International Journal of Robotic Computing
Vol. 2, No. 1 (2020) 39-57

DOI: 10.35708/RC1869-126258

Crop and Weed Classification Using Pixel-wise Segmentation on Ground and Aerial Images

Mulham Fawakherji[1], Ali Youssef[1], Domenico D. Bloisi[2], Alberto Pretto[3], and Daniele Nardi[1]

[1] Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Rome, Italy
nardi@diag.uniroma1.it
[2] Department of Mathematics, Computer Science, and Economics, University of Basilicata, Potenza, Italy
domenico.bloisi@unibas.it
[3] IT+Robotics Srl, Padova, Italy
alberto.pretto@it-robotics.it

Received (11/10/2018)
Revised (06/18/2019)
Accepted (11/25/2019)

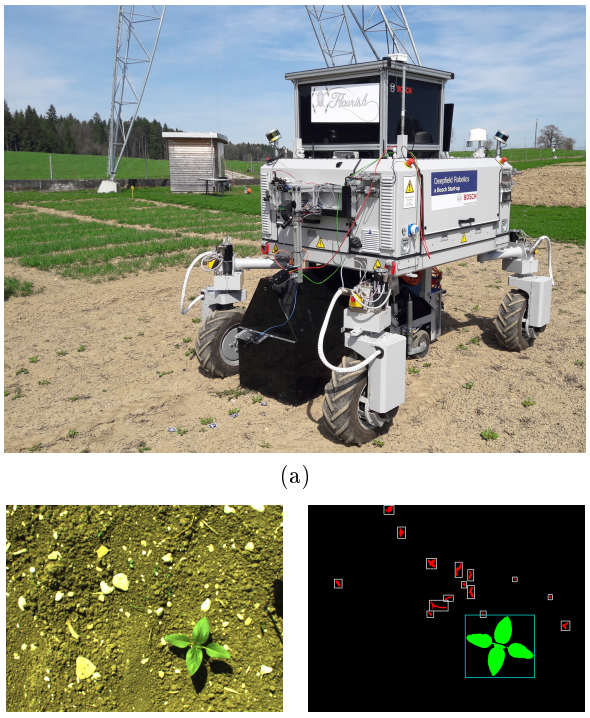
Artificial Intelligence (AI) is a key tool in agriculture for implementing sustainable strategies for weed control. In traditional weed control, the agro-chemical inputs are uniformly applied to the field, while innovative approaches using AI aim at minimizing the usage of chemical inputs thanks to local applications. In this paper, we focus on agricultural robotics systems that address the weeding problem by means of selective spraying or mechanical removal of the detected weeds. We present a set of deep learning based methods designed to enable a robot to efficiently perform an accurate weed/crop classification from RGB or RGB+NIR (Near Infrared) images. In particular, we use two Convolutional Neural Networks (CNNs) to simplify and speed up the training process. A first encoder-decoder segmentation network is designed to perform a "plant-type agnostic" segmentation between vegetation and soil. Each plant is hence classified between crop and weeds by using a second network, depending on the type of pipeline, for patch-level or pixel-level classification. We introduce also a third CNN, specifically designed for setups with limited resources, like in small UAVs (Unmanned Aerial Vehicles), that exploits the proposed encoder-decoder segmentation network to efficiently estimate crop/weeds local statistics. Quantitative experimental results, obtained using multiple publicly available datasets, demonstrate the effectiveness of the proposed approaches.

Keywords: Precision agriculture; crop/weed classification; image segmentation.

1 Introduction

Autonomous robotics applications for precision agriculture represent a concrete solution towards a sustainable agriculture and chemical treatments reduction [5]. The term *crop* defines the cultivated plant, while the term *weeds* defines unwanted plants that grow spontaneously in the field. Precision weed control is a challenging task that aims to reduce the amount of herbicides without compromising the quality of crops. Since achieving that manually is time-consuming and expensive, selective spraying or accurate mechanical removal of weeds are preferable options.

Autonomous robots equipped with automatic weed detection systems (e.g., Fig. 1a) can be used to improve the efficiency of precision farming techniques on weed control by modulating herbicide spraying appropriately to the level of weeds infestation. However, the great variety of crop and weeds shapes, size and colors,



(a) (b) (c)

Fig. 1: (a) The robot used to acquire some of the datasets used in the experiments. (b) An example of RGB image provided by the camera mounted on the robot. (c) Label mask with bounding boxes predicted by our approach for image (b). Crop pixels are colored in green while weed pixels are colored in red.

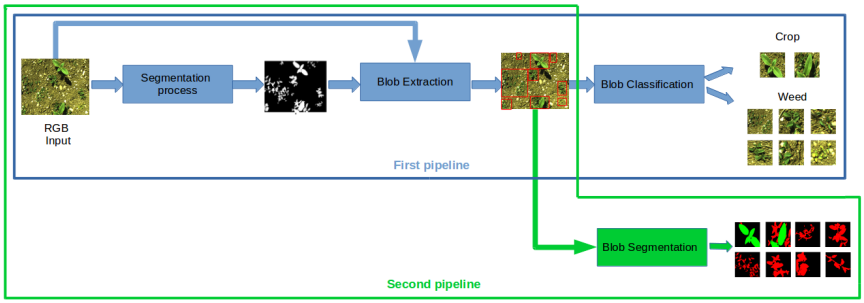


Fig. 2: The proposed three-steps crop/weed detection approaches. The two pipelines (blue and green boxes) share the first two steps. The first step is a binary pixel-wise segmentation (i.e., soil/plant) of the RGB input image. The second step concerns the extraction of the image patches to be classified. Crop/weed classification is carried out in the third step, by means of a patch-based classification CNN for the first pipeline, and a segmentation CNN applied to the extracted patches for the second pipeline.

together with the presence of overlaps between plants, makes the automatic image based crop/weeds classification problem (see Fig. 1b,c) a challenging task for autonomous farming robots [13]. Moreover, when using supervised methods, the capability to generalize of the trained models still remains an obstacle to employ farming robots in different farm conditions caused by environmental changes, different plants characteristics, and types of soil [16,12].

In this paper, we present two novel pipelines (Fig. 2) that combine a robust pixel-wise segmentation CNN, designed to be trained once to generalize the vegetation/soil segmentation problem, with a specialized crop/weeds classification network, designed to be trained for each type of cultivation with a specific, reduced size dataset. The aim of the proposed approaches is to reduce the limitations of CNNs in generalizing when a limited amount of data with pixel-wise annotations is available: pixel-wise labeling is in fact the bottleneck for most crop/weeds classification methods. Our methods relies on a robust binary segmentation that aims to be agnostic to plant species, so easily trainable also by using external, ready to use pixel-wise labeled datasets that possibly do not includes the target crop. Specifically, we use here a deep convolutional encoder-decoder architecture for robust background (soil) removal and the extraction of regions of interest (ROIs). The chosen network is based on the UNet architecture [19] with a modified VGG-16 encoder [23] followed by a binary pixel-wise classification layer. A coarse-to-fine classifier based on CNN is used to classify the extracted ROIs into crop and weeds. The classification between crop and weeds is obtained feeding a *classification* or a *segmentation* CNN (depending on the pipeline, see Fig. 2) with image patches (i.e., bounding boxes) enclosing plant instances extracted by using the obtained vegetation/soil segmentation mask. Both these architectures have the advantage of requiring smaller datasets comparing to conventional 3-classes soil/crop/weeds classification systems. They

Fig. 3: Left: a multirotor UAV. Right: the Jetson TX2 embedded board that can be easily installed on an UAV.

just require a small dataset to specialize on the target crop species. The generation of the specialized datasets is even more simple for the first pipeline (see the blue box in Fig. 2) since it is not required pixel-wise labeling, but only one label for each bounding box. We extensively tested the introduced approaches by using 4 different datasets, showing good classification results and generalization properties.

To further validate our approach and to provide a more efficient classification pipeline, we introduce also a third pipeline designed for setups with limited resources, like in small UAVs (e.g., Fig. 3 left). The third pipeline is shown in Fig. 4. It exploits the proposed encoder-decoder segmentation network and it is used to address the 3-classes segmentation problem with RGB-NIR images as input. We added a post-processing step that, starting from the segmented image, divides it into a fixed size grid and then computes for each cell the crop/weeds statistics (i.e., weed distribution and segmentation confidence). We implemented this pipeline on an NVIDIA Jetson TX2 embedded board (Fig. 3 right) and we evaluated it on real data coming from two datasets acquired by UAVs.

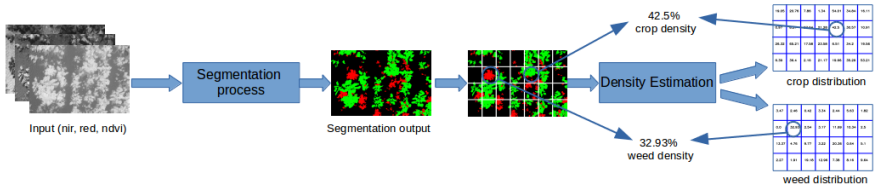


Fig. 4: Crop/weeds detection pipeline for limited resources systems.

In summary, the main contributions of this work are:

- A *context independent* background removal method that uses a CNN based pixel-wise segmentation to distinguish between soil and plants.
- Two accurate three-steps crop/weed detection pipelines based on the introduced segmentation network.
- A computationally efficient method for estimating the crop and weed distribution that exploits a modified version of the introduced segmentation network to be used on embedded GPU boards.

The reminder of the paper is organized as follows. Section 2 contains a discussion of similar approaches found in the literature. Section 3 describes the details of the proposed method, while Section 4 shows both qualitative and quantitative results obtained on publicly available data. Finally, conclusions are drawn in Section 5.

2 Related Work

The problem of vision based crop and weeds classification has been addressed in different ways. Handcrafted features are used, among others, in [3,7]. De Rainville *et al.* [3] present an unsupervised classification method based on morphological features extracted taking into account the spatial localization of vegetation in the field. Haug *et al.* [7] present a method to classify carrot plants and weeds from RGB and near-infrared (NIR) images that uses a background removal step based on the Normalized Difference Vegetation Index (NDVI) and a Random Forest classifier applied to features extracted at sparse pixel positions. The approach described in [7] has been extended in [13], where a plant arrangement prior is added to the features list used for classification, and tailored for UAVs (Unmanned Aerial Vehicles) applications in [14].

The adoption of deep CNNs in overcoming the limitations of handcrafted features has been explored, among others, in [9,18,12]. Fine-tuned pre-trained CNN models are used for plant classification of 44 different species in [9]. Potena *et al.* [18] proposes an on-line perception system for weed-crop classification that uses a cascade of two different CNNs. A shallow CNN performs vegetation detection, while a second, deeper CNN discriminates between weeds and crops. Encoder-decoder architectures such as the SegNet segmentation network [1] are used in [21,16,4]. In [21], a SegNet network is fed with three channels images that include the NIR channel, the red channel from the RGB image, and the NDVI map. A similar approach is exploited in [16], where 14 channels images that include several vegetation indices are used as input for a modified version of the SegNet network. Synthetic training datasets are used to train a SegNet network in [4] by randomizing the key features of the target environment (i.e., crop and weed species, type of soil, and light conditions). The fully convolutional network (FCN) proposed in [22] is employed in [12,11]. In [12], the authors exploits the crop arrangement as an additional source of information, by analyzing image sequences that cover a portion of the field. Class-wise stem detection and pixel-wise crop/weeds semantic segmentation is jointly addressed in [11]. Model

Algorithm 1 The proposed crop/weed classification algorithm

Input: RGB image I_{RGB}
Result: A set of classified blobs B_c
$M \leftarrow$ Segmentation of I_{RGB} using VGG-UNet
$C \leftarrow$ Contour-Extraction(M) ▷ set of contours belonging to connected regions
for i `in range len(`C`)` **do**
    $B_M[i] \leftarrow$ `BoundRect(`$C[i]$`)` ▷ $B_M[i]$ is the bounding box around the contour i
    $B_{RGB}[i] \leftarrow$ `(`$I_{RGB} \cap B_M[i]$`)` ▷ $B_{RGB}[i]$ is the corresponding bounding box from RGB image
    $B_c[i] \leftarrow$ (classify $B_{RGB}[i]$ using VGG-16 into weed or crop) ▷ for the first pipeline, or:
    $B_c[i] \leftarrow$ (segment $B_{RGB}[i]$ using VGG-UNet deep CNN into soil, weed or crop) ▷ for the second pipeline
end for

compression and mixtures of lightweight CNNs are exploited in [15] to learn from a very deep, pre-trained model a lighter model which allows real-time weed segmentation also for robots with limited computing power. Multi-spectral features and 3D surface features are exploited for plant classification in [24].

The approach described in this paper builds upon our previous work presented in [6]. Here, we introduce two new pipelines (see the blue box in Fig. 2 and Fig. 4) and we extend the experimental evaluation by using new datasets end by presenting additional results.

3 Methods

Our goal is to process an input RGB or multispectral image of the field to extract semantic information. In particular, we present three pipelines to obtain:

1. A 3-steps pixel-wise image segmentation into 3 classes (i.e., weed, crop, and soil) for the first two pipelines (see Fig. 2 and Sec. 3.1);
2. A 2-steps coarse-grained weed and crop density for the third pipeline (see Fig. 4 and Sec. 3.3).

The third pipeline is designed to reduce the computational burden while providing still accurate information with a lower resolution. We describe below each step involved in the three pipelines.

3.1 Pixel-wise Segmentation

The two pipelines of Fig. 2 take an RGB image as input and share the first two steps, i.e., vegetation segmentation and patches extraction, to achieve the classification goals exploiting two different approaches, i.e., on patch (Sec. 3.2) and pixel level (Sec. 3.2). All the steps of our full crop/weeds pixel-wise segmentation method are given in Algorithm 1.

Vegetation Segmentation To remove the background (i.e., the soil), we firstly apply a robust pixel-wise soil/plant segmentation of the RGB image in input. We use a modified version of the UNet semantic segmentation network

[19], which is composed by a contracting encoder along with a symmetric expanding decoder. In our implementation, the contracting path consists of a VGG-16 structure modified by removing the last fully connected layers and fine-tuning the other layers. The indices of spatial information in the pooling operations are spread through the expansive path, which contains a sequence of up-convolution operations of features encoded in the contracting path. The expanding decoder is designed with 4-convolutional layers, where each layer is composed of a batch normalization, 4-upsampling layers and a soft-max pixel-wise classifier. Between the contracting and expanding paths, there is a bottleneck consisting of two convolutional layers combined with batch normalization and a dropout activation function.

The lack of pixel-wise annotated datasets for each possible crop type and for different field conditions can lead to strong challenges in generalizing an end-to-end crop/weeds segmentation network. The goal of the first step is to obtain a robust binary segmentation mask that enables to generate blobs corresponding to vegetation pixels in the RGB image, so to simplify the following classification step. This is obtained by exploiting the similarities of various plants properties instead of differentiate between them. The idea is to train the first segmentation network by exploiting external, ready to use datasets coming from different contexts, containing different plants categories, types of fields, and captured under varying environmental conditions. This context-independent training possibly enables to avoid to pixel-wise annotate large amount of data acquired in the target filed, an operation that usually requires a lot of manual work.

3.2 Patches Extraction

The second step concerns the extraction of the image ROIs to be classified. This is obtained by extracting the vegetation blobs contained in the binary mask generated during the segmentation process. In this stage, the input consists in the original RGB image plus the binary mask generated during the segmentation process. A dilation operator is applied to the binary mask to gradually increase the boundaries of the foreground regions (i.e., the areas containing vegetation pixels) to reduce the holes between those regions. Then, the connected blobs from the dilated mask are extracted, and a bounding box for each blob is determined. Finally, a set of patches from the original RGB image corresponding to the bounding boxes is generated.

Patch-Wise Classification A deep CNN for crop/weed classification is employed for patch-wise classification. The image patches identified in the previous step are fed to the CNN classifier, which is based on a fine-tuned model of VGG-16 exploiting the ability of deep CNN in object classification. The VGG-16 network architecture for object classification is used as encoder. The network consists of 13-convolutional layers with a kernel of 3×3. A max-pooling operation with a kernel of 2×2 with a stride of 2 for down-sampling. Batch normalization and a ReLU activation function are used too. This step just requires a training dataset that includes labeled patches with positive and negative examples of the target crop. The annotation of such training dataset just requires to specify a label for each image that is a much faster operation than a pixel-wise annotation.

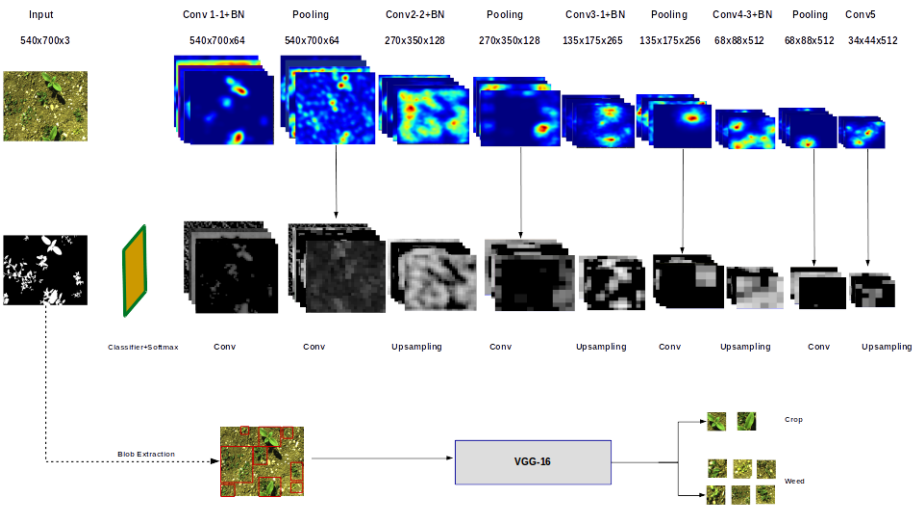


Fig. 5: Details of the architecture of the first pipeline.

Fig. 5 shows the details of the first pipeline. To create the figure, we have used a fine-tuned VGG-16 encoder model at early step of the training, showing randomly picked up filters to illustrate the ability of the network to learn weights based on neurons responses to image pixels (e.g., soil/plant pixel).

Pixel-Wise Classification In this step, we classify each pixel in the image patches identified in the previous step into one of three classes (soil, weed, and crop). To perform the segmentation, we used the same deep CNN architecture that is used for vegetation segmentation with 128×128 input size and three classes output.

3.3 Weeds Distribution Estimation

The goal of the pipeline shown in Fig. 4 is to process images, e.g., acquired by a limited resources system like a small UAV, to quickly extract high-level information about the field, like weed and crop density, instead of finding the exact location for weed inside the images. This high-level information can be sent to a ground robot equipped with more powerful computational resources to perform a fine-grained pixel-wise segmentation (e.g., Sec. 3.1) and to perform selective spraying. We first train a CNN to build a model able to detect crop and weed in the images, then we compute the distribution of the weed and the crop with a confidence based on the output of the trained model.

Crop/Weed Segmentation To perform pixel-wise segmentation of the input image into 3 classes (i.e., weed, crop, and soil), we use a modified version of the UNet semantic segmentation network [19], very similar to the one introduced in Sec. 3.1, filled with one or three channels e.g., NIR, red, and/or the Normalized Difference Vegetation Index (NDVI) [20].

Crop/Weed Distribution Estimation To compute the distribution of weed and crop inside the image, the mask generated during the segmentation process is divided using a fixed size grid (28 cells) based on the segmentation output size. Then, we compute for each cell the crop/weeds statistics (e.g., weed distribution and segmentation confidence). The distribution for a specific class in each cell is obtained by computing the number of pixels of that class inside the cell (Kc) divided by the total number of pixel inside the cell (see Eq. 1).

$$D_c = \frac{K_c}{\sum_{i=1}^{n} K_i} \tag{1}$$

where D_c represent the distribution of class c, n is the number of classes, and K_c is the total number of pixels belonging to class c inside the cell. For each cell, we also compute a per-class confidence $MCon_c$ as the average of the class probabilities $P(\cdot)$ provided by the network of Fig. 4 for all the cell pixels X_i that are classified with class c, i.e.:

$$MCon_c = \frac{\sum_{i=1}^{K_c} P(X_i = c)}{K_c} \tag{2}$$

4 Experimental Results

Experiments are organized into four case studies:

1. The first one (see Sec. 4.3) aims to demonstrate the performance of different deep CNN architecture on pixel-wise segmentation in order to classify pixels in image into two (soil and vegetation) or three classes (soil, crop, and weed).
2. In the second experiment (see Sec. 4.4), we study the effect of supporting the input of our deep CNN (described in Sec. 3.1) by increasing the number of input channels by a set of vegetation indices.
3. The third experiment (see Sec. 4.5) evaluates the performance of the two full pipelines described in Sec. 3.1, where the first one performs patch classification after background removal and blob extraction, where the second one performs pixel-wise segmentation on the extracted patches.
4. The last experiment (see Sec. 4.6) aims at measuring the accuracy of the pipeline shown in Fig. 4.

4.1 Datasets

In our experiments, we have used six datasets, some of them publicly available.

- *Sunflowers*: 500 images acquired in a sunflowers field by a custom-built agricultural field robot;
- *SugarBeets*: 10000 sugar beets images coming from the "Sugar Beets 2016" datasets [2];
- *Carrots*: 60 images acquired on a commercial organic carrots farm [8];
- *Stuttgart*: 200 sugar beets coming from the so-called Stuttgart dataset;
- *UAV1*: 155 multispectral images (NIR, Red) plus the NDVI index acquired from an UAV in a sugar beets field;

– *UAV2*: 75 multispectral images (NIR, Red) plus the NDVI index acquired from an UAV in a corn field.

All datasets are labeled with three classes (i.e., soil, crop and weed). Ground truth annotations consist in binary masks generated via manual segmentation. An intensity value of 1 in the binary masks corresponds to the segmented crop, 2 to segmented weeds, and pixels with 0 value correspond to the background soil.

4.2 Evaluation Metrics

For quantitatively evaluate the results, we used the following three metrics commonly used in the literature [17].

Mean Intersection-Over-Union (mIOU):

$$mIOU = \frac{1}{C} \sum_{j=1}^{C} \frac{TP_j}{TP_j + FP_j + FN_j} \tag{3}$$

Recall:

$$Recall_j = \frac{TP_j}{TP_j + FN_j} \tag{4}$$

Precision:

$$Precision_j = \frac{TP_j}{TP_j + FP_j} \tag{5}$$

where TP stands for True Positive, FP for False Positive, FN for False Negative, and C is the total number of classes.

4.3 Architecture Evaluation

The first experiment aims to demonstrate the performance of different deep CNN architectures on pixel-wise segmentation in order to classify pixels in image into three classes, namely soil, crop, and weed. Then, we measure the performance of the same networks on the background removal problem, i.e., pixel classification into only two classes: soil and plants. For this experiment, we use only RGB images as input to:

1. SegNet [1] based on VGG-16 encoder;
2. UNet [19];
3. UNet based on VGG-16 decoder (VGG-UNet);
4. BonNet [17];
5. Fully connected network FCN8 [10].

The training dataset is made of a set of 500 sunflowers images. Data augmentation was performed using rotations, horizontal and vertical flipping, and zooming. The final dataset was composed by 2000 images (see Fig. 6).

VGG-UNet has been trained by initializing the encoder (VGG-16) with the weights taken from training the VGG-16 on the ImageNet dataset, then we trained the whole network using Stochastic Gradient Descent (SGD) with a fixed learning rate of $1 \cdot e^{-4}$ and a momentum of 0.90. The parameters of the

Table 1: Quantitative results showing *mIOU* obtained by different networks architectures on the *Sunflowers* dataset

Architecture	**3-classes**	**2-classes**
VGG-SegNet	0.68	0.90
UNet	0.62	0.90
Bonnet	**0.80**	0.90
FCN8	0.31	0.45
VGG-UNet	0.64	**0.91**

network are updated in a way that cross entropy loss is reduced. Mini-batches composed by one image were used for training. It is worth noticing that, the dataset used for the training procedure does not present all the challenges introduced when dealing with a real-world field, because it does not contain data captured with different field conditions and at different stages of plant level. For this reason, in order to properly evaluate our approach, we used for testing 1000 images coming from the *SugarBeets* dataset [2] and other 60 images from the *Carrots* dataset. It is important to note that *Carrots* and *SugarBeets* datasets were not included in fine tuning VGG-16 encoder and were used only to evaluate the capability of VGG-UNet to generalize. Table 1 shows the results obtained by using different network architectures on the *Sunflowers* dataset. When considering only two classes, namely soil/plants, the VGG-UNet approach outperforms the other tested approaches. Fig. 7 shows qualitative results.

4.4 Multi-Channel Architecture Evaluation

To improve the segmentation performance of the VGG-UNet architecture, we increase the number of input channels by a set of vegetation indices: Excess Green (ExG), Excess Red (ExR), Color Index of Vegetation Extraction (CIVE), and Normalized Difference Index (NDI). In addition, we use the HSV (hue, saturation, value) representation of the input image, concatenating all those

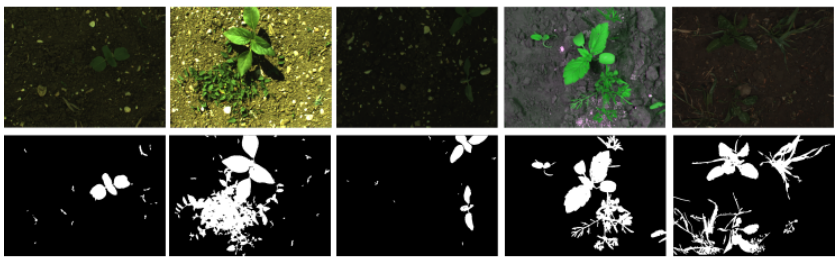


Fig. 6: Samples from the dataset used for training and testing. The first row contains the RGB images in input, while the second row shows the ground truth masks. In the first, second, and third columns sunflowers images taken under different lighting condition and different age of growth are shown. The fourth column refers to the organic *Carrots* dataset and the fifth column refers to the *SugarBeets* dataset.

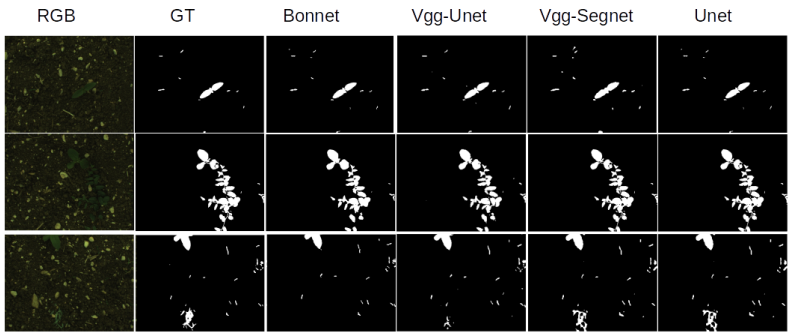


Fig. 7: Qualitative results achieved by different CNN structures. First column RGB images, second column ground truth mask, third, forth, fifth, and sixth prediction from Bonnet, VGG-UNet,VGG-Segnet, and UNet.

representations along with the input RGB image to form a multi-channel input volume.

For training our deep CNN, we have used a dataset consisting of the 80% of the images in the *SugarBeets* dataset. For testing, we used a dataset consisting of the remaining 20% of images from the *SugarBeets* dataset, plus all the images in the *Sunflowers*, *Carrots*, and *Stuttgart* datasets. Results for the 2 classes segmentation problem are reported in Table 2 (RGB input) and Table 3 (multi-channel input), while results for the 3 classes segmentation problem are reported in Table 4 (RGB input) and Table 5, respectively. Some qualitative results are reported in Fig. 8. It is noteworthy to underline that, this simple enrichment of the input has produced a substantial improvement in the results in almost all tests. Those indicators proved their ability to naturally segment vegetation and they do not present high sensitivity to soil types or weather condition, as reported in [16]

Table 2: Quantitative results obtained by VGG-UNet architecture for 2 classes segmentation on different datasets with RGB input.

Dataset	**mIOU**	**soil recall**	**plant recall**	**soil pre**	**plant pre**
SugarBeets	0,93	0.99	0.96	0.99	0.90
Stuttgart	0,66	0.98	0.59	0.99	0.37
Carrots	0,64	0.96	0.43	0.94	0.55
Sunflowers	0,63	0.95	0.39	0.96	0.53

Table 3: Quantitative results obtained by VGG-UNet architecture for 2 classes segmentation on different datasets with multi-channel input.

Dataset	mIOU	soil recall	plant recall	plant pre	plant pre
SugarBeets	0,92	0.99	0.98	0.99	0.86
Stuttgart	0,85	0.99	0.89	0.99	0.77
Carrots	0,67	0.94	0.75	0.99	0.60
Sunflowers	0,70	0.99	0.78	0.94	0.67

Table 4: Quantitative results obtained by VGG-UNet architecture for 3 classes segmentation on different datasets with RGB input.

Dataset	mIOU	soil recall	crop recall	weed recall	soil pre	crop pre	weed pre
SugarBeets	0,71	0.99	0.94	0.80	0.95	0.90	0.68
Stuttgart	0,45	0.98	0.66	0.68	0.97	0.36	0.34
Carrots	0,35	0.923	0.36	0.29	0.94	0.43	0.32
Sunflowers	0,39	0.92	0.33	0.37	0.96	0.32	0.29

Table 5: Quantitative results obtained by VGG-UNet architecture for 3 classes segmentation on different datasets with multi-channel input .

Dataset	mIOU	soil recall	crop recall	weed recall	soil pre	crop pre	weed pre
SugarBeets	0,75	0.998	0.94	0.80	0.99	0.92	0.70
Stuttgart	0,60	0.984	0.75	0.71	0.90	0.63	0.59
Carrots	0,40	0.93	0.43	0.38	0.98	0.32	0.37
Sunflowers	0.41	0.95	0.39	0.37	0.97	0.49	0.40

4.5 Pixel-wise Segmentation Evaluation

The third experiment aims to evaluate the performance of the two pipelines, the first one depending on blobs classification after background removal, the second one being based on patch segmentation after background removal.

The evaluation of the two approaches based on object-wise classification accuracy are given in the form of two confusion matrices. The confusion matrix for the first approach is shown in Fig. 9, while the confusion matrix for the second approach is shown in Fig. 10.

For the first approach (i.e., background removal plus classification), the result for correctly detected crops was 87%, while the 13% of crop was detected as weed and 22% of weed was detected as crop. This is mainly due to the overlapping problem between weed and crop. The 32% of soil detected as weed is due to inaccuracies in the binary masks coming from the segmentation process. The dilation operation carried out to increase the boundaries of the foreground regions during the blob detection process increases the number of soil pixels included in weeds blobs.

In the second approach (i.e., background removal plus segmentation), we overcome the problem of overlapping between weeds and crop inside the blob, as can be noted looking at the confusion matrix in Fig. 10. Just 2% of crop

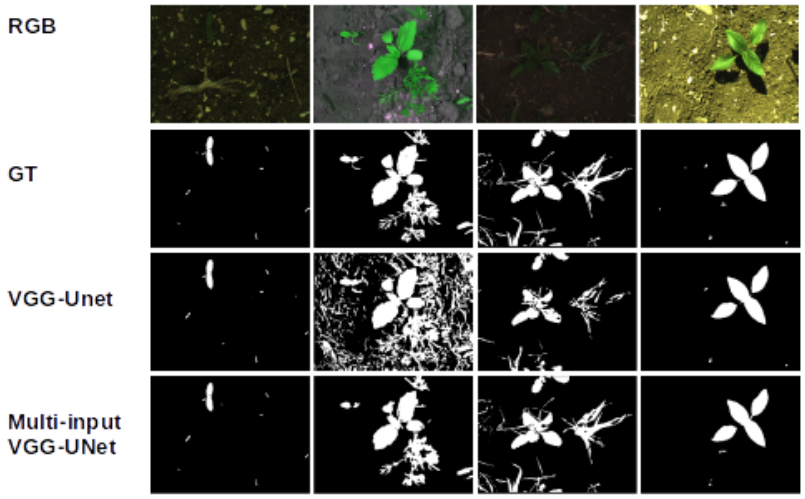


Fig. 8: Qualitative results obtained by VGG-UNet and multi-channel VGG-UNet.

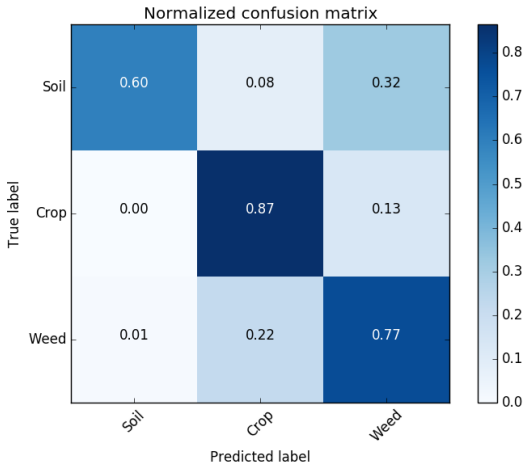


Fig. 9: Confusion matrix obtained by evaluating the first full pipeline (i.e., background removal plus classification).

was detected as weed and 3% of weed was detected as crop. This is due to the fact that, in the first approach, we classify the whole blob into one of two class types (i.e., crop or weed), while in the second approach each pixel in the blob is classified into one of the two class types (i.e., crop or weed). Qualitative results are shown in Fig. 11.

4.6 Weeds Distribution Evaluation

In the last experiment, we have used the two datasets *UAV1* and *UAV2* (see Sec. 4.1 and Fig. 12). For the first dataset, we took 100 images and we applied

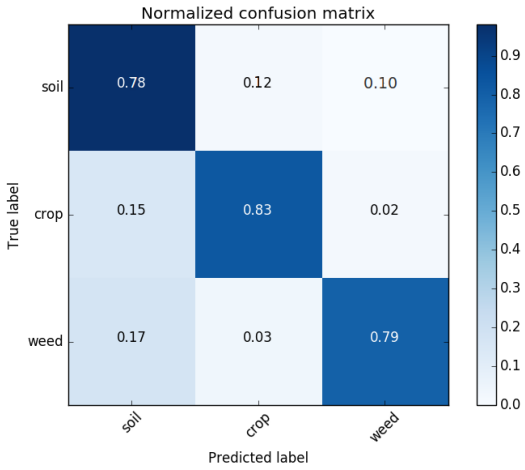


Fig. 10: Confusion matrix obtained by evaluating the second full pipeline (i.e., background removal plus segmentation).

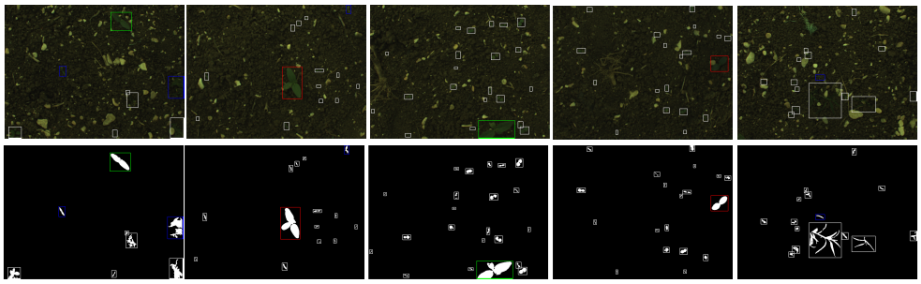


Fig. 11: Qualitative results achieved using segmentation plus classification. White boxes denote true positive samples (weed), green boxes true negative samples (crop), blue boxes false positive samples, and red false negative samples.

to them a data augmentation process. In particular, we exploited rotations and horizontal and vertical flipping to create an augmented training dataset composed by 420 images. The remaining 55 images from the original dataset were used for testing purposes. The second dataset (*UAV2*) was used for testing only. We evaluated the generalization capability of the net for classifying the pixels in two classes, namely soil and plant.

Network Training. We trained the proposed VGG-UNet by initializing the encoder (VGG-16) with the weights taken from training the VGG-16 on the ImageNet dataset, then we trained the whole network using Stochastic Gradient Descent (SGD) with a fixed learning rate of $1 \cdot e^{-4}$ and a momentum of 0.90. The parameters of the network are updated in a way that cross entropy loss is reduced. Mini-batches composed by one image was used for training.

Qualitative results of using different input types for pixel-wise segmentation into soil, weed, and crop on sugar beet dataset are shown in the first row of Fig. 13. The segmentation performance of VGG-UNet decreased when we used Red

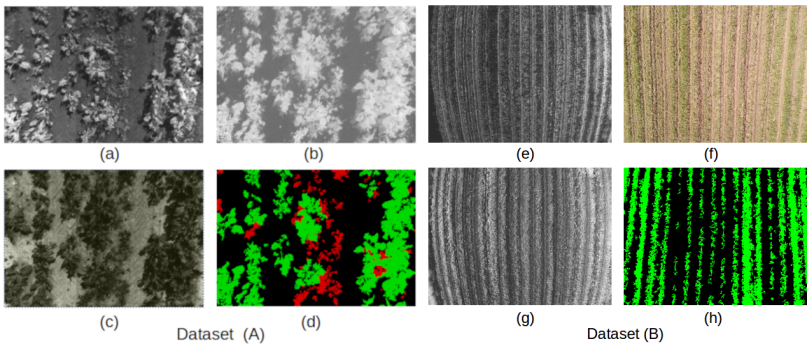


Fig. 12: Image samples from the datasets *UAV1* and *UAV2*. (a), (b), (c), and (d) represent NIR, NDVI, RED, and ground truth from sugar beet dataset. (e), (f), (g), and (h) represent NIR, RGB, RED, and ground truth from corn dataset.

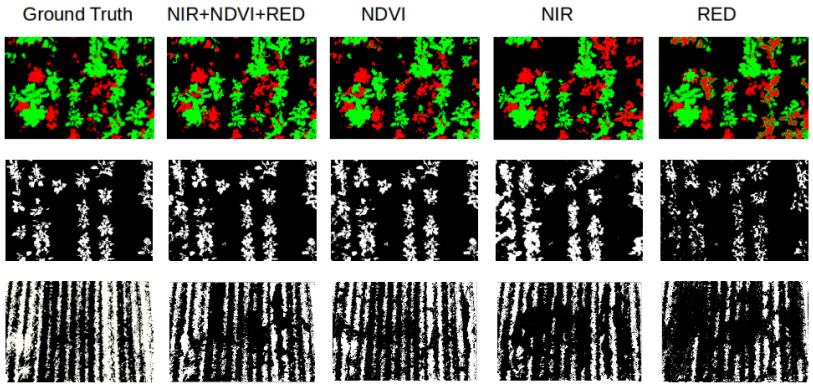


Fig. 13: Qualitative results achieved by CNN with different input type. First and second rows represent the output for 3 classes segmentation on sugar beet dataset, the third row 2 classes segmentation output on corn dataset.

channel alone as input. The best results was achieved using the three channels together as input, while for pixel-wise segmentation into soil, and plant sugar beet dataset the best result for NDVI channel. Table 6 shows the quantitative results obtained by the used VGG-UNet architecture. We have performed also a quantitative comparison between the used VGG-UNet architecture and the SegNet network. Table 7 contains the results for the comparison. VGG-UNet performs slightly better on the sugar beet dataset.

The final output of our approach is shown in Fig. 14, where the first row shows the segmentation output, the second and third rows show the crop and weed distribution. The number in each cell describes the crop distribution (in the second row) and the weed distribution (in the third row). The green circles show the high density of crop, while the red ones show the high density of weed. The performance of VGG-UNet architecture has been tested on an embedded GPU board, namely Jetson TX2, and the processing time per image was 0.6

Table 6: Quantitative results illustrating mean accuracy obtained by VGG-UNet using different input on the sugar beets and corn datasets.

	3-classes	**2-classes**	
Input	sugar beet	sugar beet	corn
Red	0.84	0.91	0.73
NIR	0.90	0.96	0.85
NDVI	0.93	**0.99**	**0.92**
RED+NIR+NDVI	**0.95**	0.98	0.88

Table 7: Quantitative results illustrating mean accuracy obtained by SegNet and VGG-UNet architecture on the sugar beet dataset.

Architecture	**3-classes** RED+NIR+NDVI input	**2-classes** NDVI input
SegNet	0.93	0.98
VGG-UNet	**0.95**	**0.99**

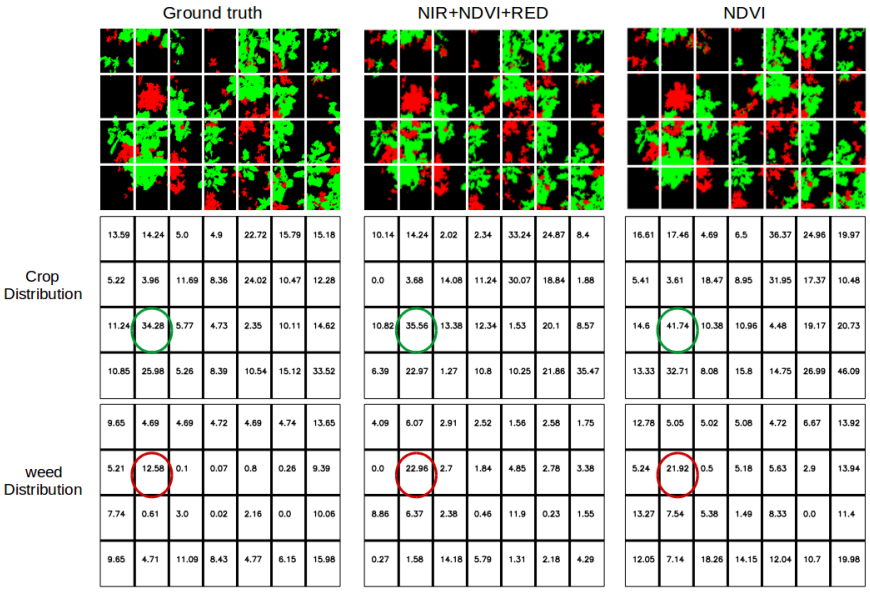


Fig. 14: Results from computing weed and crop distribution with different input types. The first column represent the ground truth. The second and third columns shows the prediction output from VGG-UNet when using NIR+NDVI+RED and NDVI as input. The second and third rows represents crop and weed distribution.

seconds when we use three channels together as input, and around 0.2 seconds when we used each channel alone as input.

5 Conclusions

In this paper, we have described two novel pipelines that combine a robust pixel-wise segmentation CNN, designed to be trained once to generalize the vegetation/soil segmentation problem, with a specialized crop/weeds classification network, designed to be trained for each type of cultivation with a specific, reduced size dataset. Our goal is to reduce the limitations of CNNs in generalizing when a limited amount of data with pixel-wise annotations is available.

Starting from the consideration that pixel-wise labeling is the bottleneck for most crop/weeds classification methods, we adopt an approach based on a robust binary segmentation to be agnostic with respect to plant species. Our network is trainable also by using external, ready to use pixel-wise labeled datasets that possibly do not includes the target crop. Specifically, we use a deep convolutional encoder-decoder architecture for robust background (soil) removal and the extraction of regions of interest (ROIs).

A coarse-to-fine classifier based on CNN is used to classify the extracted ROIs into crop and weeds. We describe two different methods for obtaining the classification between crop and weeds, i.e., by feeding a *classification* or a *segmentation* CNN with image patches (i.e., bounding boxes) enclosing plant instances extracted by using the obtained vegetation/soil segmentation mask. Both the presented architectures have the advantage of requiring smaller datasets comparing to conventional 3-classes soil/crop/weeds classification systems. They just require a small dataset to specialize on the target crop species.

We extensively tested the introduced approaches by using four different datasets, showing good classification results and generalization properties.

As future directions, we aim to insert between the segmentation and classification steps an automatic alignment process to improve the classification accuracy of our pipeline.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 (2015)
2. Chebrolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., Stachniss, C.: Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. The International Journal of Robotics Research (2017)
3. De Rainville, F.M., Durand, A., Fortin, F.A., Tanguy, K., Maldague, X., Panneton, B., Simard, M.J.: Bayesian classification and unsupervised learning for isolating weeds in row crops. Pattern Analysis and Applications **17**(2), 401–414 (2014)
4. Di Cicco, M., Potena, C., Grisetti, G., Pretto, A.: Automatic model based dataset generation for fast and accurate crop and weeds detection. In: IROS. pp. 5188–5195 (2017)
5. Duckett, T., Pearson, S., Blackmore, S., Grieve, B., Wilson, P., Gill, H., Hunter, A., Georgilas, I.: Agricultural Robotics: The Future of Robotic Agriculture. UK-RAS White Papers, UK-RAS Network (2018)
6. Fawakherji, M., Youssef, A., Bloisi, D.D., Pretto, A., Nardi, D.: Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation. In: Proc. of the IEEE International Conference on Robotic Computing (IRC) (2019). https://doi.org/10.1109/IRC.2019.00029

7. Haug, S., Michaels, A., Biber, P., Ostermann, J.: Plant classification system for crop/weed discrimination without segmentation. In: WACV. pp. 1142–1149. IEEE (2014)
8. Haug, S., Ostermann, J.: A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In: Computer Vision - ECCV 2014 Workshops. pp. 105–116 (2015)
9. Lee, S.H., Chan, C.S., Wilkin, P., Remagnino, P.: Deep-plant: Plant identification with convolutional neural networks. In: ICIP. pp. 452–456. IEEE (2015)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
11. Lottes, P., Behley, J., Chebrolu, N., Milioto, A., Stachniss, C.: Joint stem detection and crop-weed classification for plant-specific treatment in precision farming. arXiv preprint arXiv:1806.03413 (2018)
12. Lottes, P., Behley, J., Milioto, A., Stachniss, C.: Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. IEEE Robotics and Automation Letters (RA-L) **3**, 3097–3104 (2018)
13. Lottes, P., Hoeferlin, M., Sander, S., Müter, M., Schulze, P., Stachniss, C.: An effective classification system for separating sugar beets and weeds for precision farming applications. In: ICRA. pp. 5157–5163 (2016)
14. Lottes, P., Khanna, R., Pfeifer, J., Siegwart, R., Stachniss, C.: Uav-based crop and weed classification for smart farming. In: ICRA. pp. 3024–3031 (2017)
15. McCool, C., Perez, T., Upcroft, B.: Mixtures of lightweight deep convolutional neural networks: Applied to agricultural robotics. IEEE Robotics and Automation Letters **2**(3), 1344–1351 (2017)
16. Milioto, A., Lottes, P., Stachniss, C.: Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In: ICRA. pp. 2229–2235 (2018)
17. Milioto, A., Stachniss, C.: Bonnet: An open-source training and deployment framework for semantic segmentation in robotics using cnns. arXiv preprint arXiv:1802.08960 (2018)
18. Potena, C., Nardi, D., Pretto, A.: Fast and accurate crop and weed identification with summarized train sets for precision agriculture. In: International Conference on Intelligent Autonomous Systems. pp. 105–121 (2016)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015)
20. Rouse, Jr., J.W., Haas, R.H., Schell, J.A., Deering, D.W.: Monitoring vegetation systems in the great plains with ERTS. In: Proc. of the 3rd Earth Resource Technology Satellite (ERTS) Symposium. vol. 1 (1974)
21. Sa, I., Chen, Z., Popović, M., Khanna, R., Liebisch, F., Nieto, J., Siegwart, R.: weednet: Dense semantic weed classification using multispectral images and mav for smart farming. IEEE Robotics and Automation Letters **3**(1), 588–595 (2018)
22. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Strothmann, W., Ruckelshausen, A., Hertzberg, J., Scholz, C., Langsenkamp, F.: Plant classification with in-field-labeling for crop/weed discrimination using spectral features and 3d surface features from a multi-wavelength laser line profile system. Computers and Electronics in Agriculture **134**, 79–93 (2017)