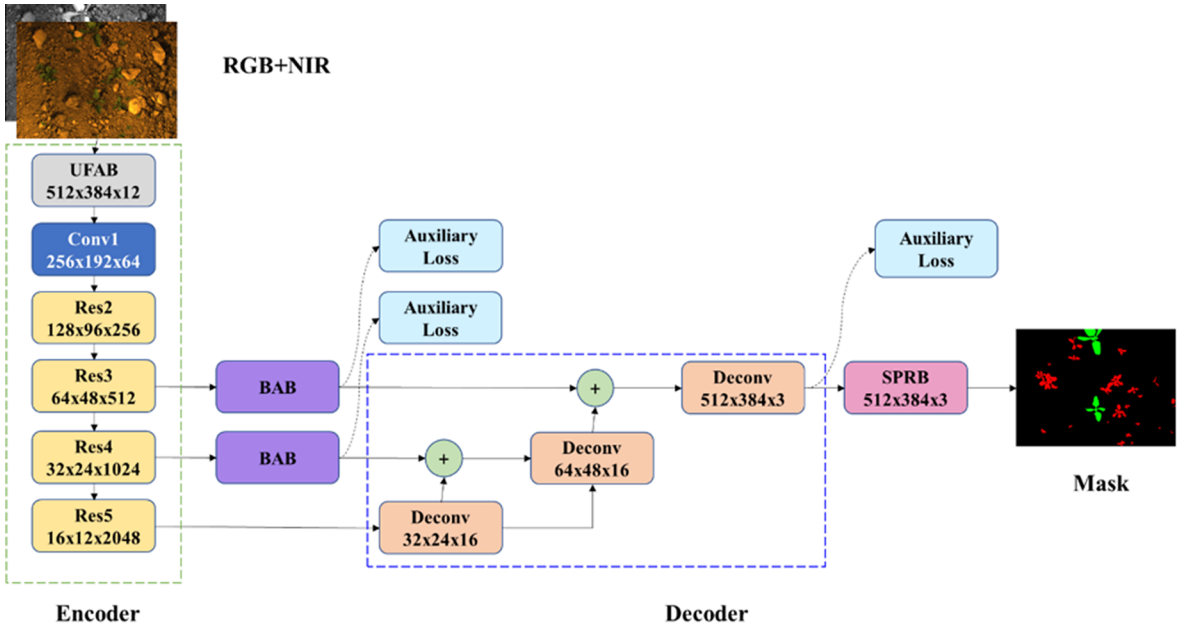

Fig. 1. Overview of the proposed weed/crop segmentation network, which includes the Res50 as backbone network. The additional components are described in the following section.

2.2.3. (Color + NIR)-based indices

(Color + NIR)-based indices (Hamuda et al., 2016; Milioto et al., 2018; Xue and Su, 2017) have been proven to be effective at enhancing the inherent color property of objects. For example, ExG (Woebbecke et al., 1995) uses chromatic coordinates to express the new representation through the transformation formula 2 g-r-b, which shows a clear contrast from the non-plant background. CIVE (Kataoka et al., 2003) emphasizes the green information, and obtains the discriminate features to segment the crop from the soya bean and sugar beet image. NDI (Woebbecke et al., 1993) measures the difference between red, green, and blue bands of the image, and demonstrates the vivid distinction between the soil background and plant objects. Others, such as (Burgos-Artizzu et al., 2011; Hague et al., 2006; Hunt et al., 2005; Meyer et al., 2004), reduce the influence of illumination, and present good performance at separating the plant pixels. But the above indices still have limitations when the dominant plant color is green for the discrimination of weed and crop. To address this problem, (Milioto et al., 2018) calculates the 14 color indices from the RGB data, and feeds them into the deep neural network to achieve accurate weed recognition under various weather and lighting conditions. All those color-based indices are handcrafted, and may not be easy to properly apply, without specific knowledge of their properties.

Previous works (Pei and Mai, 2008; Pei et al., 2008) show that the feedforward neural network can be treated as a nonlinear function approximation. Inspired by those works, we introduce a universal function approximation block (UFAB) to perform the basic arithmetic operations and automatic mapping of the multi-band input to latent color-based indices through the trainable network. Different from (Pei and Mai, 2008; Pei et al., 2008), our block consists of stacked 1 times 1 convolutional layers (Lin et al., 2013) instead of the conventional fully connected layer, to weaken the chance of "parameters blow up", and to speed up the convergence. Also, we exploit the cascade dense connections block (Huang et al., 2017). In dense connections block, all the preceding layers are concatenated to make the input of the succeeding layer, which reduces the redundancy in the pool of nonlinear functions with diverse information paths, and strengthens the model diversity.

2.2.4. Attention mechanisms

Attention is the basic property of the human visual system. It is well known that people do not process the entire scene at once; they intuitively glimpse a series of salient regions, and make a judgment by aggregating various local information. In accordance with the sampling attention location, attention mechanisms can be split into two types, soft attention and hard attention. Soft attention (Xu et al., 2015) computes the weight distribution "softly" throughout the entire patches of the source input, emphasizing the prominent position, and suppressing the irrelevant counterparts. On the other hand, hard attention just selects only one patch at a time, and uses Monte Carlo sampling to gradually approach the global best, which effectively reduces computational complexity. Also, it should be integrated with complicated algorithms, like the recurrent neural network (Ba et al., 2014; Gregor et al., 2015) or reinforcement learning (Xu et al., 2015).

In computer vision, (Wang et al., 2017a) designed a cascade bottom-up and top-down attention module named residual attention (RA), to extract the dense discriminative features for image classification. (Jetley et al., 2018; Zhang et al., 2018b) learn a soft weight distribution from the intermediate stages in the CNN pipeline, by combing with global feature vectors to feed into the final classification layer. Moreover, rather than only inferring attention maps in the spatial dimension, (Wang et al., 2017c; Woo et al., 2018) extend the attention process into channel and temporal dimensions, respectively. Similar to (Fu et al., 2019), we integrate the bridge attention block to full use of the attention mechanism in the shortcut of the encode and decode part, to alleviate GPU memory usage, and elaborately capture the global spatial information from the moderately low to high level, to precisely locate the weed position.

2.2.5. Spatial pyramid refinement block for multi-scale feature fusion

Fusing the low- to high-level features produces consistent improvement in the segmentation task. (Liu et al., 2015) successfully applies the global context pooling to exploit global semantic information in the fully convolutional network. (Zhao et al., 2017) builds the pyramid scene parsing network to provide additional contextual information for precise scene parsing. In recent study, the dilated (atrous) convolution (Chen et al., 2014) makes great progress in enlarging the receptive field, without increasing the computational burden. Based on this, (Chen et al., 2017a; Chen et al., 2017b; Wang et al., 2017a; Yang et al., 2018) propose atrous spatial pyramid pooling (ASPP), to select the multiple dilation rates for generating densely spatial- and scale-sampled features, and effectively boost the model ability to recognize

objects of arbitrary size. In our work, we fuse multi-scale representation into the decoder, to obtain more precise segmentation boundaries.

3. The proposed approach and implementation detail

This section describes the details of the proposed semantic segmentation network. Fig. 1 shows the overall structure of our network, where the simplest version of FCN8s (Long et al., 2015) is chosen as our baseline for pixel-wise classification. First, we alter the encoder part to take RGB + NIR images as inputs. Thereafter, UFAB is located to automatically generate proper RGB + NIR indices, and followed by Res50-based backbone for extracting features with diverse scales in the encoder part. Then, several attention units are applied to model the long-term spatial dependency information. In the decoder part, the output of a specified encoder layer is up-sampled and followed by deconvolution to be combined with its corresponding BAB result successively to restore the input resolution. Finally, the robust segmentation results are given by fusing the multi-scale contextual information.

3.1. Description of the improved backbone network

In our work, the backbone network is Res50, and we integrate the two methods to model a larger receptive field to detect the tiny weed, as well as compose robust spatial independent information.

3.2. Hybrid dilated convolution(HDC)

We follow the idea of HDC that aggregates the convolutional layers with different dilation rates to cover the entire information, without any holes or missing edges. In detail, the setting of HDC is followed by (Wang et al., 2017b)'s "dilation-bigger". The dilation rate for convolutional layers in res4_1 and res4_2 is 2, and they are 5, 9, 1, and 2, for following convolutional layers in res4_3, res4_4 and res4_5, respectively. The final res5 has three residual blocks, and is attained with higher dilation rate of "5, 9, 17" to get larger receptive field features.

3.3. DropBlock

We attempt to join the dropout with constant probability into the backbone network, but fail to make sure of the network convergence. For this reason, we introduce the DropBlock into our backbone network, which randomly zeroes out the continuous regions, breaks the close relation of local semantic information, and forces the model to "look around" the remaining regions. Specifically, we set the DropBlock in the output of the convolutional layer at the res4, res5 block, and discard the downsample convolutional layer, to avoid overfitting and learn uncorrelated features. The keep probability is 0.8, block size 7, to break more spatial correlation information.

3.4. (RGB + NIR)-based indices with the universal function approximation block (UFAB)

We consider the universal function approximation block (UFAB) as depicted in Fig. 2, which automatically approximates any nonlinear color-based indices by training. The UFAB calculates the input of RGB + NIR images by several multilayer perceptron units, and provides a rich index pool to precisely classify the crop and weed pixels in a complex environment. Commonly, the single multilayer perceptron unit is defined by:

$$H_{i,j} = \sigma\left(\varphi\left(\sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{l=1}^{L} w_{k,l}^{c} x_{i+k-1,j+l-1} + b^{c}\right)\right) \quad (1)$$

where $H_{i,j}$ represents the output of a multilayer perceptron unit, x corresponds to the input signal, and w and b denote the weight and bias values, respectively, in convolution filters. For any convolution block in UFAB, kernel size $K = L = 1$, c is the number of channels, and σ is the nonlinear activation function. We place the BatchNormalization φ to alleviate the covariance shift, and speed up convergence. To diversify the choice of mapping functions, we construct the dense connection (Huang et al., 2017) by stacking the multilayer perceptron unit. The dense unit can be represented by:

$$x^{\ell} = H^{\ell}([x^0, x^1, x^2, ...,x^{\ell-1}]) \quad (2)$$

where x^{ℓ} denotes the output of the ℓ-th multilayer perceptron unit, H is the multilayer perceptron unit we mentioned before, and $[\cdot,\cdot,..., \cdot]$ denotes the concatenate operation to merge multiple outputs into a single tensor. The layer ℓ receives the forward information from 0, ..., $\ell - 1$, and shares the values with the next layer. So, we can obtain the pool of nonlinear indices in the ℓ-th layer, in which the indices are taken directly from the input bands to the full composition of $\ell - 1$ convolutional layers. The learned convolution blocks automatically approximate the proper and rich color-based indices.

In our experiment, we set hyperparameter c and layer ℓ to 4 and 2, respectively. In addition, the σ nonlinear activation is implemented by sigmoid activation function to model the latent nonlinear color-indices from RGB + NIR input sources.

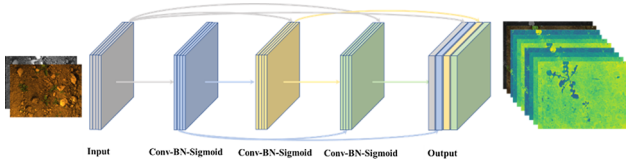

Fig. 2. Universal function approximation block, which learns the latent information from RGB-NIR channels, and automatically learns latent color-based index features by network training.

3.5. Bridge Attention Block (BAB)

Fig. 3 shows the overview BAB with the following approaches:

3.5.1. Spatial attention module

We first extract the features from coarse to fine scale at the end of each block in the backbone network. Then, the extracted features are shrunk to the same number of channels via a global convolutional module (Peng et al., 2017). Next, the channel-adapted feature $x_i \in R^{H \cdot W \cdot C}$ is sent to three branches, and the linear transformations f, g, and h are taken to conduct learnable tensor F, G, and H, which represent the key, value, and query tensor, respectively (Zhang et al., 2018b). Then the key, value and query tensors are transposed and resized to $F' \in R^{C \cdot N}$, $G' \in R^{N \cdot C}$ and $H' \in R^{N \cdot C}$, where $N = H \cdot W$. Thereafter, we perform dot product for F', and G' tensors, and take the normalization operation to generate the soft spatial attention map $S \in R^{N \cdot N}$. Finally, the enhanced features are computed by matrix multiplication: $A = S \cdot H$. Note that the scale normalization $S = \frac{1}{C}(F' \cdot G')$ is applied to emphasize the outstanding region of interest, where C is the number of channels to normalize the attention weight.

3.5.2. Boundary refinement module

For the sake of computation simplicity, we apply the spatial attention module in coarse resolution, while the spatial information would be significantly lost and lead to performance decay when the feature map is reduced. We follow the idea (Peng et al., 2017) of applying a boundary refinement module to polish the boundaries, and retain detail information. The module consists of two additional branches; one branch has two convolutional layers denoted by B(), and the other branch is for shortcut connection. Therefore, the output $\widehat{A} = A + B(A)$, the stacked convolutional layer models the boundary alignment and gradually refines the spatial context to provide better discriminative features $\widehat{A}$. Meanwhile, the shortcut connection allows the attention

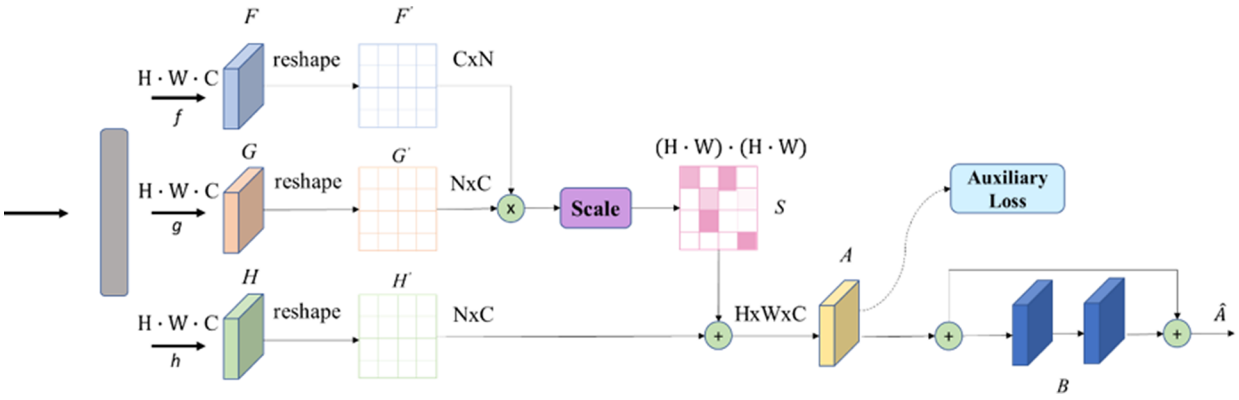

Fig. 3. Bridge Attention Block, where F, G, and H represent the different discriminative features generated by the convolutional layer. We apply auxiliary loss to the end of feature A, which is then followed by two standard 3×3 convolutional layers with ReLU nonlinear activation function to refine the feature boundaries.

information to skip the convolutional layer, and selectively emphasizes the boundary locations.

3.5.3. Implement detail

The BAB is applied in the outputs of res3_4, and res4_6, and the outputs of the two-stage features reduce to 16-channel features by a 5×5 global convolution module. Each linear transformation function $f, g,$ and h is implemented by another 1×1 convolutional, BatchNormalization, and ReLU layer. The nonlinear activation function is absent to preserve semantic consistency in BAB. In boundary refinement modules, we simply use two 3×3 convolutional layers for $B()$ to further refine pixel-wise object boundaries. We do not use the BAB in res5_3 layer; the output feature of res5_3 layer is 16×12. Applying the BAB in res5_3 produces unfavorable effect, and causes performance degradation; for further discussion, see Section 4.4.1.

3.6. Spatial Pyramid Refinement Block (SPRB)

We devise a SPRB Fig. 4 succeeding the output of the decoder, which consists of several branches to directly obtain precise pixel-wise output by aggregating multi-scale contextual information. Each branch has a convolutional layer with one kernel size, where each branch denotes a feature map with different size of receptive field to cover the local object from small to large areas, with the merit of fusing various kernel size of convolution filters. The outputs of these parallel layers are concatenated along with the dimension of depth, and conduce to a single feature map via a 1×1 convolutional layer. Specially, we exploit nonlinear activation for 1×1 convolution to sum over the score of the majority of regions to generate final logits, and discard the unnecessary information.

Because of the computation efficiency, we fuse 4 scales corresponding to kernel sizes of 3×3, 5×5, 7×7, and 9×9 with the channel size 4 to enforce local consistency in high resolution. Also, we perform additional loss in front of the block to force the network to generate more reliable localization results.

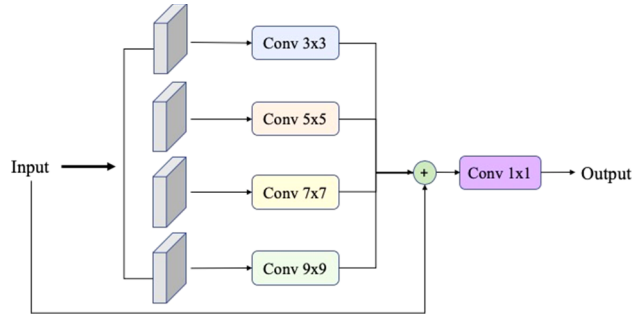

Fig. 4. Detail of the spatial pyramid refinement block, where each branch has various kernel size with the same number of channel sizes. Then to avoid accuracy degradation, we add the shortcut connection.

4. Experiment results

In this section, we first briefly introduce our dataset for experiments. Then, we show several experimental results, and explain the improvement for using different components in our proposed network.

4.1. Dataset overview

We conduct all the experiments on two public datasets named "Bonn", and "Stuttgart". The Bonn dataset was captured on a sugar beet farm near Bonn in Germany, and covers the whole growth stage of plants. While "Stuttgart" records the plant and weed distribution, there is no available timestamp for sugar growth, and the images are mixed in different growth stage. The two datasets amount to around 5 TB, including visual and geo-location ABD real-time kinematic (RTK) data. All the visual data is captured by CCD camera, and provides 4-channel (RGB, NIR) multi-spectral information. Each image has its corresponding segmentation mask, and Fig. 5 shows the samples from two datasets. There are four categories of objects of crop, soil dicot weed, grass weed, and background. We combine the two types of weed into a single category "weed", because of the few samples, and reconstruct the dataset. Note that the Bonn dataset records the data from the emergence of plants and stop until maturity, which contains the full timestamp of data across each growing stage of plants, and the number of weed pixels is fewer than that of crop pixels (Table 1). In order to build a robust network, we extract 80% images in each stage from the Bonn dataset, and resize the multi-spectral image into 512×384 size. In summary, we use 80% Bonn images (7,270) for training, and the remaining 20% Bonn (1,800) and 100% Stuttgart data (2,584) to evaluate the network generalization performance.

4.2. Training strategy and metrics

We implement our methods on Keras (Gulli and Pal, 2017), and

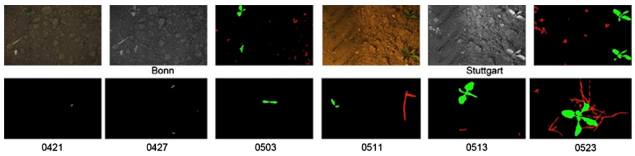

Fig. 5. The first row shows the Bonn and Stuttgart datasets, which include RGB, NIR, and pixel-wise label, while the second row displays six out of the twenty growth stages of plant masks; the label is the time of taking pictures.

Table 1
The summary statistics of crop and weed distribution in the Bonn and Stuttgart datasets.

	Bonn	Stuttgart
Total image	9,070	2,584
Crop pixels	1.7%	1.5%
Approx. crop size	(0.5–100) cm^2	(3–25) cm^2
Weed pixels	0.7%	0.7%
Approx. weed size	(0.5–60) cm^2	(2–60) cm^2

train the network architecture by Adam optimizer (Kingma and Ba, 2014). In order to avoid overfitting, we use standard online augmentation methods to increase the number of training data, such as random horizontal, vertical flipping, and shift random pixels in the x, y direction. Additionally, we introduce random brightness jittering within the range $[-0.1, 0.2]$, to alleviate the impact of natural illumination. All the models have trained for 300 epochs with the mini-batch size of 8 on TITAN XP. The learning rate starts at 0.3, and is multiplied by 0.8 every 30 epochs to stabilize the network converges. We use 3 types of loss to leverage the error with ground truth label.

$$\mathscr{L}_{total} = \alpha\mathscr{L}^{S} + \beta\mathscr{L}^{SPRB} + \gamma\mathscr{L}^{BAB}_{3\_4} + \delta L^{BAB}_{4\_6} \quad (3)$$

where $\mathscr{L}^{S}$ represents the primary loss of semantic segmentation near the output, $\mathscr{L}^{SPRB}$ is the auxiliary loss in SPRB, and $\mathscr{L}^{BAB}_{3\_4}$ and $L^{BAB}_{4\_6}$ are the losses of spatial attention modules in stage res3_4 and res4_6 for directly leveraging the ground truth mask with predict output to force the BAB to learn better long-range contextual information. All the auxiliary supervision structures follow the design (Zhang et al., 2018a), which is constructed with a 1convolutional layer, followed by an upsampling, and a softmax layer. To balance the different loss components, we set the hyperparameter α, β, γ, and δ as (16, 16, 2, and 1), respectively, according to the resolution of feature maps extracted from the layers. The input RGB + NIR images are linearly rescaled to range [0, 1], and we choose three standard semantic segmentation metrics for performance evaluation, which are mean intersection over union (mIOU), mean pixel accuracy (mPA), and per class iou (cIOU).

4.3. The experimental results on UFAB

4.3.1. UFAB hyperparameter selection

In UFAB, c is the number of channels, and ℓ means the number of stacked convolutional layer; substantially increasing c and ℓ strengthens the capacities for fitting nonlinear functions, but potentially adds more parameters, and increases the risk of falling into the trap of sub-optimization. In particular, we evaluate the difference hyperparameter c, and layer ℓ, and draw Table 2. The mIOU increases at the beginning, but slowly degrades by increasing the number of channels c, because it could improve the inter-class representation and approximate more color indices at the same time; but larger c would make the network explore more indices with a limited number of training data, which has negative effects on the training results. A similar result to c is observed by increasing ℓ. The nonlinearity in the pool could be enhanced by increasing ℓ, as well as the number of indices. This implies that too complicated nonlinearity due to the successive compositions in the dense net is not that helpful. We expect that the best performance can be achieved depending on the number of input bands, and the structure of UFAB.

Table 2
The design choices for the hyper-parameter c, ℓ on UFAB.

c	ℓ	mIOU(%)	Parameters(M)
2	2	87.96	24.13
4	2	88.42	24.14
6	2	88.14	24.15
8	2	88.12	24.17
16	2	88.09	24.22

a) Fixing the number of ℓ

c	ℓ	mIOU (%)	Parameters(M)
4	2	88.42	24.14
4	4	88.40	24.16
4	6	88.36	24.19
4	8	88.31	24.21

b) Fixing the number of c

4.3.2. Selection of nonlinear activation function in UFAB

We find the UFAB is sensitive to the choice of nonlinear activation function to approximate any color indices. We first attempt to use the rectified linear unit (ReLU) as nonlinear function, which has a wider range of applications in the convolutional neural networks. The ReLU activation function accesses the simple calculations and is able to accelerate the convergence speed, but performance degradation was found, which we suspect comes from ReLU preserving the positive part, and pruning the negative part to zero. In other words, the negative part goes to "die", and the weights are not altered in the consequent training iteration (Xu et al., 2015). For this reason, we select saturated activation function (sigmoid) (Xu et al., 2015) at the end of the convolutional layer, to give the chance of recovering the values. The sigmoid activation can be expanded into a higher order (theoretically infinite order) polynomial, and approximate any nonlinearity. In addition, it restricts the value range from (0 to 1), which seems like normalization for the image to ensure that each input pixel has a similar data distribution, so we are able to safely use the pre-trained ImageNet weights for model initialization. Another merit is better explanatory power; it is easy to visualize the output feature map as images, to understand the internal workings of the UFAB. Fig. 6 shows that the result supports our qualitative analysis; the sigmoid activation function performs better than ReLU in approximating any color-based index function.

4.4. The experimental results on BAB

4.4.1. The attention method in BAB

In terms of our weed/crop segmentation, the BAB emphasizes the relevant local information, and suppresses unnecessary clutter in the processing feature map at a certain position. The selection of attention mechanism is key to exploiting rich context-aware information. We use two groups of soft attention mechanisms to evaluate the segmentation network for our weed/crop task. The first group uses simple dot scale (Zhang et al., 2018b) and dot softmax (Jetley et al., 2018); they attempt to compose the alignment score function to consider the relative importance between key array and value array. Another group consists of RA (Wang et al., 2017a) and CBAM (Woo et al., 2018); they stack a series of convolutional layers, instead of computing the alignment score to get the high-attention image regions. We remove other components,

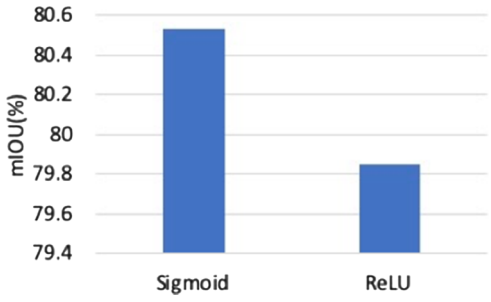

Fig. 6. The performance comparison of Sigmoid and ReLU activation functions. We use FCN8s with vanilla Res50 as baseline, and apply the UFAB at the front-end of the network. The same hyperparameters, $c = 4$ and $\ell = 2$ in UFAB are used, and each method repeats three times to get the average.

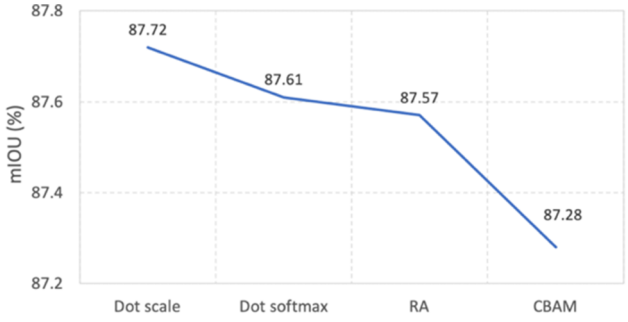

Fig. 7. The performance of four attention mechanisms.

Table 3
The mIOU accuracy of dot scale and softmax method in BAB. We compared the performance in different resolutions of the feature map, where res3_4, res4_6, and res5_3 corresponding to the feature map sizes 64×48, 32×24, and 16×12, respectively.

Method	Stage	mIOU(%)
Base		87.29
Dot scale	res3_4	87.32
	res4_6	87.72
	res5_3	87.09
	res3_4,res4_6	88.11
Dot softmax	res3_4	87.48
	res4_6	87.61
	res5_3	85.17
	res3_4,res4_6	86.77

and replace the spatial attention module by those two group methods; Fig. 7 shows the comparison result. Interestingly, the second group methods present a worse result than the first in our problem. The reason seems to be related to the number of parameters, and the size of object to be detected. The dot scale/softmax attention has few parameters, and during training, the message seems to be easy to deliver back and forth. Our weed object is relatively small, and is therefore prone to losing spatial information by the stacking of down- and up-convolutional layers. Further, we discuss the position of BAB in Table 3. In most cases, the mIOU increased by attaching the attention modules, but the performance dropped in stage res5_3, both in dot scale and softmax methods, whose resolution is too small to provide precise spatial information. Attaching to multi-stage feature maps consistently increases accuracy in the dot scale method.

4.4.2. Attention visualization

We visualize the BAB block, and express the reason why our network is able to capture the global contextual information no matter how far away the related feature is. For the spatial attention module in BAB, the size of the spatial attention map is $N_1 \times N_2$, where each N_1 and N_2 equals the size of feature map, i.e., $(H \times W)$. N_2 represents the sub-attention map for a point P in N_1, which shows the long-range dependency in N_1 corresponding to the selected point P. In other words, the locations with the similar pixel distribution to P are highlighted in N_2. Fig. 8 shows the coefficient matrix N_2 (32×24) with its corresponding point in N_1. For example, the green point #1 in the image (row 1 and column 1) marks weed pixels. The corresponding sub-attention map (row 1, column 5) emphasizes the most corresponding weed areas. In addition, we can see the place (row 1, column 5, marked as red circle) that shows significant difference from the surrounding area, even though some of the weed pixels are far away from the marked point #1. A similar phenomenon appears on the image (row 2, column 5, marked as red circle) again, where the attention map focuses on not only the principal weed positions at bottom-left, but also glances at the area at top-left (the small weed areas), to explore the global contextual information during the forward propagation. The pixel with high weight encourages the network to localize semantic relevant regions, and capturing a large amount of long-range contextual information to improve feature representation.

4.5. Improvement by additional components

We conducted the ablation experiment on the Bonn dataset to show how the additional components improve the performance. We fixed the hyperparameters, and continuously added one of the proposed components. Table 5 summarizes the performance. When the hybrid dilation convolution is introduced into the backbone network, it shows the significant increment of around 4.65%, compared to the baseline Res50, in which the receptive field covers a wider area in terms of capturing detailed information from the small weed. The DropBlock strategy contributes around 1% increment by regularizing the convolutional layer, and focuses the model to learn more invariant features from non-continuous regions. Then, the BAB gains another 0.64% increment by including rich long-range contextual information with a small amount of parameter increase. UFAB and SPRB further result in around 1% improvement to provide more precise results. Note also that the whole of our proposed methods achieves 7% performance improvements compared to the baseline network, with only 0.04 million additional parameter increase.

4.6. Comparative study

In this section, we first compare our UFAB method with the conventional hand-crafted color indices. Then, we evaluate our model on the Bonn and Stuttgart datasets. Finally, we further discuss why the weed/crop segmentation task is a challenge, and how robust our network is.

4.6.1. Comparing UFAB-based color indices with the conventional (RGB + NIR)-based handcrafted indices

To evaluate the effectiveness, we select several well-known color-based functions as input to compare with our UFAB block. First, we just generate the RGB-based indices according to Table 4, and concatenate them with original multi-spectral (RGB + NIR) bands to replace the UFAB in the network(RGB mix in Fig. 9 and Table 6). Second, we continue to concatenate three multi-spectral (RGB + NIR)-based indices with RGB-based indices, in order to make full use of the multi-spectral information. For fair comparison, we stack one 1×1 convolutional layer with the number of channels 8 to decrease the channel dimension for the generated multispectral indices, and keep the same number of output channels as our proposed UFAB. Strictly speaking, this set of 11 indices include both handcrafted and learned ones (RGB + Multispectal). Then, we train the remaining part, and evaluate the performance on the Bonn test dataset. Table 6 shows the comparison result; the handcrafted color indices features achieve 87.81% accuracy in mIOU, which gains 0.2% improvement compared with the baseline (neither UFAB nor color indices, just RGB + NIR). Increasing the multi-spectral information brings more advantage and gives 88.20% mIOU accuracy, but is still lower than our proposed method by around 0.5% difference. The conventional method only obtains 72.30% weed IOU accuracy, compared to ours of 73.50%. The UFAB increases the performance for both the weed and crop objects, it introduces more potential discriminate features to distinguish the green plants (crop/weed) from the undesired regions (soil background regions). Moreover, the network learned indices seems to maintain “the object integrity”, the adjacent pixels are more likely to belongs to the same object. In Fig. 9, we explore the segmentation result in terms of different approaches, it is obvious that the weed object is complete split from other regions and conventional methods fail to remove the crop pixels from weed object, where the UFAB enforce consistency over visuality similar regions. More evidence can find in columns 6–9, our method efficiently suppresses the irrelevant background pixels, while preserving the

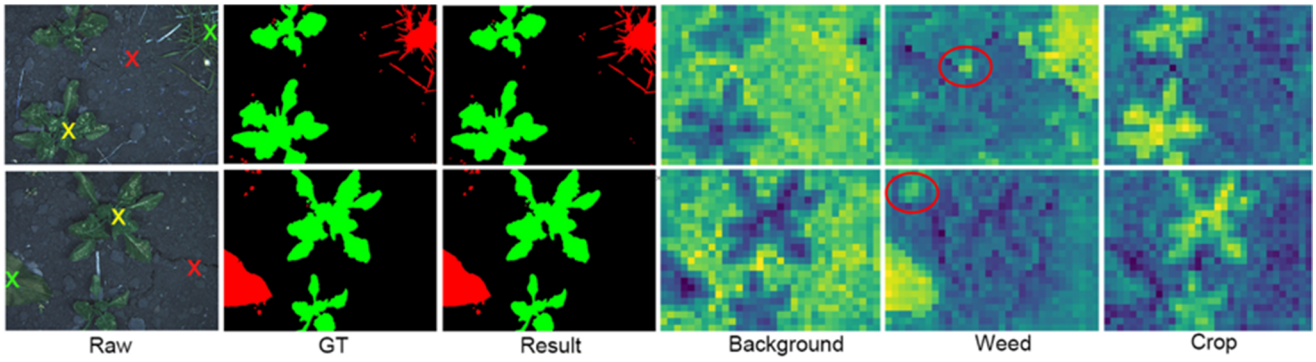

Fig. 8. Visualizing the spatial attention map, we select three points from the sub-attention map N_1, and find their coefficient matrix N_2. The sub-attention map highlights the relevant areas, and suppresses the counterpart. The marked points of red, green, and yellow represent the positions of background, weed, and crop. Columns 4 to 6 are their coefficient matrix N_2. Best seen in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
The conventional color index functions. We preprocess the image by the formula, then normalize the value to range [0, 1].

Description	RGB-based Index Function
NDI	$128 \cdot (((G - R)/(G + R)) + 1)$
ExG	$2g - r - b$ $r = R^*/(R^* + G^* + B^*)$ $g = G^*/(R^* + G^* + B^*)$ $b = B^*/(R^* + G^* + B^*)$
ExR	$1.3R - G$
ExGR	$ExG - ExR$
CIVE	$0.441R - 0.811G + 0.385B + 18.78745$
NGRDI	$(G - R)/(G + R)$
MExG	$1.262G - 0.884R - 0.331B$
VEG	$G/(R^a B^{1-a}),\ a = 0.667$
	(RGB + NIR)-based Index Function
RVI	R/NIR
DVI	$NIR - R$
NDVI	$(NIR - R)/(NIR + R)$
	$G,\ R,\ B,\ NIR \in [0,\ 255]$, $R^*, G^*, B^* \in [0,\ 1]$.

important visual properties of "weed" and "crop" objects against background "soil".

4.6.2. The performances on Bonn dataset

We compare the performance of our network with other prevailing DNN-based semantic segmentation architectures to show the effectiveness. We use the same improved res50 as the backbone on the Bonn testing dataset. In particular, for PSPNet (Zhao et al., 2017), we apply the pyramid pooling module with four pooling sizes of "1, 2, 3, and 6", and zero paddings along the width direction. SegNet (Badrinarayanan et al., 2017), UNet (Ronneberger et al., 2015), and FCN8s (Long et al., 2015) fuse with the multi-scale features from res3_4 and res4_6 layer. We keep the same hyperparameters to train those methods with the same epochs. Table 7 illustrates our network performance compared to others. It is worth mentioning that our proposed method exceeds others, and achieves the best result by 89.01% in terms of mIOU by as much as a large margin of around 2% than the second one. In particular, the improvement on the per-class IOU accuracy demonstrates well that our method effectively captures the discriminative features to split weed and crop from background pixels. The more impressive evidence can be found in Fig. 10, where the other methods easily misclassify the crop and weed (row 1) at the overlapped area, while our network differentiates them successfully. Our method performs well to "see" globally over the whole semantic information to segment long and thin weed (row 3), even though it is hard for the convolutional layers (Russakovsky et al., 2015) to detect the small thin structures.

4.6.3. Result on Stuttgart dataset

To further evaluate the model generalization ability, we try to make inference on 2,584 images from the Stuttgart dataset by using our proposed network. Table 8 shows the comparison, where our framework is resilient to the unknown data, and outperforms the other well-known architectures by as much as around 2% margin. The network was just trained with the Bonn training dataset; there was no domain adaptive structure (Huang et al., 2018; Sankaranarayanan et al., 2018; Zhang et al., 2018c) to transfer learnable features from source to target domain. Even though there was performance degradation while detecting the weed from the background in the feature domains, our network still provides better generalization capability to tackle the related task with unseen datasets.

4.6.4. Performance of mis-annotated samples

Our algorithm has failed to increase the accuracy further, due to the noisy or ambiguous annotated labels among the boundary of weed and crop. But we found interesting behavior of our trained network. For example, the long and narrow weed overlaps the healthy crop, but is labeled as crop in the first column in Fig. 11. In this case, our model is still robust to obtain the weed pixels, and distinguishes well the clear boundaries that separate weed and crop. Another example is shown in the second column, where there is nothing present in the bottom-left position of the RGB image, but the ground truth mask exposes the weed object in the corresponding position. Column 3 displays a more

Table 5
The performance persistently improves by attaching additional components.

Description	mIOU(%)	mPA(%)	Parameters(M)
Res50	81.72	88.84	24.10
Res50 + HDC	86.37	91.29	24.10
Res50 + HDC + DropBlock	87.29	92.49	24.10
Res50 + HDC + DropBlock + BAB	87.93	92.60	24.11
Res50 + HDC + DropBlock + UFAB + BAB	88.17	92.73	24.13
Res50 + HDC + DropBlock + UFAB + BAB + SPRB	88.72	93.08	24.14

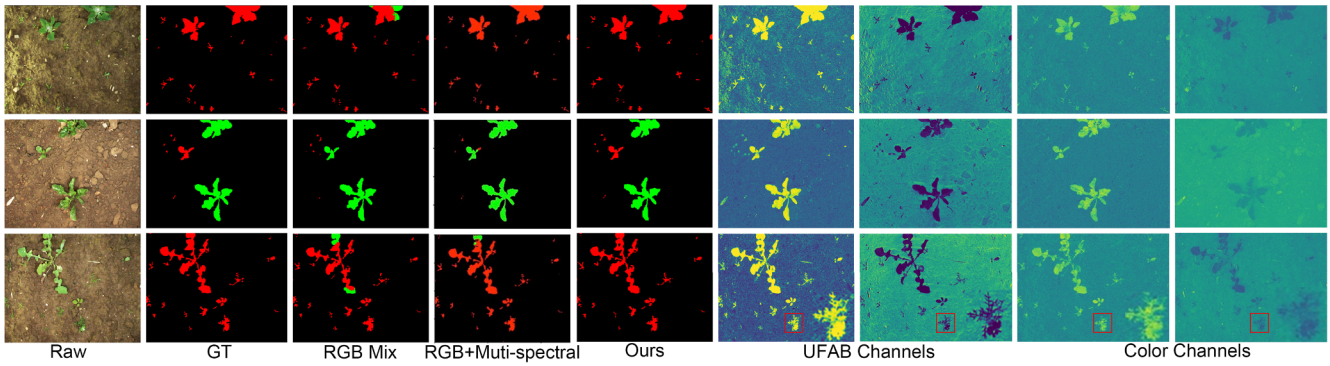

Fig. 9. We visualize the results of the UFAB-based color indices and the conventional approaches, Columns 3–5 show the segmentation performances, and the other 4 columns present the pairs of typical high-contrast channels for comparison. Our UFAB looks better at separating the weed/crop from background.

Table 6
The performance for conventional and UFAB features. The first row represents the baseline method with only RGB-NIR input features. The color index-based approach shows improvement, while integrating more color index-based features. Our method (UFAB) performs better at obtaining discriminate features for capturing weed objects.

Description	mIOU(%)	Bg(%)	Weed(%)	Crop(%)
Baseline	87.60	99.61	71.88	91.31
RGB-Mix	87.81	99.73	71.33	92.34
RGB + Multi-spectral	88.20	99.74	72.30	92.56
Ours	88.72	99.73	73.50	92.93

Table 7
Comparison of the seven models. All models use the same backbone network, except for *RSS**. *RSS**(Milioto et al., 2018) may have different training and validation data from ours, due to the incomplete description in the paper.

Model	mIOU(%)	Bg(%)	Weed(%)	Crop(%)	mPA(%)
PSPNet	81.86	99.59	56.38	89.61	87.10
UNet	86.94	99.69	70.62	90.51	92.27
SegNet	82.35	99.58	59.45	88.02	88.14
FCN8s	84.74	99.61	65.94	88.67	91.00
RSS∗	80.80	99.48	59.17	83.72	——
Ours	89.01	99.73	75.26	92.04	93.44

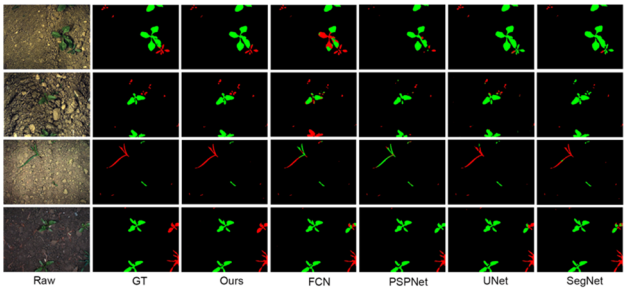

Fig. 10. We illustrate four test images, and compare the prediction results from five models. Note that each raw image in column 1 is captured in different background environment. The results in row 4 show that by capturing the local and global information, our method provides better performance.

interesting example. This frequently occurs in the mature period, where the crop has greater lamina, and covers the stem. The ground truth mask is wrongly annotated as background in the large crop area, but our model still tries to annotate correctly. We removed the wrong annotated labels in the mature period from the Bonn testing dataset, and reevaluated the performance. The mIOU accuracy was increased by around 1% from (89.01 to 90.22)%. It is evident that the mis-annotated labels have a strong effect on the performance.

Table 8
The performance on the Stuttgart dataset. Res50∗ is the improved Res50 backbone network that is integrated with HDC and DropBlock. The RSS∗ has different backbone structure; we directly use the experiment result from the paper (Milioto et al., 2018).

Model	Backbone	mIOU (%)
Ours	Res50	70.82
	Res50∗	72.94
PSPNet	Res50	51.41
	Res50∗	60.18
UNet	Res50	69.10
	Res50∗	70.24
SegNet	Res50	65.81
	Res50∗	67.84
FCN8s	Res50	56.90
	Res50∗	70.72
RSS∗	——	61.12

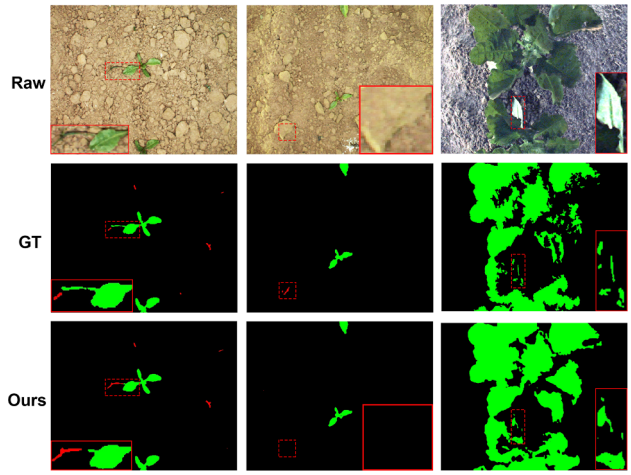

Fig. 11. The samples of mis-annotated labels. We zoom out the area of interest (dashed box), and compare with our predicted result. Best seen in color.

5. Conclusion

In this paper, we have presented an improved semantic segmentation network for crop/weed recognition. Our network is integrated with the hybrid dilated convolutional layer and DropBlock to enlarge the receptive field and learn robustness features. We introduce another two blocks, the bridge attention block and spatial pyramid refinement, to capture the global discriminative features, and register them with local precision. Moreover, the front-end universal function approximation block avoids the problem of manually selecting the proper color indices, and effectively generates the transformed color features to differentiate

plant pixels from the background. Our elaborate experiments show that the proposed network achieves outstanding performance compared to other conventional methods, and each component has notable potential to boost the segmentation accuracy. In the future, we will exploit the domain knowledge in weed detection, and model the network to learn more correlated spatial information.

CRediT authorship contribution statement

Jie You: Conceptualization, Methodology, Software, Writing - original draft. **Wei Liu:** Formal analysis, Investigation, Data curation. **Joonwhoan Lee:** Validation, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2019R1A6A1A09031717).

Appendix A. Supplementary material

Supplementary data associated with this article can be found online at https://github.com/kehuantiantang/A-DNN-based-Semantic-Segmentation-for-Detecting-Weed-and-Crop.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compag.2020.105750.

References

Ba, J., Mnih, V., Kavukcuoglu, K., 2014. Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755.
Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Machine Intell. 39, 2481–2495.
Burgos-Artizzu, X.P., Ribeiro, A., Guijarro, M., Pajares, G., 2011. Real-time image processing for crop/weed discrimination in maize fields. Comput. Electron. Agric. 75, 337–346.
Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.
Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Machine Intell. 40, 834–848.
Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.
Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L., 2016a. Attention to scale: Scale-aware semantic image segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649.
Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2016b. Encoder-decoder with atrous separable convolution for semantic image segmentation, in. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818.
Couprie, C., Farabet, C., Najman, L., LeCun, Y., 2013. Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572.
Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2012. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1915–1929.
Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. Int. J. Comput. Vision 59, 167–181.
Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154.
Ghiasi, G., Lin, T.Y., Le, Q.V., 2018. Dropblock: A regularization method for convolutional networks. Adv. Neural Informat. Process. Syst. 10727–10737.
Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D., 2015. Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623.
Gulli, A., Pal, S., 2017. Deep Learning with Keras. Packt Publishing Ltd.
Hague, T., Tillett, N., Wheeler, H., 2006. Automated crop and weed monitoring in widely spaced cereals. Precision Agric. 7, 21–32.
Hamuda, E., Glavin, M., Jones, E., 2016. A survey of image processing techniques for plant extraction and segmentation in the field. Comput. Electron. Agric. 125, 184–199.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.
Huang, H., Huang, Q., Krahenbuhl, P., 2018. Domain transfer through deep activation matching, in. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 590–605.
Hunt, E.R., Cavigelli, M., Daughtry, C.S., Mcmurtrey, J.E., Walthall, C.L., 2005. Evaluation of digital photography from model aircraft for remote sensing of crop biomass and nitrogen status. Precision Agric. 6, 359–378.
Jetley, S., Lord, N.A., Lee, N., Torr, P.H., 2018. Learn to pay attention. arXiv preprint arXiv:1804.02391.
Jiang, F., Pang, Y., Lee, T.N., Liu, C., 2019. Automatic object segmentation based on grabcut. In: Science and Information Conference. Springer, pp. 350–360.
Kataoka, T., Kaneko, T., Okamoto, H., Hata, S., 2003. Crop growth estimation system using machine vision. In: Proceedings 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003), vol. 2, IEEE. pp. b1079–b1083.
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Informat. Process. Syst. 1097–1105.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.
Lin, G., Milan, A., Shen, C., Reid, I., 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934.
Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.
Liu, W., Rabinovich, A., Berg, A.C., 2015. Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.
Lottes, P., Behley, J., Chebrolu, N., Milioto, A., Stachniss, C., 2018a. Joint stem detection and crop-weed classification for plant-specific treatment in precision farming. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 8233–8238.
Lottes, P., Behley, J., Chebrolu, N., Milioto, A., Stachniss, C., 2020. Robust joint stem detection and crop-weed classification using image sequences for plant-specific treatment in precision farming. J. Field Robot. 37, 20–34.
Lottes, P., Hoeferlin, M., Sander, S., Müter, M., Schulze, P., Stachniss, L.C., 2018b. An effective classification system for separating sugar beets and weeds for precision farming applications. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 5157–5163.
Lottes, P., Stachniss, C., 2017. Semi-supervised online visual crop and weed classification in precision farming exploiting plant arrangement. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 5155–5161.
Meyer, G.E., Neto, J.C., Jones, D.D., Hindman, T.W., 2004. Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. Comput. Electron. Agric. 42, 161–180.
Milioto, A., Lottes, P., Stachniss, C., 2018. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 2229–2235.
Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.
Pal, N.R., Pal, S.K., 1993. A review on image segmentation techniques. Pattern Recognit. 26, 1277–1294.
Pei, J.S., Mai, E.C., 2008. Constructing multilayer feedforward neural networks to approximate nonlinear functions in engineering mechanics applications. J. Appl. Mech. 75, 061002.
Pei, J.S., Mai, E.C., Wright, J.P., 2008. Mapping some functions and four arithmetic operations to multilayer feedforward neural networks. In: Health Monitoring of Structural and Biological Systems 2008. International Society for Optics and Photonics, pp. 693512.
Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large kernel matters–improve semantic segmentation by global convolutional network, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353–4361.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 234–241.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115, 211–252.
Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., Chellappa, R., 2018. Learning from synthetic data: Addressing domain shift for semantic segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3752–3761.
Shuai, B., Zuo, Z., Wang, B., Wang, G., 2017. Scene segmentation with dag-recurrent neural networks. IEEE Trans. Pattern Anal. Machine Intell. 40, 1480–1493.
Slaughter, D., Giles, D.K., Downey, D., 2008. Autonomous robotic weed control systems: A review. Comput. Electron. Agric. 61, 63–78.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017a. Residual attention network for image classification, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G., 2017b. Understanding convolution for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1451–1460.

Wang, X., Girshick, R., Gupta, A., He, K., 2017c. Non-local neural networks, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803.

Woebbecke, D.M., Meyer, G.E., Von Bargen, K., Mortensen, D., 1995. Color indices for weed identification under various soil, residue, and lighting conditions. Trans. ASAE 38, 259–269.

Woebbecke, D.M., Meyer, G.E., Von Bargen, K., Mortensen, D.A., 1993. Plant species identification, size, and enumeration using machine vision techniques on near-binary images. In: Optics in Agriculture and Forestry. International Society for Optics and Photonics, pp. 208–219.

Woo, S., Park, J., Lee, J.Y., So Kweon, I., 2018. Cbam: Convolutional block attention module, in. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19.

Xu, B., Wang, N., Chen, T., Li, M., 2015. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057.

Xue, J., Su, B., 2017. Significant remote sensing vegetation indices: A review of developments and applications. J. Sensors 2017.

Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K., 2018. Denseaspp for semantic segmentation in street scenes, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3684–3692.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Yu, F., Koltun, V., Funkhouser, T., 2017. Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 472–480.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A., 2018a. Context encoding for semantic segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7151–7160.

Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2018c. Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318.

Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T., 2018c. Fully convolutional adaptation networks for semantic segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6810–6818.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.