2020 IEEE Conference on Control Technology and Applications (CCTA)
August 24-26, 2020. Montréal, Canada

Data Augmentation Using GANs for Crop/Weed Segmentation in Precision Farming

Mulham Fawakherji
Dept. of Computer, Control, and
Management Engineering
Sapienza University of Rome
Rome, Italy
fawakherji@diag.uniorma1.it

Ciro Potena
Engineering Department
Roma Tre University
Rome, Italy
cpotena@os.uniroma3.it

Ibis Prevedello
Dept. of Computer, Control, and
Management Engineering
Sapienza University of Rome
Rome, Italy
ibiscp@gmail.com

Alberto Pretto
IT+Robotics S.r.l.
Padova, Italy
alberto.pretto@it-robotics.eu

Domenico D. Bloisi
Dept. of Mathematics,
Computer Science, and Economics
University of Basilicata
Potenza, Italy
domenico.bloisi@unibas.it

Daniele Nardi
Dept. of Computer, Control, and
Management Engineering
Sapienza University of Rome
Rome, Italy
nardi@diag.uniorma1.it

***Abstract*—Farming robots need a fast and robust image segmentation module to apply targeted treatments, which require the ability to distinguish, in real time, between crop and weeds. Existing solutions make use of visual classifiers that are trained on large annotated datasets. However, generating large datasets with pixel-wise annotations is an extremely time-consuming task. In this work, we tackle the crop/weed segmentation problem by using a synthetic image generation method to augment the training dataset without the need of manually labelling the images. The proposed approach consists in training a Generative Adversarial Network (GAN), which can automatically generate realistic agricultural scenes. As a difference with respect to common GAN approaches, where the network learns how to reproduce an entire scene, we generate only instances of the objects of interest in the scene, namely crops. This allows to build a generative model that is more compact and easier to train. The generated objects are then placed into real images of agricultural datasets, thus creating new images that can be used for training. To evaluate the performance of the proposed approach, quantitative experiments have been carried out using different segmentation network architectures, showing that our method well generalizes across multiple architectures.**

I. Introduction

The interest in robots for precision agriculture is growing due to the need to minimise the energy and waste in agri-food production. Precision agriculture aims at developing monitoring and intervention techniques to improve efficiency and to reduce the environmental impact. In particular, monitoring data is obtained though the deployment of sensing technologies and automation [1]. In this context, robotics can play a key role, leading to a full automation of time consuming farming activities such as data collection, data analysis, and the targeted application of chemicals.

In order to build effective farming robots, one of the main challenges to address is the crop and weeds detection task, where the robot should be able to identify the vegetation and to distinguish between crop and weeds (see Fig. 1). Moreover, this process has to be carried out in real-time in order to

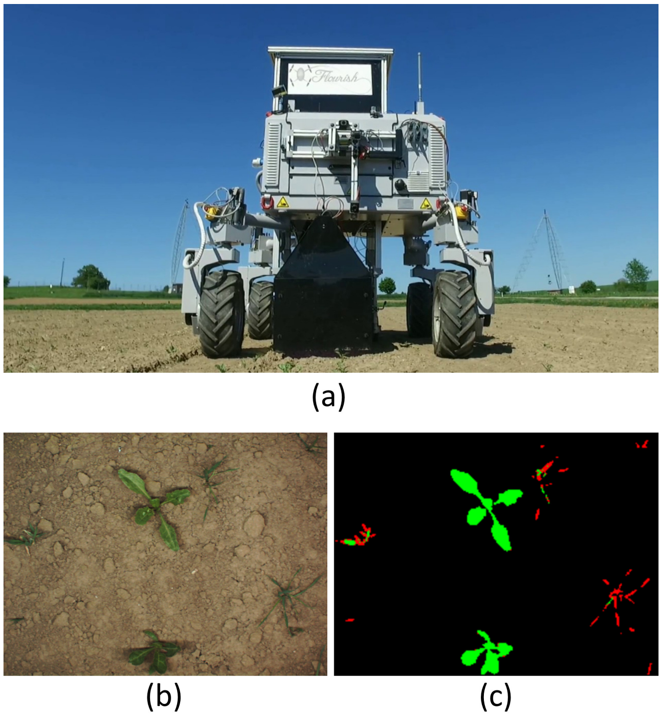

Fig. 1. a) The BoniRob farming robot used in the Flourish project to acquire the datasets for the experiments. The black structure under the robot includes the sensors and the lighting devices used to acquire the data. b) An image of the ground captured by the robot. c) Result of the crop/weed segmentation process of image b), where green pixels represent crop and red pixels represent weeds.

978-1-7281-7140-1/20/$31.00 ©2020 IEEE

trigger the proper weeding actions. Machine learning methods, specifically Convolutional Neural Networks (CNNs), have been used to accomplish the crop/weed classification task (e.g., [2]–[4]). These methods enable to train highly discriminative visual models that can be used to distinguish among different plant species with great accuracy. However, CNNs are data-driven methods, meaning that their performance depends on the size and variety of the training dataset [5]. Creating a good dataset for agricultural purposes is not easy: 1) Data must be collected across different seasons, crop growth-stages, and weather conditions; 2) training images need to be annotated at pixel level, which is an extremely time-consuming task.

In this paper, we focus on increasing the size of existing agricultural datasets without any manual annotation effort. The idea is to generate synthetic images similar to the real ones provided by existing datasets [6]. Specifically, the generated images will look real to a human eye and will be useful samples to improve accuracy when used for training classification algorithms, in our case a CNN for semantic segmentation. To achieve our goal, we propose a novel architecture for the production of sugarebeet images composed of two main steps. In the first step, a GAN [7] generates a large amount of synthetic crop instances. In the second step, the generated crop instances are inserted into real images coming from existing agricultural datasets, thus creating new annotated images.

Using GANs to improve the training results of a neural network is a well-known technique in machine vision [8], [9]. However, as a difference with respect to other approaches in the literature, where GANs are usually exploited to reproduce an entire scene [10], we aim at generating specific objects only and at inserting them into real images. This allows to train more compact generative models, thus to create photo-realistic training samples in a faster and more effective manner. The proposed approach is evaluated on images from a sugarbeet precision farming scenario. We generate synthetic datasets to training different state-of-the-art segmentation networks, showing how the proposed method for synthetic image generation properly generalizes across different segmentation approaches.

The rest of the paper is organized as follows. After discussing related work in Section II, the proposed approach is described in Section III. Experimental results are shown in Section IV, while conclusions are drawn in Section V.

II. Related Work

Several crop/weed segmentation algorithms have been proposed in the past few years, demonstrating the large interest of the robotic community in this topic. Haug *et al.* [11] propose a machine vision approach for plant classification. The method is capable of distinguishing carrot and weeds by using RGB and near infra-red (NIR) images. The reported accuracy is around 93.8%. McCool *et al.* [12] propose a three stage approach. They start from a pre-trained model with state-of-the-art performance but a high computational cost. Then, a model compression is performed, retrieving a faster CNN. Finally, they combine several lightweight models into a mixture model obtaining an accuracy of 93.9%.

Mortensen *et al.* [13] use a CNN to estimate the in-field biomass and crop composition. Their method is a modified version of the well-known VGG-16 deep neural network. The reported accuracy is 79% with 7 classes of objects. Lottes *et al.* [14] propose a Random Forest classifier based on a tailored feature extraction and on an additional Markov random field process to smooth the resulting segmentation mask. They report a crop weed estimation accuracy around 96%. Sa *et al.* [3] propose an aerial segmentation algorithm. The method exploits multi-spectral imagery and a CNN, reporting an error margin $\leq$ 2%. In [15], we propose a pipeline with multiple data channels to support the input of the CNN by using more vegetation indices. This helps to achieve a good generalization to different crop types.

A. Data Augmentation

Despite the promising results, the above cited methods require a training stage that use large annotated datasets, which are difficult to generate. To cope with this problem, several solutions have been proposed in the literature. In [4], we propose a novel dataset summarization technique. This approach builds a summarized dataset from a big, unlabelled, dataset. The summarized dataset will contain the most informative images and requires a reduced labelling effort. In [16], we propose to use a state-of-the-art graphic engine to generate synthetic but realistic farming scenes. The generated scene, together with ground truth data, are used to train the final CNN or to supplement an existing real dataset.

Milioto *et al.* [2] propose a CNN that requires little data to adapt to a new, unseen environment. The CNN is fed by additional, task-relevant background knowledge helping to speed up the training and to better generalize to new crop fields. The reported results show a segmentation accuracy around 96% and a fast re-adaptation to unseen environments. Giuffrida *et al.* [8] exploit a conditional GAN to generate 128×128 synthetic *Arabidopsis* plants, with the possibility of deciding the desired number of leaves of the final plant. The method is tested by using a leaf counting algorithm showing how the addition of synthetic data helps to avoid overfitting and to improve the accuracy.

The approach proposed in this paper is similar to existing methods for data augmentation, but differs for an important detail: We want to create a synthetic sample for the object of interest and put it into an image from an existing dataset instead of creating a fully synthetic new image to be added to the existing dataset.

III. Methods

The main goal of this work is to develop an algorithm that can learn how to generate synthetic images from agricultural scenes similar to real ones. The resulting images will be used to train a CNN for crop/weed segmentation. The role of the synthetic images is to increase the number of training samples, which in turn allows to improve the segmentation accuracy of the CNN.

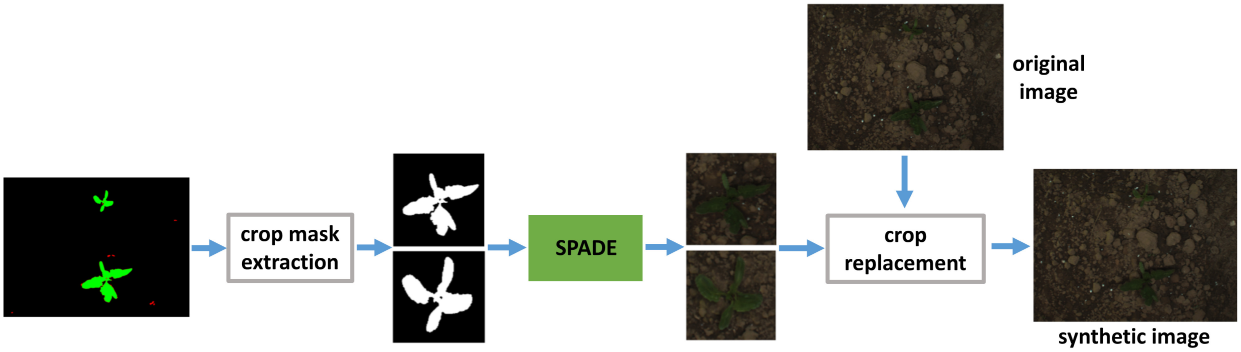

Fig. 2. Dataset creation for segmentation training. First the crop masks are taken from the full mask, then new RGB crops are generated from these masks and past back to the original image.

In this section, we describe the three major steps involved in the development of our sample generation algorithm:

(A) The RGB image generation process.
(B) The evaluation of the GAN training using appropriate metrics.
(C) The composition of the synthetic farming scenes.

A. RGB Crop Image Generation

The first step of our approach concerns the generation of agricultural scenes populated by synthetic crop samples. A similar problem is usually addressed by training a generative model capable to reproduce an entire scene, which requires deep models, a large amount of training data, and a high computational power. However, by considering that the goal is to achieve data augmentation for the training dataset, a full scene generation is redundant. Moreover, since the accuracy of CNNs can vary a lot from class to class, it is interesting to augment the number of training samples only for the classes with a lower classification accuracy.

In a crop/weed segmentation scenario, a network usually tends to accurately detect the terrain, while it usually misclassifies the pixels belonging to crop and weeds due to their similar visual appearance. In such a scenario, it is useless to increase the number of terrain samples, while increasing the crop samples can provide a significant information. Consequently, we propose to only learn generative model to build new instances for specific object classes, i.e., the ones with low segmentation accuracy.

Let $m \in \mathbb{L}^{w \times h}$ be a semantic segmentation mask, where w and h represent its width and height, respectively. Our goal is to learn a mapping function that converts m into a photo-realistic image. To achieve our aim, we employ the SPatially-Adaptive DEnormalization (SPADE) architecture [17], which is a type of conditional GAN (cGAN) where the generative model is trained by conditioning the shape of the generated object. Therefore, by providing as an input constraint the object shapes extracted from real objects, the generative model will synthesize new realistic objects while keeping their original footprint onto the image. Differently from other existing cGANs, this type of network is a Semantic Image Synthesis that converts a semantic segmentation mask into a photo-realistic image. In other words, its input/output behavior is the opposite of an image segmentation network.

In the SPADE architecture, the image encoder encodes a real image to a latent representation for generating a mean and a variance vector. This allows to change the style (i.e., color, texture, etc.) of the image. The generator uses the segmentation map in each SPADE ResNet blocks. The discriminator classifies the concatenation of the segmentation map with the original (or generated) image as real or fake.

The SPADE generator is built based on pix2pixHD [18]. It starts with random noise in the input and uses the semantic map at every SPADE ResBlk upsampling layer. Also, using SPADE, it is possible to separate between semantic and style control. By modifying the semantic map, it is possible to change the final content. By modifying the random vector, it is possible to change the style of the image.

Also the discriminator architecture follows pix2pixHD presenting a multi-scale design that includes instance normalization, with the difference that spectral normalization is applied to all the convolutional layers of the discriminator. In particular, it is based on Patch-GAN [19] and it takes as input the concatenation of the segmentation map with the image. The encoder is composed of six convolutional layers with stride 2 followed by two linear layers. It is responsible for producing the mean (μ) and covariance (σ^2).

B. GAN Training Evaluation

Evaluating GANs is a very challenging task and several aspects need to be taken into consideration when defining metrics that can produce meaningful scores, such as distinguishing generated from real samples, detecting overfitting, and identifying mode dropping and mode collapsing.

For most of the GANs presented in the literature, network inspection is qualitative only, based on a manual inspection to check the fidelity of the generated sample. This kind of evaluation is still considered the best approach, but it is time-consuming, subjective, and often it can also be misleading. In this paper, we employ an empirical evaluation. The basic idea is to use samples generated by the network and samples collected from the real dataset, to extract features from both of them, and then to calculate performance using specific

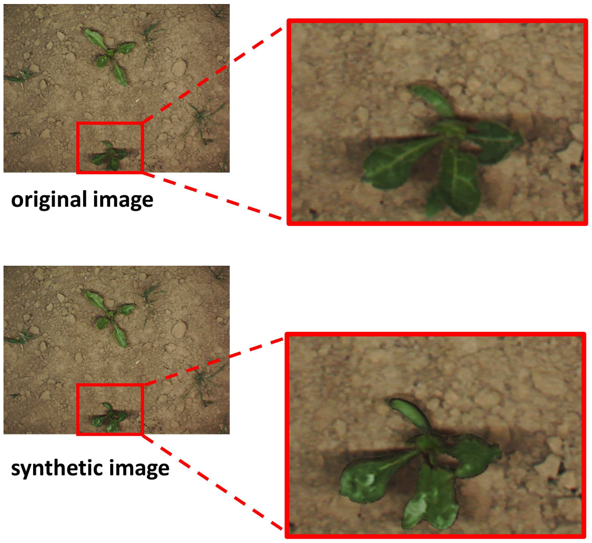

Fig. 3. Difference between the original image and the synthetic one obtained by inserting in the original image a plant sample generated using the GAN.

TABLE II
BONNET IOU SEGMENTATION RESULTS

Dataset	Ground IoU	Weed IoU	Crop IoU	IoU	GA
Original	0.995	0.34	0.84	0.723	0.964
Synthetic	0.992	0.31	0.85	0.716	0.959
Mixed	**0.998**	**0.53**	**0.92**	**0.813**	**0.978**

TABLE III
SEGNET IOU SEGMENTATION RESULTS

Dataset	Ground IoU	Weed IoU	Crop IoU	IoU	GA
Original	0.98	0.30	0.73	0.67	0.972
Synthetic	0.979	0.24	0.70	0.63	0.96
Mixed	**0.995**	**0.33**	**0.90**	**0.74**	**0.989**

TABLE IV
UNET-RESNET IOU SEGMENTATION RESULTS

Dataset	Ground IoU	Weed IoU	Crop IoU	IoU	GA
Original	0.98	0.33	0.94	0.75	0.97
Synthetic	0.979	0.37	0.91	0.76	0.98
Mixed	**0.997**	**0.47**	**0.94**	**0.80**	**0.983**

metrics. In particular, we employ six metrics: Inception Score, Mode Score, Kernel MMD, Wasserstein distance, Fréchet Inception Distance (FID) and 1-nearest neighbor (1-NN). The description of the above listed metrics can be found in [20].

C. Scene Generation

The final step in our approach concerns how to use the crop RGB images described in Section III-A to create a realistic agricultural scene. To do so, we follow the pipeline reported in Fig. 2. First, we get the crop masks from the full mask image. The mask is then resized to the SPADE network input size, which is 256×256 pixels. The SPADE network then generates a random, photo-realistic crop instance by using the shape of the input mask and a random noise signal. The generated image is then up-sampled again to its original size and replaced into the source image. Despite the simplicity of the proposed approach, the up-sampling procedure may let the synthetic image to loose some details. Fig. 3 shows an example of the obtained synthetic image.

IV. EXPERIMENTAL RESULTS

Quantitative experiments have been carried out to support the main claim made in this paper: A synthetic photo-realistic generative model can augment a real-world dataset to train a crop/weed segmentation system. This results in 1) a lower annotation effort, since less real images need to be labelled, and 2) an improvement of the network generalization capabilities and performance.

TABLE I
UNET IOU SEGMENTATION RESULTS

Dataset	Ground IoU	Weed IoU	Crop IoU	IoU	GA
Original	0.997	0.294	0.901	0.732	0.974
Synthetic	0.992	0.248	0.826	0.69	0.962
Mixed	**0.997**	**0.553**	**0.946**	**0.831**	**0.995**

To test our method, we have used the open-source Sugar Beet 2016 dataset [6]. The dataset has been collected by a BOSCH Bonirob farm robot (see Fig. 1a) moving on a sugar beet field across different weeks. The dataset is composed of a set of images taken by a 1296×966 pixels 4-channels (RGB-NIR) JAI AD-13 camera, mounted on the robot and facing downward. Specifically, from the Sugar Beet 2016 dataset we took a total of 1600 images, randomly chosen among different days of acquisition in order to contain different growth-stages of the target crop. We denote the reduced Sugar Beet 2016 dataset as *Original* and then we split it into a traning set (1000 images), a validation set (300 images), and a test set (300 images). In addition, we have created a *Synthetic* dataset composed of 1000 images generated by using the proposed architecture. In particular, for each image in the original training set, we created a synthetic image by replacing all the crops. Finally, we also created a third dataset, denoted as the *Mixed* dataset, composed by the union of the *Original* and the *Synthetic* datasets.

We have used the three datasets to train four state-of-the-art semantic segmentation networks, namely Unet [21], Unet-Resnet (i.e., Unet with Resnet50 back-end), Bonnet [22], and Segnet [23]. To evaluate the semantic segmentation output, we have taken into account 300 real images in all the experiments, computing the following metrics: Per-class Intersection over Union (denoted as *Ground IoU*, *Weed IoU*,*Crop IoU*), Average Intersection over Union (denoted as *IoU*), and Global classification Accuracy (denoted as *GA*). An example of segmentation results is shown in Fig. 4. Tables I, II, III, and IV show the quantitative results for the semantic segmentation.

For all the architectures, the results show that the IoU value increases drastically by using the original dataset augmented with the synthetic one compared to using only the original dataset. Moreover, using only the synthetic dataset also leads to a competitive performance when compared with the use

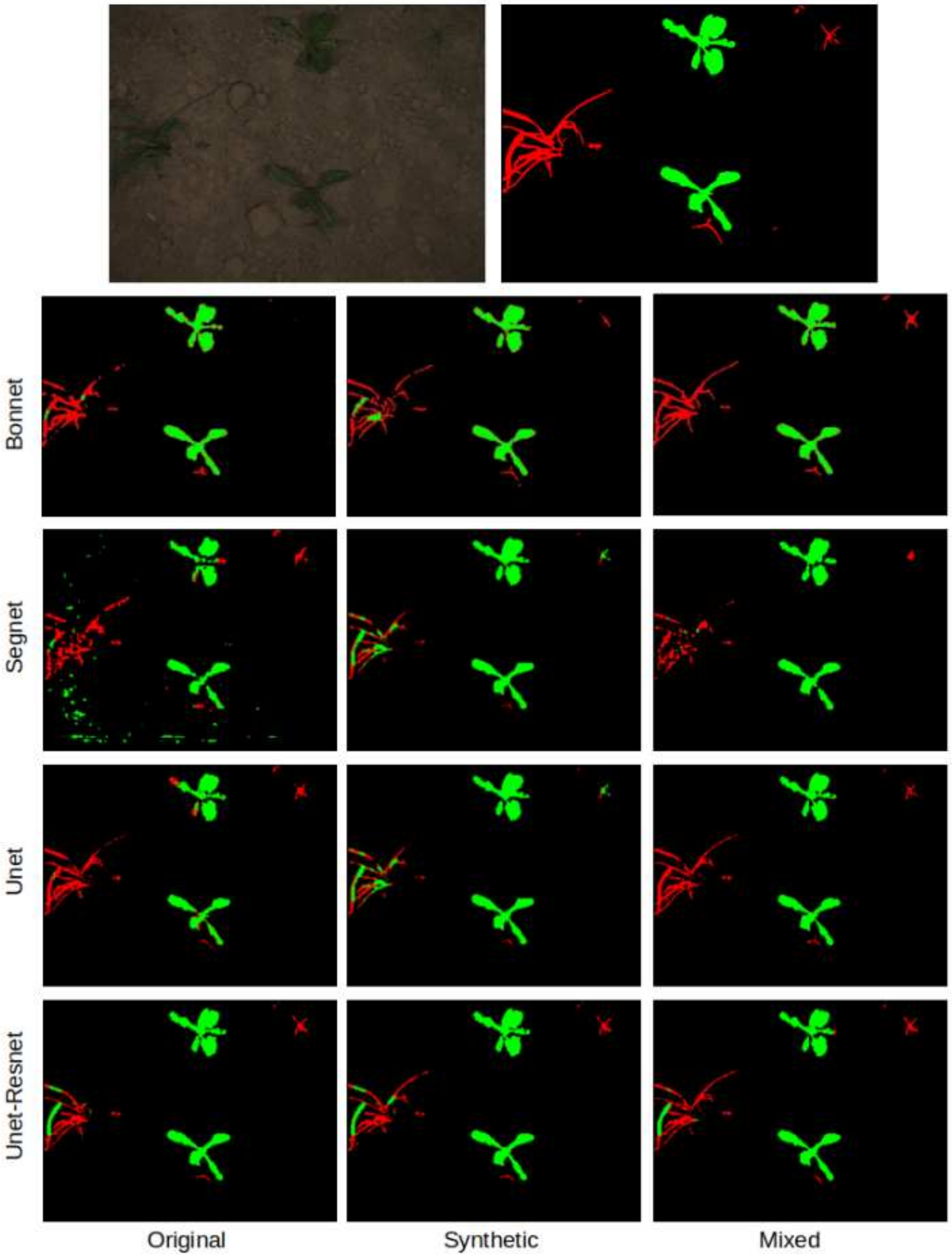

Fig. 4. Examples of a segmented image obtained by using four different segmentation networks trained with three different datasets. The first row of the image shows the input RGB image and its corresponding ground truth. The remaining rows show the segmentation results generated by the different networks on the $Original$, $Synthetic$, and $Mixed$ training datasets.

of the original one only. In the case of the Unet-Resnet architecture, using only the synthetic dataset overcomes the performance obtained by using only the original one.

It is worth noticing that, the term synthetic refers to the substitution of all crops that are more than 50% inside the frame by synthetic ones. For the crops that are mostly out of the frame, the original is kept. In fact it is necessary to have the root of the plant to center the mask for the synthetic image generation.

The source code for the approach described in this paper as well as the data generated using it are publicly available and can be downloaded from.
https://github.com/ibiscp/Synthetic-Plants

V. Conclusions

This paper introduces a data augmentation strategy that leverages a GAN to generate specific synthetic objects, superimposed over real images, in order to generate novel semi-synthetic datasets. The idea is to guide the process of the generative sub-network of the GAN by exploiting a set of segmentation masks of the objects of interest, obtained from the labeling process of a reduced set of real images. The obtained synthetically augmented dataset can then be used to train a semantic segmentation network. We have applied this method to the crop/weed segmentation process, a well-known problem in agricultural robotics, showing that the GAN augmented datasets can improve the performance of four different state-of-the-art segmentation architectures. Plants are overall well segmented even if sometimes their shape is not totally recovered, giving the fact that we are repeatedly using the same plant shapes during training. However, these small inaccuracies do not negatively affect the ability of the net to detect instances of plants.

Future developments include (i) the increase in the generative model output size, allowing to generate higher resolution photo-realistic objects, (ii) the generation of the Near Infra-Red image for the crop, which is commonly used to improve the segmentation performance of the network, (iii) testing the proposed approach in different application scenarios, and (iv) performing an ablation study that consider the use of different masks.

References

[1] T. Duckett, S. Pearson, S. Blackmore, B. Grieve, and M. Smith, "White paper - agricultural robotics: The future of robotic agriculture," 2018.

[2] A. Milioto, P. Lottes, and C. Stachniss, "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, 2018, pp. 2229–2235.

[3] I. Sa, M. Popović, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, and R. Siegwart, "Weedmap: a large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming," *Remote Sensing*, vol. 10, no. 9, p. 1423, 2018.

[4] C. Potena, D. Nardi, and A. Pretto, "Fast and accurate crop and weed identification with summarized train sets for precision agriculture," in *International Conference on Intelligent Autonomous Systems*. Springer, 2016, pp. 105–121.

[5] J. Xie, M. Kiefel, M. Sun, and A. Geiger, "Semantic instance annotation of street scenes by 3d to 2d label transfer," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3688–3697.

[6] N. Chebrolu, P. Lottes, A. Schaefer, W. Winterhalter, W. Burgard, and C. Stachniss, "Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields," *The International Journal of Robotics Research*, 2017.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[8] M. Valerio Giuffrida, H. Scharr, and S. A. Tsaftaris, "Arigan: Synthetic arabidopsis plants using generative adversarial network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2064–2071.

[9] L. Sixt, B. Wild, and T. Landgraf, "Rendergan: Generating realistic labeled data," *Frontiers in Robotics and AI*, vol. 5, p. 66, 2018.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[11] S. Haug, A. Michaels, P. Biber, and J. Ostermann, "Plant classification system for crop/weed discrimination without segmentation," in *IEEE winter conference on applications of computer vision*. IEEE, 2014, pp. 1142–1149.

[12] C. McCool, T. Perez, and B. Upcroft, "Mixtures of lightweight deep convolutional neural networks: Applied to agricultural robotics," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1344–1351, 2017.

[13] A. K. Mortensen, M. Dyrmann, H. Karstoft, R. N. Jørgensen, R. Gislum *et al.*, "Semantic segmentation of mixed crops using deep convolutional neural network," in *Proc. of the International Conf. of Agricultural Engineering (CIGR)*, 2016.

[14] P. Lottes, M. Hörferlin, S. Sander, and C. Stachniss, "Effective vision-based classification for separating sugar beets and weeds for precision farming," *Journal of Field Robotics*, vol. 34, no. 6, pp. 1160–1178, 2017.

[15] M. Fawakherji, A. Youssef, D. Bloisi, A. Pretto, and D. Nardi, "Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation," in *3rd IEEE International Conference on Robotic Computing, IRC 2019, Naples, Italy, February 25-27, 2019*, 2019, pp. 146–152.

[16] M. Di Cicco, C. Potena, G. Grisetti, and A. Pretto, "Automatic model based dataset generation for fast and accurate crop and weeds detection," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5188–5195.

[17] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[20] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, and K. Q. Weinberger, "An empirical study on evaluation metrics of generative adversarial networks," *CoRR*, vol. abs/1806.07755, 2018.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[22] A. Milioto and C. Stachniss, "Bonnet: An Open-Source Training and Deployment Framework for Semantic Segmentation in Robotics using CNNs," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2019.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015. [Online]. Available: http://arxiv.org/abs/1511.00561