Computers and Electronics in Agriculture 194 (2022) 106719

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Deep learning-based segmentation of multiple species of weeds and corn crop using synthetic and real image datasets

Artzai Picon [a,c], Miguel G. San-Emeterio [a], Arantza Bereciartua-Perez [a], Christian Klukas [b], Till Eggers [b], Ramon Navarra-Mestre [b,*]

[a] TECNALIA, Basque Research and Technology Alliance (BRTA), Parque Tecnológico de Bizkaia, C/ Geldo. Edificio 700, E-48160 Derio, Bizkaia, Spain
[b] BASF SE, Speyererstrasse 2, 67117 Limburgerhof, Germany
[c] University of the Basque Country, Bilbao Engineering Faculty, Edificio I. Plaza Ingeniero Torres Quevedo, 48013 Bilbao, Spain

ARTICLE INFO

Keywords:
convolutional neural network
deep learning
multi-weed classification
plant safety digitalization
weed semantic segmentation

ABSTRACT

Weeds compete with productive crops for soil, nutrients and sunlight and are therefore a major contributor to crop yield loss, which is why safer and more effective herbicide products are continually being developed. Digital evaluation tools to automate and homogenize field measurements are of vital importance to accelerate their development. However, the development of these tools requires the generation of semantic segmentation datasets, which is a complex, time-consuming and not easily affordable task.

In this paper, we present a deep learning segmentation model that is able to distinguish between different plant species at the pixel level. First, we have generated three extensive datasets targeting one crop species (*Zea mays*), three grass species (*Setaria verticillata, Digitaria sanguinalis, Echinochloa crus-galli*) and three broadleaf species (*Abutilon theophrasti, Chenopodium albums, Amaranthus retroflexus*). The first dataset consists of real field images that were manually annotated. The second dataset is composed of images of plots where only one species is present at a time and the third type of dataset was synthetically generated from images of individual plants mimicking the distribution of real field images.

Second, we have proposed a semantic segmentation architecture by extending a PSPNet architecture with an auxiliary classification loss to aid model convergence. Our results show that the network performance increases when supplementing the real field image dataset with the other types of datasets without increasing the manual annotation effort. More specifically, the use of the real field dataset obtains a Dice-Søensen Coefficient (DSC) score of 25.32. This performance increases when this dataset is combined with the single-species class dataset (DSC=47.97) or the synthetic dataset (DSC=45.20). As for the proposed model, the ablation method shows that by removing the proposed auxiliary classification loss, the segmentation performance decreases (DSC=45.96) compared to the proposed architecture method (DSC=47.97).

The proposed method shows better performance than the current state of the art. In addition, the use of proposed single-species or synthetic datasets can double the performance of the algorithm than when using real datasets without additional manual annotation effort.

1. Introduction

Presence of weed communities in crop fields has a negative impact van Heemst (1985) on crop yield as they compete with crops for soil, nutrients and sunlight, causing crops to grow slower and smaller. Therefore, new herbicides are continuously being developed to provide greater efficiency and safety. The development of these herbicides requires an extensive set of trials to quantify their actual performance under different conditions. Currently, these trials are conducted by manual visual assessments that rely on the expertise of field researchers. In these trials, quantifying the coverage of the different species, estimating their growing stage and measuring the impact of the herbicide on the crop and on the weed species are key elements in a visual assessment. However, these estimates are affected by human variability, so the automation of these tasks has been a topic of interest to researchers in recent years. In recent years, the computer vision community has

* Corresponding author.
E-mail address: ramon.navarra-mestre@basf.com (R. Navarra-Mestre).

https://doi.org/10.1016/j.compag.2022.106719
Received 2 August 2021; Received in revised form 8 January 2022; Accepted 14 January 2022
Available online 7 February 2022

tackled the task of weed identification. The main issues for computer vision algorithms are related to the natural conditions of field images and the living nature of plants. The main challenges for the development of successful models are the different lighting conditions, the overlapping of leaves, the inhomogeneity of weed patch densities, the different scales between different images, and the fact that leaves of different species can be very similar to each other (McCarthy et al. (2010)). In addition, the large extent of plantings often poses a computational time problem. More recently, deep learning has proven to be a disruptive technology for agronomy, overcoming the limitations of classical methods Picon et al. (2020), as CNNs enable richer expressive power to represent image content (Kamilaris and Prenafeta-Boldú (2018,)). Thus, CNNs have achieved considerable success in plant species classification (Dyrmann et al. (2016a,)), crop disease detection (Picon et al. (2018, 2019)), plant segmentation (Romera-Paredes and Torr (2016)), and weed characterization (Teimouri et al. (2018)), among other applications. However, CNNs have some disadvantages. One of them is the large number of manually annotated images required to generate a model (Medela et al. (2019, 2020,)). Manually annotating the required images is a time-consuming and sometimes unfeasible task. In 2016, Mortensen et al. (2016) presented their seminal work on semantic segmentation of crops and weeds using deep learning. They obtained a pixel accuracy of 0.79 for segmentation of different crop species, with no success in detecting weed species. However, their results indicated the great potential of using CNNs for the task of crop and weed identification. In subsequent studies (Dyrmann et al. (2016b)), they improved the method to distinguish corn from 23 different weed species by correctly labeling pixels as "corn" or "weed" in a two-class classification problem. They obtained a per-pixel accuracy of 0.94. These techniques required extensive and good quality manual annotation to perform supervised training. In 2018 Sa et al. (2018) achieved an F1 score of 0.80 on crop and weed segmentation with their modified VGG-16 called *weedNet* and Milioto et al. (2018) achieved a Intersection-Over-Union metric of 0.81 on per-pixel classification of crop, weed, and soil. Previous work does not address the segmentation of multi-species weeds, which is a more complex and demanding problem, especially since different plant species can only be distinguished by specific and subtle taxonomic keys that may not even always be visible in an image. In this paper we address multispecies segmentation using a combined approach. On the one hand, we propose a novel methodology to include synthetic and single-species datasets to reduce the need for manual annotation and, on the other hand, we propose a novel architecture to successfully perform multispecies semantic segmentation.

2. Dataset generation methodologies

Semantic segmentation models require complex and time-consuming manually annotated images, which greatly increases the difficulty of generating a functional dataset. In this paper, we propose three different methodologies to minimize the need for manually annotating images. First, we introduce some improvements to speed up the manual annotation of real image datasets (dataset A). In addition, we propose two different methods to generate datasets that do not require manual annotation: a) A method to generate synthetic datasets based on single plant images (dataset B) and b) a methodology to generate real field datasets consisting of multiple plant images of a single weed species (dataset C). These methods are described below.

2.1. Dataset A: Real field dataset

In order to generate a suitable training and validation image set, an extensive image acquisition campaign was conducted in 2017 at two different locations: Limburgerhof (DE) and Utrera (ES). Twenty-four 2.0x2.5 m plots were planted. In these plots, two rows of maize (*Zea mays*) were planted along with 6 different weed species, three "grass-leaved" (*Setaria verticillata, Digitaria sanguinalis, Echinochloa crus-galli*) and three "broad-leaved" (*Abutilon theophrasti, Chenopodium album, Amaranthus retroflexus*). Each plot was photographed with a top view and a perspective view using two different devices: a Canon EOS 700D SLR camera and a Samsung A8 cell phone. To facilitate image acquisition, a metal structure was created to hold two cell phones and two SLR cameras to simultaneously acquire a top view (2.0 m height, 18 mm focal length) and a perspective view (1.6 m height, 30° angle, 18 mm focal length). Images were taken twice a day, three times a week for a period of 9 weeks in order to collect the different phenological stages of corn and weeds. The trials started on 09/05/2017 and ended on 06/07/2017. After removing images containing artifacts, a total number of 1679 images were manually segmented into the 7 target classes named after their corresponding EPPO codes (ZEAMX, SETVE, DIGSA, ECHCG, ABUTH, CHEAL, AMARE). Examples of annotated images are shown at Fig. 1. The Fig. 1 shows some random examples of images from the dataset A.

Manual segmentation of a natural dataset is a complex process involving extensive manual work. Although the targeted weeds were planted at specific positions, the wild growth of unknown weeds in the experimental plots made this task more complex. To address this problem, two new classes were added (generic broadleaf weed and generic grass-leaved weed) that allowed the annotation of unknown or untargeted weeds. The developed network topology was adapted to ignore these noisy annotations. Zhou et al. (2019) showed that when labeling very different objects, as is the case for the scene images, expert annotators make a pixel error of 17.6%. In our use case, We can expect this error to be higher, specially for images with higher leaf density and leaf overlap in our plant dataset. Because of this, we propose two annotation improvements to reduce the annotation time and increase the accuracy of manual segmentation:

- a) Automated ground segmentation: A robust and easy to implement color-based segmentation algorithm was developed to automatically eliminate the presence of soil and automatically subtract it from the manual segmentation. The algorithm is based on thresholding over the *Lab* color space where pixels with positive a-channel values are removed from the segmentation and thus a refined segmentation is obtained.
- b) Support for overlapping plants: Especially in late phenological stages, overlapping plants make it more complicated to accurately segment all classes. To alleviate this, we allow the possibility to mark one annotation as being inside another. In this case, this annotation will be removed from the segmentation of the other classes. This simplifies the annotation process, as it is not necessary to precisely annotate all species, but only those that overlap from the top.

Using these two algorithms for refinements greatly simplifies the annotation process. Fig. 2 depicts the refined manual segmentation after removing the soil and overlap. After annotation corrections, the final composition of the dataset is depicted in Table 1. Even with the proposed enhancements, annotation per image took an average of 15 min without taking into account review time. Without the enhancements, image annotation time was 40 min. (see Fig. 3).

For training, and to avoid bias, the experimental plots were separated into training, test and validation. Eight plots were used for training, two for validation and two for testing. (see Table 2).

2.2. Dataset B: Synthetic dataset

Dataset A has several drawbacks, as annotation is difficult and error-prone due to the considerable complexity of the dataset. As a consequence, the number of images annotated for training and testing is limited and noisy. To mitigate this, we propose the use of synthetic images containing image communities generated by individual plant images. To this end, we conducted an additional individual plant acquisition campaign. This dataset introduces three new weed species:

Fig. 1. Examples of annotated crops at different growing stages.

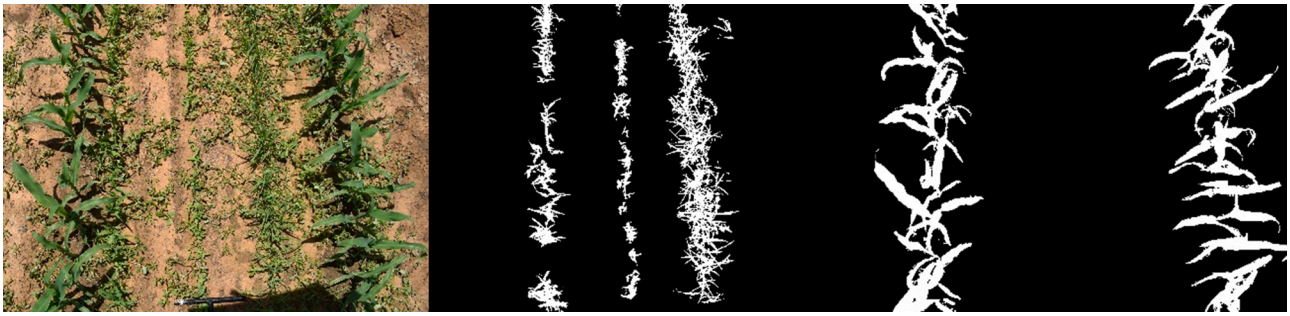

Fig. 2. Examples of annotated images after annotation improvements.

Table 1
Percentage of pixels for the different species in Dataset A.

Species	EPPO-CODE	Type	Coverage
Zea mays	ZEAMX	Crop	15.52%
Setaria verticillata	SETVE	Grass	3.30%
Digitaria sanguinalis	DIGSA	Grass	0.28%
Echinochloa crus-galli	ECHCG	Grass	2.19%
Abutilon theophrasti	ABUTH	Broad	1.51%
Chenopodium albums	CHEAL	Broad	0.57%
Amaranthus retroflexus	AMARE	Broad	4.36%
Other broad	-	Broad	21.48%
Other grass	-	Grass	5.97%
Soil	-	-	55.04%

Chenopodium sp., *Datura stramonium* and *Fallopia convolvulus*. These are images of a single plant in an 80x80cm greenhouse plot based on two greenhouse locations. Different species were planted in each: AMARE, DIGSA, ECHCG and SETVE in greenhouse 1; ABUTH, CHESS, DATST, POLCO and ZEAMX in greenhouse 2. There were a total of 8 weeds and 1 crop. Of each species, 30–36 individual plants were planted. A single image was taken each working day (Monday through Friday) for each of the individual plants, from day 0 to day 80. Not all made it to the last day, so the final data set B contains 6906 images of individual plants of 9 different species and at different growing stages. Since only one plant appears in each image, all images in dataset B can be automatically labeled. Using the vegetation segmentation algorithm we presented in the previous section, we were able to automatically label and segment the entire dataset. Based on these images from individual plants, we created a first version of a synthetic plant community generator algorithm from a collection of individual plants and empty background images.

To generate these images, several random regions associated with three fundamental parameters describing that region are created. The model parameters are: plant species, growing stage and density. The plant species are grown following a Monte-Carlo approach as a function of the region parameters (Fig. 5). The process of this algorithm is as follows (i) growing regions are created as ellipses of random size; (ii) each ellipse is randomly assigned a class (species), an age (days after planting) and a density (real number between 0 and 1); (iii) a location point within the image is randomly sampled for each candidate plant (iv) depending on the location point, a candidate would be within a growing region or not (in that case the potential candidate is rejected); (v) if the candidate is located within an ellipse, the algorithm randomly samples a number between 0 and 1 and compares it with the ”density” parameter of its growth region if the sampled number is greater than the ”density” threshold, the candidate is rejected; (vi) the algorithm chooses from the candidate repository a candidate image that fits the growth region requirements and places it in the plot image.

With this method we generate images in which several plant species are present in different growing stages with non-homogeneous densities. For the experiments proposed in this project we generated a Dataset B consisting of 5000 synthetic images. Of the 5000 generated plot images, 80% were reserved for training, 10% for validation and another 10% for testing. Random examples of the images in dataset B are shown in Fig. 4. In this figure, it can be seen that the proposed method adequately fulfills the geometry and distribution of the plants. However, some images present an unrealistic color integration between species and background that can be improved in the future with more realistic color balance integration or with approaches based on generative adversarial networks (GAN) as, for example, those used by Qin et al. (2020), Yu et al. (2021), Picon et al. (2021)

2.3. Dataset C: Field dataset of a single species.

Synthetic dataset B may have problems mimicking the growth of real plant communities and their adequate overlap, while dataset A has unbalanced classes and noisy annotations. Therefore, we prepared an additional dataset C. This dataset contains images of plants growing in a controlled environment, with only one species in each plot. The fields of the plots were checked daily and each time a plant of another species grew, it was manually removed. Having only one species per plot means that all images are already labeled and therefore semi-automatic segmentation can be performed. There are plots of three densities: high, medium and sparse. The images were taken in two campaigns, one in Utrera (ES) with 4245 images and the other in Limburgerhof (DE) with 818 images. There are substantial differences between the ES and DE images, especially in the soil/background, although the concept is the same. Some examples of images from dataset C are shown in Fig. 6.

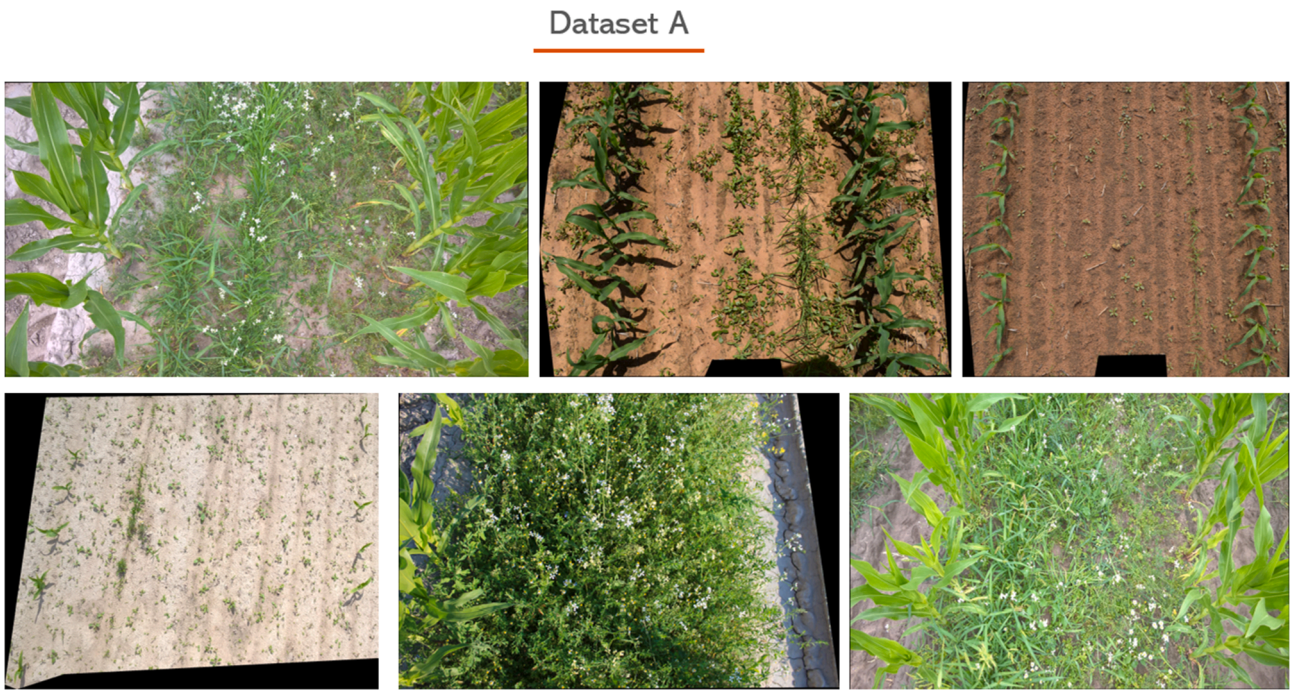

Fig. 3. Six random example images extracted from Dataset A.

Table 2
Percentage of pixels for the different species in dataset B.

species	EPPO-CODE	Type	Coverage
Zea mays	ZEAMX	Crop	24.84%
Setaria verticillata	SETVE	Grass	5.67%
Digitaria sanguinalis	DIGSA	Grass	5.15%
Echinochloa crus-galli	ECHCG	Grass	4.95%
Datura stramonium	DATST	Broad	20.98%
Abutilon theophrasti	ABUTH	Broad	13.84%
Chenopodium albums	CHEAL	Broad	0%
Amaranthus retroflexus	AMARE	Broad	4.47%
Fallopia convolvulus	POLCO	Broad	16.12%
Other	-	-	0%
Soil	-	-	3.99%

Using a leaf segmentation algorithm as described in the previous subsections, we automatically generate labeled masks for each image that act as semantic segmentation labels. Although this segmentation method makes some errors (at the pixel level), we can consider the dataset C as accurately annotated. Table 3 shows the species and percentages of this dataset.

DData sets B and C are similar but complement each other in their differences: Dataset B is more realistic in terms of plant community growth, as it presents several species in the same image, and Dataset C presents better textures, overlays, shadows, and shapes (more information) from actual field images, even though only one species is present.

3. Network Architecture

In this section we describe the two architectures that will be used in this work. First, a baseline segmentation architecture based on state of the art PSPNet network is proposed. This baseline method is extended for the required task by incorporating an additional classification loss function that is able to enhance.the convergence of the training process and thus increase the performance of the model.

3.1. Baseline architecture (PSPNet)

The selection of the most appropriate network architecture was made after considering the particular characteristics of the discriminative patterns of the plant species. First of all, color does not provide too much discriminatory information, except in the case of flowering growth stages or subtle specific coloration in the early growth stages for some species. Therefore, decision making should be based on shape and edge analysis. The proposed baseline network should focus on the following features:

- High feature resolution: Leaves of different weed species can be very similar. Sometimes there are images where the difference between two types of leaves is just ∼20 pixels and can be based on subtle attributes such as apex termination, plant edge shape that only appear on the specific part of the plant. This means that our model needs to learn filters to detect these slight differences, ”focusing” on small groups of pixels and be able to add this contextually relevant information to the other pixels of the plant located far away from the informative attribute.
- Multi-scale detection: This is a key feature, as the scale of the leaves changes from one image to another. Normally we can find different plants at different growth stages in the same image. This means that the model must be able to recognize the same leaf type at different phenological stages and different sizes coexisting in the same image. In addition, it is useful for agricultural researchers to study the evolution of plant species growth, so we want our model to be able to deal with the multiscale problem.

Based on these assumptions, Pyramid Scene Parsing Network (PSPNet) Zhao et al. (2017) was selected as the most promising network. PSPNet is a deep learning model released in 2017 specializing in semantic segmentation for scene understanding. This means classifying each pixel as part of an object, taking into account the color, shape and location of each element. Their idea was to create a standard semantic segmentation network that would add two main features: multiscale (hence the pyramidal module) and contextual information. In the PASCAL VOC 2012 dataset Everingham et al. (2010) this network

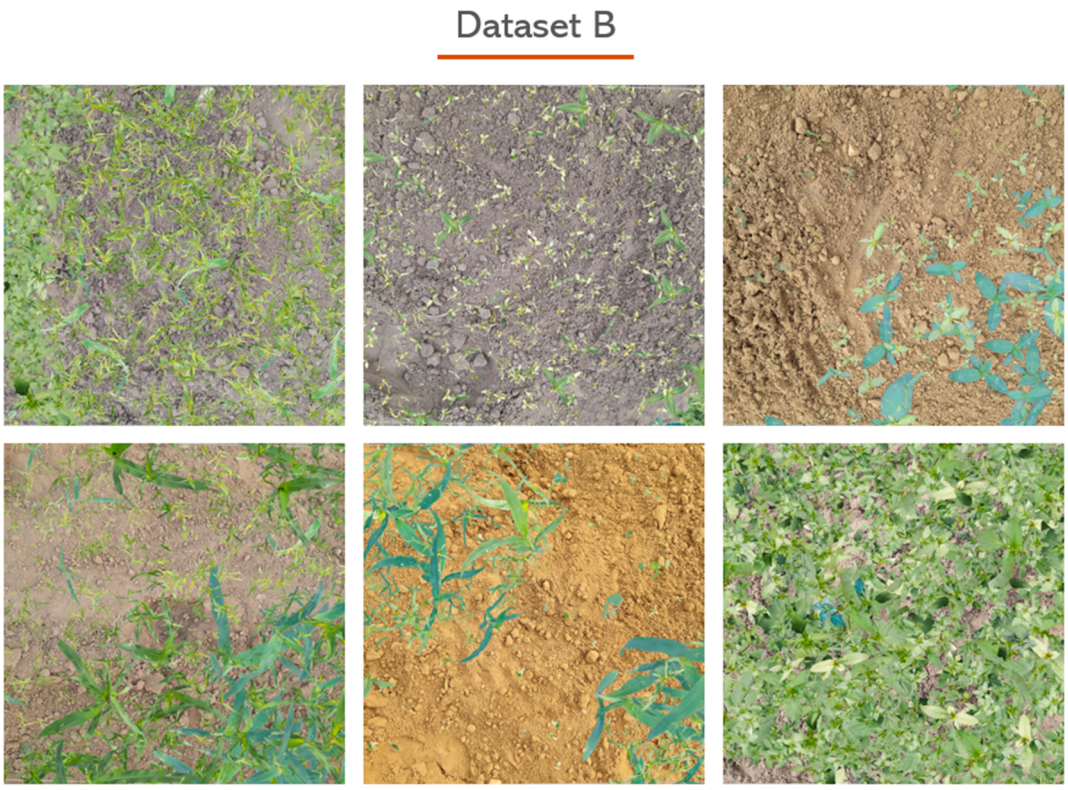

Fig. 4. Six random example images extracted from Dataset B.

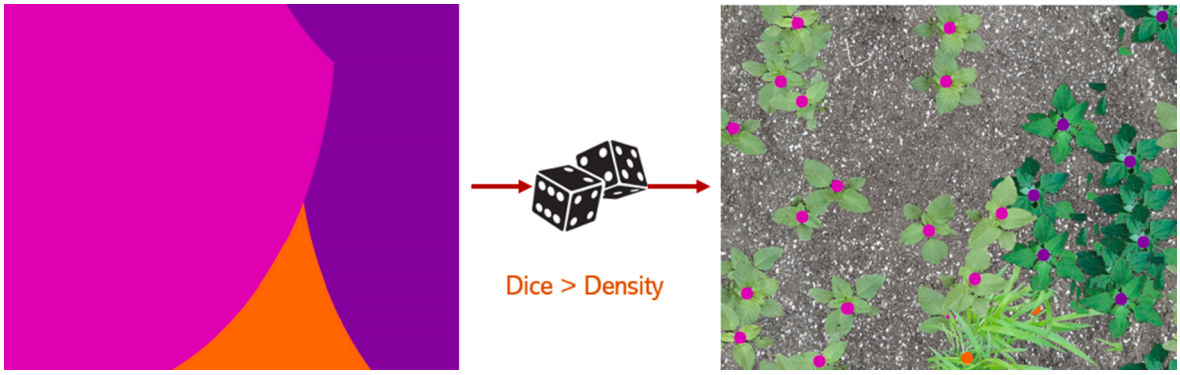

Fig. 5. The spatial distribution of the different candidate plants is generated using a Monte Carlo approach. These images show how the algorithm works. In the image on the left we can see several growth regions - ellipses - of three different classes (magenta, purple and orange). The Monte Carlo algorithm randomly selects different locations for the possible candidates and compares a random sampling with the "density" parameter of the ellipse. This sampling determines whether the candidate "grows" or not. In the image on the right we can see a synthetic image created with the proposed algorithm.

outperformed several models such as DeepLab Chen et al. (2018) or Piecewise Lin et al. (2016), proving to be a promising candidate architecture. Moreover, PSPNet seems to fit the parameters we need to address our weed species identification problem, as it has a pyramid clustering layer (multiscale), is specialized in semantic segmentation (high resolution) and scene parsing (contextual information).

In order to validate this approach, a small dataset A was selected and initial experiments were performed using PSPNet, fully convolutional densenet (Jégou et al. (2017)), Ronneberger et al. (2015). The results showed that although densenet and unet obtained the best separation accuracy between vegetation and soil, resulting in more accurate segmented contours, only PSPNet was able to extract accurate species information. Based on these preliminary results, PSPNet was selected as the reference architecture.

3.2. *Dual PSPNet proposal*

As we have already discussed, PSPNet offers many advantages, such as its ability to extract information at multiple scales. However, it fails when focusing on subtle edges and boundaries that may appear in few pixels in the analyzed image, while maintaining a good understanding of pixel clustering. Therefore, we propose a modification of the original neural network: a dual-task PSPNet. This modification would change the original feature extraction backbone to a classification network. We train the two tasks simultaneously, providing two truth labels for the two tasks and two separate loss functions that are combined.

A simplified schematic of this network is depicted in Fig. 7. It performs two tasks on the same pipeline at the same time. One of our main interests was to use high-resolution images to notice the minimal differences in the shape of the leaves. The first layer of the network has an

Fig. 6. Six random example images drawed from Dataset C.

Table 3
Percentage of pixels for the different species in Dataset C.

species	EPPO-CODE	Type	Coverage
Zea mays	ZEAMX	Crop	3.18%
Setaria verticillata	SETVE	Grass	4.68%
Digitaria sanguinalis	DIGSA	Grass	4.53%
Echinochloa crus-galli	ECHCG	Grass	4.62%
Datura stramonium	DATST	Broad	1.46%
Abutilon theophrasti	ABUTH	Broad	7.03%
Chenopodium album	CHEAL	Broad	1.54%
Amaranthus retroflexus	AMARE	Broad	2.71%
Fallopia convolvulus	POLCO	Broad	3.60%
Soil	-	-	66.64%

input shape of $(473, 473, 3)$ but we did not want to lose any information by resizing the images. The solution was to extract slightly overlapping tiles from the input image of the same dimensions as the input shape of the network.

Since the network performs two tasks, we need two truth labels for the same original mosaic: one for classification and one for segmentation. The classification ground truth indicates which weed species are present in the mosaic, while the segmentation ground truth has each pixel annotated with its corresponding class.

The classification loss indicates which species are present in the tile, the network is forced to learn from the tile as a whole regardless of the percentage that species occupies in the tile. It learns *what is a plant species*, increasing the importance of the leaf as an object. We intend this task to highlight pixel grouping in training and implicitly guide the segmentation task.

For the second task, we provide a pixel-wise annotated ground-truth, so that the network can learn in detail. As a result, the network provides 11 masks with the same shape as the input image (a mosaic of the original), one mask for each possible class. By focusing on each pixel - and its neighbors - the model can learn about the edges and boundaries that differentiate similar weed species.

3.3. Architectural details

The architecture of the proposed neural network is inherited from the original PSPNet Zhao et al. (2017), presenting some variations with respect to the original model. The main schematic depicted in Fig. ?? shows several elements encompassed in two modules. The pipeline is described below from left to right. Each letter refers to the different parts depicted in Fig. 7:

(a) The first element is the input image. The receptive field of our model is much smaller than the original image, so we extract mosaics of size $(443, 443, 3)$ from the image to maintain high resolution and isolate geometric information from the plot distribution.

(b) The second module is a pretrained ResNet50 He et al. (2016) that serves as both a backbone and a classifier. Its first layer has a $(443, 443, 3)$ shape. This module outputs a feature map and thereafter takes two different paths. First it will be interpreted, after a fully connected layer, to form the output of the classification task -(c)-. The loss function selected for the classification task is "weighted binary cross entropy", using "sigmoid" as the last activation layer to support the presence of multiple classes simultaneously, as well as to support class imbalance. The other way is to pass the feature map through the pipeline to the segmentation part.

(c) The classification task of the model analyzes the input image tile by tile and predicts the presence of the different classes in each small portion of the image. By reconstructing the final image from the predicted mosaics, we obtain a classification output. This output is interpreted from the same feature map that will go down the pipeline, so it has the information for both classification (which species are present) and segmentation (which class corresponds to which pixel).

(d) Once the information has passed through the classification backbone, it is reduced to a feature map that enters the pyramid pooling module. The loss function of this block is the "weighted categorical cross-entropy". This module consists of four parts:

(1) The first is proper pyramidal pooling, consisting of four separate filters with different receptive fields that scan the entire feature map and create four other arrays. This is key for multiscale feature detection, as the network can integrate information of different scales and sizes.

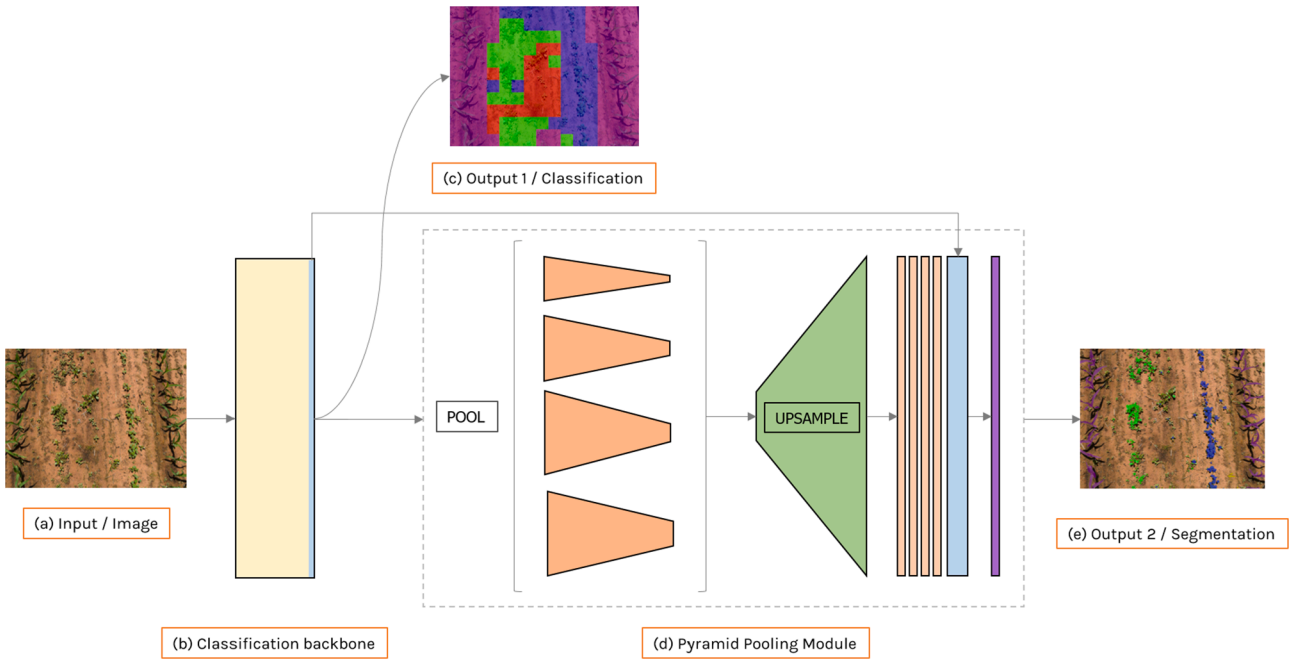

Fig. 7. DUAL PSPNet scheme: It uses a ResNet50 as backbone -(b)-. The trained classification network generates a feature map that is passed to the pyramidal pooling module -(d)-. This pyramidal module extracts multiscale features, context information and concatenates that information to perform semantic segmentation. From a single input -(a)- this CNN gives two outputs: classification (c) and segmentation (e). These two tasks are controlled by two different loss functions.

(2) After pooling, some resampling layers restore each array to the size of the original feature map. This is done by bilinear interpolation.

(3) In the next part, we stack the upsampled arrays of the pyramid pooling, one after the other, and concatenate them with the feature map which is the output of the ResNet50 explained in (b). This creates a module that includes all the multiscale information and the contextual information. This concatenated feature map has the same size as the one created by the feature extractor -(b)-.

(4) Convolution layer receives the feature map from the concatenation network and creates the final pixel prediction map (e). The last activation layer is "softmax".

(e) Finally, the network generates a mask for each class as output of the segmentation task. A final post-processing in which we combine these masks leads to the final image -(e)-

Both tasks are trained by a simple weighted sum of the loss functions for the classification task and the segmentation task simultaneously. Unlike the classical PSPNet, in which the network is divided into two different problems that are trained sequentially with only one task and related loss function active at a time as the training strategy, our proposal extends the network architecture to a true dual-task network in which the network weights are simultaneously optimized against the two loss functions and thus allowing the classification loss to guide the segmentation loss.

The selected loss functions are weighted cross-entropy functions in which each class sample is associated with a weight. The sample weight is related to the dataset to which the target belongs, with samples in datasets B and C being larger than those in dataset A, which is manually segmented. Dataset A features pixels that have been classified as "other" or unknown by a human. For those pixels, the weight is decreased by 100 to reduce their influence on training.

4. Training procedure

To avoid bias, each data set was divided into 80% of the images for training, another 10% for validation, and a final 10% for testing. As the different plots were acquired multiple times on different days, the dataset splitting was performed by ensuring that images belonging to the same plot on different days were assigned to the same dataset split. Data augmentation was applied each time the generator obtained a new image. The transformations we applied to augment the data were: rotation, height and width shift, zoom, vertical and horizontal shift, pixel intensity shift (color shift), and Gaussian blur. Shearing is not recommended, as our method extracts tiles from the image and it is important to maintain spatial coherence. Our code was implemented in TensorFlow. We chose Stochastic Gradient Descent as optimizer for both tasks, using a learning rate of $lr = 0.001$ with a $decay = 10^{-6}$ per epoch, $momentum = 0.9$ and *Nesterov's acceleration*. Balanced Accuracy (BAC) and Dice-Sørensen Coefficient (DSC) were selected as the most suitable algorithm performance metrics, in order to account for the class imbalance present in the datasets (in such cases, the use of "regular" accuracy is discouraged) (Sánchez-Peralta et al. (2020)). Training was performed by a NVIDIA Tesla V100 with 16 GB of memory. Considering the size of the input images we set the *batch size* to 6. Following the same methodology described by Johannes et al. (2017) and Picon et al. (2018) we used Dataset A validation subset and the computed values of BAC and DSC to calculate the threshold values that maximizes the performance over the validation set for the different weed species. Metrics reported are obtained over the testing subset of the Dataset A.

5. Results

5.1. Evaluation metrics

Performance is analysed by the evaluation of Balanced Accuracy (BAC) and Dice-Sørensen Coefficient (DSC). Balanced Accuracy (BAC) is calculated as:

$$BAC = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2}$$

Where *TP* stands for *true positives*, *FN* for *false negatives*, *TN* for *true negatives* and *FP* for *false positives*. The Balanced Accuracy is obtained by

calculating the arithmetic average between sensitivity (proportion of actual positives correctly identified) and specificity (proportion of actual negatives correctly identified). Final result is calculated for each class, analyzing one class against all the others. This particular metric was selected because it prevents result-misrepresentation due to class inbalance. If we took accuracy as our metric, the final result would be distorted as it counts also the correctly identified soil pixels. The Dice-Sørensen Coefficient (DSC) is used to objectively compare the similarity of two sets of data:

$$DSC = \frac{2TP}{2TP + FP + FN}$$

In our particular case, these metrics measure the similarity between the prediction and the ground-truth masks. A result of near $DSC = 0$ would mean that the model cannot distinguish the elements between the two sets. In the contrary case with a value near $DSC = 1$, we would have a model that can perfectly discriminate one set from the other. As it was in the other case, *TP* stands for *true positives*, *FN* for *false negatives* and *FP* for *false positives*. Also, like in the case of balanced accuracy, the final result is calculated for each class, so we compute the *DSC* to weight the similarity of a single weed species against all other elements in the image. The peculiarity of this metric is that it is not affected by true negatives. In our dataset we would find many pixels in the "soil" and "other" classes that are not weed species. If we were to take those correctly predicted pixels into account in the calculation, we would be cheating ourselves about the performance of the model. Therefore, these metrics give us some information about the model's performance in discriminating between different weed species. In other words, the DSC only gives importance to pixels of the evaluated class that were correctly classified (true positives), while penalizing when the prediction is wrong.

5.2. Experiments

In this section we describe the experiments we performed to validate the hypotheses risen in this paper: the benefit of the proposed dual task network architecture and the benefits of using additional data sets that do not require manual annotation to train the algorithms and increase their performance.All experiments were tested using the test subset of data set A, as they represent real field conditions. The results of the actual metrics are shown in the Table 4. e designed two sets of experiments. One set focused on validating the performance of the proposed PSPNet Dual and the other focused on measuring the influence on different combinations of data sets.

5.2.1. Analysis of the influence of the datasets

These experiments were designed to measure the influence of the three different types of data sets on the final performance of the algorithm with the ultimate goal of reducing the need for manually annotated data sets. The experiments are described below:

-**Dual A**: In this experiment the training was performed on dataset A. This dataset had several problems: low number of images, high complexity, inaccurate annotation and high class imbalance.
-**Dual B**: In this experiment, synthetic data set B was used for training. As the training dataset was different from dataset A, a decrease in performance was expected due to the change of domain, as they have differences in spatial distribution, illumination, background and scales. The annotation in this case is not prone to expert error and the annotated plants belong to the error-free species. The information on the shapes and edges of the self leaves is still adequate for training with near perfect ground truth annotation.
-**Dual C**: In this experiment, the single-species data set C is used for training. Although the plants in data set C are obtained under real conditions, the interaction of plant communities cannot be obtained from this data set.
-**Dual A + B**: In this experiment, images from data set A are used for training. However, dataset A is complemented by dataset B. Dataset B allows reducing the effect of class imbalance and poor annotation quality of dataset A by incorporating synthetic images.
-**Dual A + C**: In this experiment, images from data set A are used for training. However, dataset A is supplemented with dataset C. Dataset C allows to reduce the effect of class imbalance and poor annotation quality of dataset A by including the images of a single species from dataset C..
-**Dual A + B + C**: The last experiment complements the images from data set A with images from data set B and data set C.

All experiments were performed using the dual task PSPNet detailed in Section 3.2.

Table 4
Results on the performance of each model at predicting all weed species attending to Balanced Accuracy (BAC) and Dice-Sørensen Coefficient (DSC). All the results correspond to the model's performance tested on the testing sub-set of Dataset A. Further explanation on every model can be found at subSection 5.2. The orange bar at the column "ALL" indicates the best model.

BAC (%)		MODELS							
		PSPNet A+C	Dual A	Dual B	Dual C	Dual A+B	Dual A+C	Dual A+B+C	ALL
CLASSES	ABUTH	81.2	63.1	59.6	70.5	74.3	77.4	74.6	
	AMARE	76.9	62.9	50.9	59.7	71.9	79.6	68.3	
	CHEAL	77.3	56.8	51.3	59.6	65.8	66.0	56.6	
	DIGSA	55.3	52.2	53.9	69.8	66.0	64.0	57.3	
	ECHCG	74.6	65.9	57.7	74.9	75.2	73.9	73.6	
	SETVE	76.4	63.3	50.8	63.8	78.9	77.9	69.8	
	ZEAMX	88.3	81.1	69.7	84.5	88.1	91.5	88.8	
	MEAN	75.71	63.61	56.27	68.97	74.31	75.76	69.86	

DSC (%)		MODELS							
		PSPNet A+C	Dual A	Dual B	Dual C	Dual A+B	Dual A+C	Dual A+B+C	ALL
CLASSES	ABUTH	65.0	23.6	9.6	17.7	44.2	58.0	47.3	
	AMARE	46.8	25.6	6.7	14.0	34.9	57.5	30.6	
	CHEAL	20.5	3.9	0.4	1.1	22.6	28.4	17.3	
	DIGSA	10.9	5.0	3.6	1.7	29.8	33.5	14.2	
	ECHCG	39.8	28.5	5.5	13.2	48.7	28.1	26.4	
	SETVE	53.0	24.2	7.0	31.2	54.8	43.8	38.3	
	ZEAMX	81.3	68.0	53.3	72.1	81.4	86.5	81.1	
	MEAN	45.33	25.54	12.30	21.57	45.20	47.97	36.46	

5.2.2. Analysis of the influence of the network architecture

Another experiment was conducted to validate that the proposed dual PSPNet presents better performance than the normal single-task PSPNet. For this purpose, we conducted two additional experiments:

- **PSPNet A + C**: This experiment uses a reference PSPNet trained with images from both dataset A (real dataset) and dataset C (single weed dataset per image). As detailed in Section 5.2.3, we selected dataset A + C because it was the best combination of datasets for training.
- **Dual PSPNet A + C**: This experiments is identical to *PSPNet A + C* but changing the classical PSPNet network architecture by the proposed dual task PSPNet.

5.2.3. Experiments results

This section presents and discusses the results obtained in the experiments described above. These results are summarized in Table 4.

As can be seen, training with only dataset A gives poor results (DSC=0.26). This may be due to the aforementioned problems of poor annotation quality, class imbalance and image sparsity that occurs in dataset A. The experiments also clearly show poor results when training only with source datasets B (DSC=0.12) or C (DSC=0.19). This demonstrates the existence of a domain shift among the different datasets. Therefore, we can conclude that training with only one type of dataset does not lead to adequate performance in plant species segmentation.

However, when data set A is combined with any of the supporting data sets, the domain shift is reduced by doubling the performance, as shown in Table 4. This performance improvement does not require extensive annotation of the datasets. The best combination of data sets was obtained by including data sets A and C for training.

It should be noted that synthetic dataset B does not provide better performance than dataset C. This may be because the integration of individual species within the composite synthetic image produced some unrealistic images. This may be resolved in the future by using more advanced color balancing methods or using generative adversarial networks to generate more plausible synthetic images.

Therefore, analyzing metrics alone is not enough to check which model performs better. A visual examination of the predictions is needed. Fig. 9 shows six representative examples of images (a) with the annotation masks (b) and the predictions of PSPNet A + C (c) and Dual PSPNet (d). The figure shows examples at different stages of growth. We can see that in the early stages the Dual network is able to notice subtle differences that PSPNet cannot. At medium stages the performance of the two models is quite similar, finding that Dual predicts corn yield better since it does not label as a crop the plants between the leaves of Zea mays. At high growth stages, Dual performs better because it has a higher clustering of pixels: a better understanding of the leaf as a concept. This allows the network to correctly predict corn in large images even when analyzing tile by tile. To understand the actual performance of the model, we also analyzed the class confusion matrix shown in Fig. 8. In this matrix we can appreciate that, when taking the best performing models, crops and broad-leaved plants (ABUTH, AMARE, CHEAL) are clearly distinguished while grass-leaved plants (DIGSA, ECHCG, SETVE) are not. This prediction error is due to small visual dissimilarities that challenge even human experts.

Results show that the use of the proposed dual classification and segmentation network obtained an average Dice-Sørensen Coefficient (DSC) of ~ 0.48 against the $\sim 45\%$ obtained when using the classical architecture and balanced accuracy is also improved. Detailed results can be seen on Table 4.

5.3. Validation of algorithm predictions on technician evaluations

The last step in the evaluation of the developed algorithm was to compare the predicted coverage with expert technician evaluations. Fig. 10 depicts the correlation between the model predictions and technician evaluations. This comparison between model predictions and ground truth was performed with simple linear regressions. The data used for this analysis were the test subset of the data set A. To measure fitting quality we used the coefficient of determination R^2. It provides a measure of the proportion of variance in the dependent variable, with a value of $R = 1$ being a perfect fit and a value of $R = 0$ indicating completely independent variables.

Analysis of the data showed a near perfect fit $R^2 = 0.98$ for the corn crop and coefficients of determination greater than 0.85 for the broad weeds: ABUTH and AMARE. In the case of CHEAL, the low coefficient of determination is caused by the lack of presence of CHEAL in data set A, as shown in Table 1. The coefficient of determination in the case of DIGSA and SETVE is higher than 0.75. However, the algorithm fails in detecting ECHCG, as it is confused with the other weeds (SETVE and DIGSA), where there are several images in which the algorithm declares a good amount of coverage, while the ground truth says that the

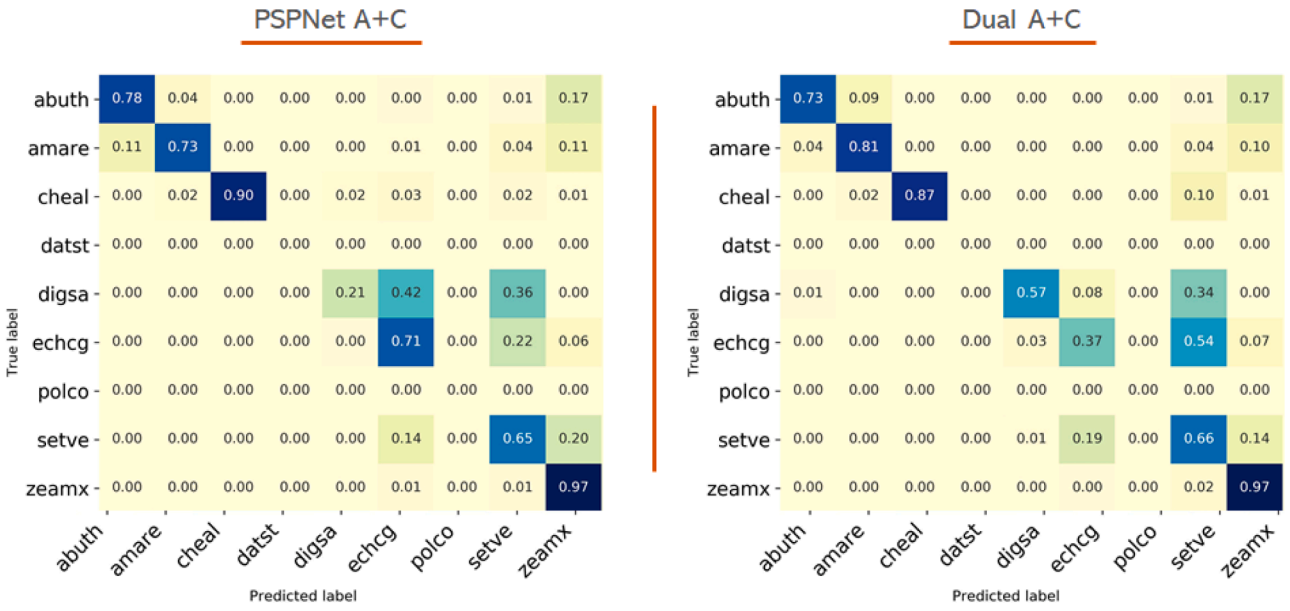

Fig. 8. Confusion matrix comparison. On the left, the confusion matrix for the PSPNet A + C model. On the right, the one corresponding to model Dual A + C.

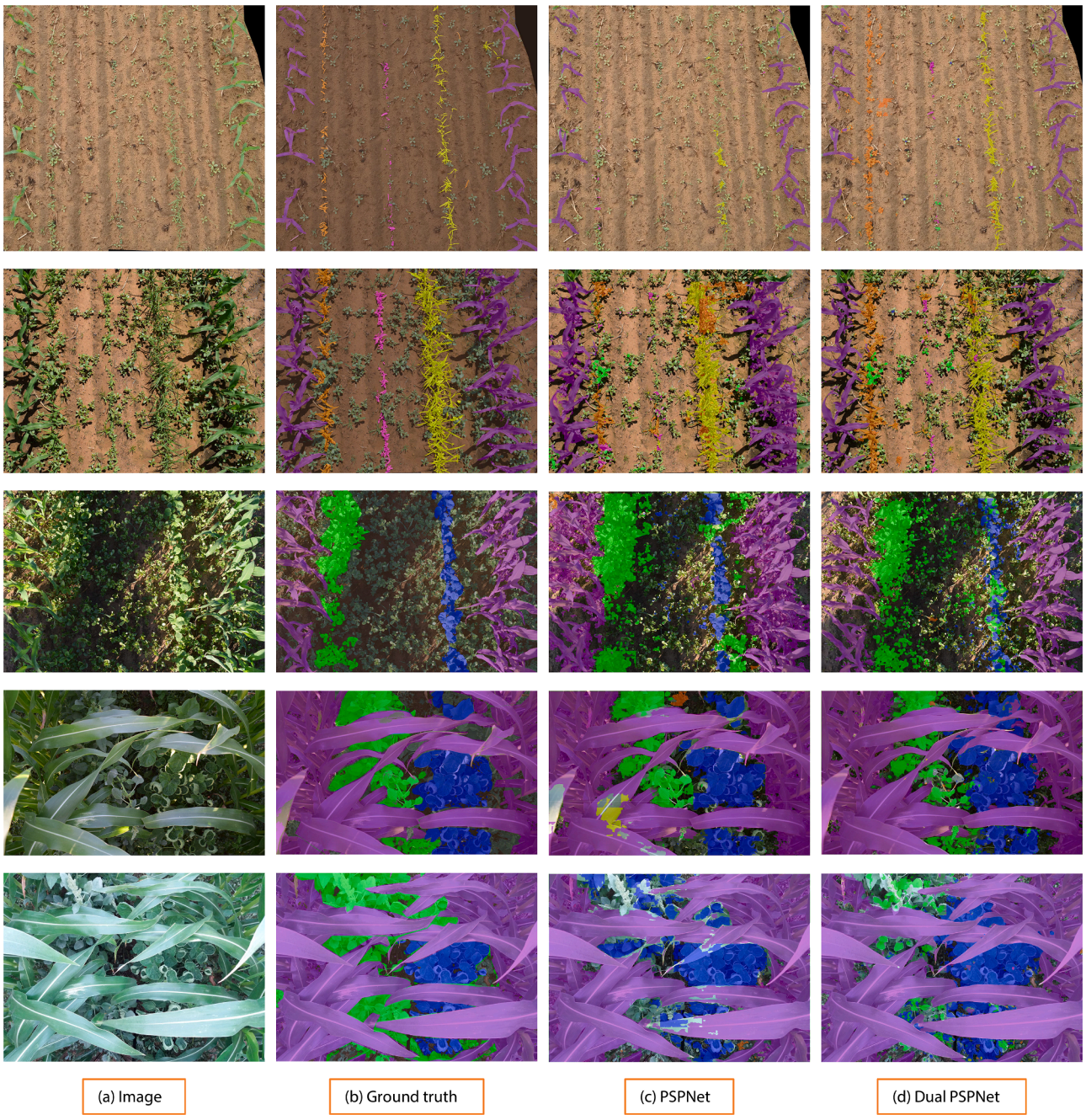

Fig. 9. Comparison between predictions of the standard PSPNet A + C -(c)- and our Dual-task PSPNet A + C -(d)-. Although both models get a similar score on overall BAC, Dual PSPNet makes more realistic predictions.

coverage is zero.

This analysis shows that crop coverage and weed coverage can be accurately predicted both for corn crop and broad leaf weeds. In the case of grasses, the confussion among the different classes increases the error on the coverage detection for ECHCG.

6. Conclusions

In this work we have proposed, for the first time, a dual-task network architecture that is suitable for segmenting different weed species under field conditions. The proposed dual-task model performs better with different species than current segmentation networks, being able to distinguish between different weed species and crops, showing a high degree of invariance to illumination, leaf overlap, background and scale.

We have also proposed three methods to reduce the cost of vegetation image annotation, helping to reduce the annotation effort. The first method focuses on refining the manually annotated data using a color-based segmentation method (dataset A). This reduces the average annotation time per image from 40 min to 15 min. The other two proposed methods do not require manual labeling, as they focus on generating single-species datasets (dataset C) or synthetic data from greenhouse images (dataset B).

Our results show that when training the proposed model with only the manually labeled dataset (dataset A) a Dice-Sørensen Coefficient (DSC) score of 25.32 is obtained. This performance increases when this dataset is combined with the proposed datasets that do not require costly manual annotation, such as the single weed class dataset (DSC=47.97) or the synthetic dataset (DSC=45.20). The ablation study shows that

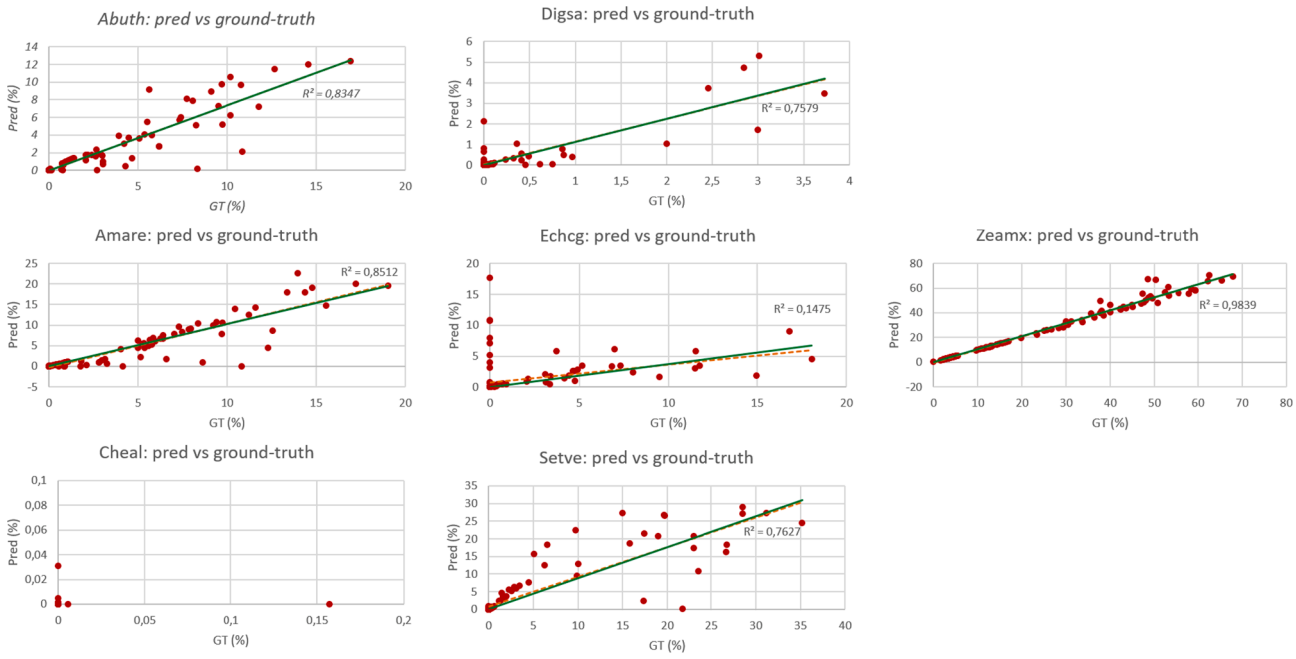

Fig. 10. Assessment of weed coverage on testing Dataset A. Comparison between the model's coverage estimation (Pred) and the manual pixel-wise annotations (GT). Left column shows the regressions for the "broad leaf" weeds. Central column shows the results for the "grass leaf" weeds. Right column shows the results for the corn crop.

when the proposed dual task network is not used, the performance decreases (DSC=45.96). We have also validated that, due to the domain drift problem, the use of data set B (DSC=12.30) and data set C (DSC=21.57) alone is not adequate for model performance.

Although synthetic dataset B improves segmentation performance, it does not outperform the single-species dataset (dataset C), probably because the synthetic image generation method cannot accurately integrate the color balance between different images. Future work will focus on the use of generative adversarial networks and more advanced color balancing techniques to generate more realistic images.

Although good segmentation is obtained for crop and broadleaf species, there is still room for improvement in detecting "grass-like" weed species since it does not classify them correctly due to their high visual similarity. The use of synthetic images provides good results in combination with a real dataset.

The proposed algorithm has been validated for experimental trials of digitization and automation of herbicide research and development by aiding the automation of these processes.

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

We would like to thank BASF technicians Rainer Oberst, Gerd Kraemer, Hikal Gad, Javier Romero and Juan Manuel Contreras, as well as Amaia Ortiz-Barredo from Neiker for their support in the design of the experiments and the generation of the data sets used in this work. This was partially supported by the Basque Government through ELKARTEK project BASQNET(ref K-2021/00014).

References

Argüeso, D., Picon, A., Irusta, U., Medela, A., San-Emeterio, M.G., Bereciartua, A., Alvarez-Gila, A., 2020. Few-shot learning approach for plant disease classification using images taken in the field. Comput. Electron. Agric. 175, 105542.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40, 834–848.

Dyrmann, M., Karstoft, H., Midtiby, H.S., 2016a. Plant species classification using deep convolutional neural network. Biosyst. Eng. 151, 72–80. https://doi.org/10.1016/j.biosystemseng.2016.08.024 http://www.sciencedirect.com/science/article/pii/S1537511016301465.

Dyrmann, M., Mortensen, A.K., Midtiby, H.S., Jørgensen, R.N. et al., 2016b. Pixel-wise classification of weeds and crops in images by using a fully convolutional neural network. In: Proceedings of the International Conference on Agricultural Engineering, Aarhus, Denmark, pp. 26–29.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88, 303–338.

Grinblat, G.L., Uzal, L.C., Larese, M.G., Granitto, P.M., 2016. Deep learning for plant identification using vein morphological patterns. Comput. Electron. Agric. 127, 418–424. https://doi.org/10.1016/j.compag.2016.07.003 http://www.sciencedirect.com/science/article/pii/S0168169916304665.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

van Heemst, H., 1985. The influence of weed competition on crop yield. Agric. Syst. 18, 81–93. https://doi.org/10.1016/0308-521X(85)90047-2 http://www.sciencedirect.com/science/article/pii/0308521X85900472.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 11–19.

Johannes, A., Picon, A., Alvarez-Gila, A., Echazarra, J., Rodriguez-Vaamonde, S., Navajas, A.D., Ortiz-Barredo, A., 2017. Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case. Comput. Electron. Agric. 138, 200–209. https://doi.org/10.1016/j.compag.2017.04.013 http://www.sciencedirect.com/science/article/pii/S016816991631050X.

Kamilaris, A., Prenafeta-Boldú, F., 2018. A review of the use of convolutional neural networks in agriculture. The Journal of Agricultural Science 1–11.

Lin, G., Shen, C., Van Den Hengel, A., Reid, I., 2016. Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3194–3203.

McCarthy, C.L., Hancock, N.H., Raine, S.R., 2010. Applied machine vision of plants: a review with implications for field deployment in automated farming operations. Intel. Serv. Robot. 3, 209–217. https://doi.org/10.1007/s11370-010-0075-2.

Medela, A., Picon, A., 2020. Constellation loss: Improving the efficiency of deep metric learning loss functions for the optimal embedding of histopathological images. Journal of Pathology Informatics 11.

Medela, A., Picon, A., Saratxaga, C.L., Belar, O., Cabezón, V., Cicchi, R., Bilbao, R., Glover, B., 2019. Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, pp. 1860–1864.

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.

Mortensen, A.K., Dyrmann, M., Karstoft, H., Jørgensen, R.N., Gislum, R. et al., 2016. Semantic segmentation of mixed crops using deep convolutional neural network. In: CIGR-AgEng Conference, 26–29 June 2016, Aarhus, Denmark. Abstracts and Full papers. Organising Committee, CIGR 2016, pp. 1–6.

Picon, A., Alvarez-Gila, A., Irusta, U., Echazarra, J., 2020. Why deep learning performs better than classical machine learning.

Picon, A., Alvarez-Gila, A., Seitz, M., Ortiz-Barredo, A., Echazarra, J., Johannes, A., 2018. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. Comput. Electron. Agric.

Picon, A., Medela, A., Sánchez-Peralta, L.F., Cicchi, R., Bilbao, R., Alfieri, D., Elola, A., Glover, B., Saratxaga, C.L., 2021. Autofluorescence image reconstruction and virtual staining for in-vivo optical biopsying. IEEE Access 9, 32081–32093.

Picon, A., Seitz, M., Alvarez-Gila, A., Mohnke, P., Ortiz-Barredo, A., Echazarra, J., 2019. Crop conditional convolutional neural networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions. Comput. Electron. Agric. 167, 105093.

Qin, Z., Zhang, P., Wu, F., Li, X., 2020. Fcanet: Frequency channel attention networks. arXiv:2012.11879.

Romera-Paredes, B., Torr, P.H.S., 2016. Recurrent instance segmentation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016. Springer International Publishing, Cham, pp. 312–329.

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.

Sa, I., Chen, Z., Popović, M., Khanna, R., Liebisch, F., Nieto, J., Siegwart, R., 2018. weednet: Dense semantic weed classification using multispectral images and mav for smart farming. IEEE Robot. Autom. Lett. 3, 588–595.

Sánchez-Peralta, L.F., Picón, A., Antequera-Barroso, J.A., Ortega-Morán, J.F., Sánchez-Margallo, F.M., Pagador, J.B., 2020. Eigenloss: Combined pca-based loss function for polyp segmentation. Mathematics 8, 1316.

Teimouri, N., Dyrmann, M., Nielsen, P., Mathiassen, S., Somerville, G., Jørgensen, R., 2018. Weed growth stage estimator using deep convolutional neural networks. Sensors 18, 1580. https://doi.org/10.3390/s18051580.

Yu, N., Liu, G., Dundar, A., Tao, A., Catanzaro, B., Davis, L., Fritz, M., 2021. Dual contrastive loss and attention for gans. arXiv:2103.16748.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A., 2019. Semantic understanding of scenes through the ade20k dataset. Int. J. Comput. Vision 127, 302–321.