DEEP CONVOLUTIONAL NEURAL NETWORKS FOR WEED DETECTION IN AGRICULTURAL CROPS USING OPTICAL AERIAL IMAGES

W. Ramirez[1,*], P. Achanccaray[1], L. F. Mendoza[1], M. A. C. Pacheco[1]

[1] Applied Computational Intelligence Lab, Pontifical Catholic University of Rio de Janeiro - williamramiez2694@aluno.puc-rio.br, (pmad9589, mendonza, marco)@ele.puc-rio.br

KEY WORDS: Semantic Segmentation, Remote Sensing, Deep Neural Network, Precision Farming.

ABSTRACT:

The presence of weeds in agricultural crops has been one of the problems of greatest interest in recent years as they consume natural resources and negatively affect the agricultural process. For this purpose, a model has been implemented to segment weed in aerial images. The proposed model relies on DeepLabv3 architecture trained upon patches extracted from high-resolution aerial imagery. The dataset employed consisted in 5 high-resolution images that describes a sugar beet agricultural field in Germany. SegNet and U-Net architectures were selected for comparison purposes. Our results demonstrate that balancing of data, together with a greater spatial context leads better results with DeepLabv3 achieving up to 0.89 and 0.81 in terms of AUC and F1-score, respectively.

1. INTRODUCTION

Agriculture is one of the most important activities for the economic sector because it supplies a primary need for humankind. For this reason, in conjunction with technology development, there is a constant development of new techniques and studies to improve and optimize agricultural production.

The pipeline of agricultural production is composed of the following processes: land preparation, planting, irrigation, fertilization, and collection, among others. These processes involve large investments in terms of natural and economic resources. Frequently, these investments do not return the expected assets due to different factors like pests, diseases, extreme weather conditions, and the presence of weeds (i.e. plants considered undesirable in a particular condition). Weeds are plants unwanted in human-controlled settings that consume important natural resources from other crops affecting their development.

The use of Unmanned Aerial Vehicles (UAVs) in the agricultural sector has increased in recent years due to its versatility in areas of difficult access, low cost and ease of capturing information from large areas of land, thanks to the advance of the systems of data collection (Bu et al., 2017). UAVs are being employed for monitoring crops' development (Shakhatreh et al., 2019), pests/diseases detection (Lim et al., 2018), weeds localization (Sa et al., 2018a), detection of plants (Pignatti et al., 2019), among others. For these purposes, passive and active sensors are mounted on UAVs to capture spectral and structural information from agricultural fields. For instance, Infrared (IR) cameras capture the range of the electromagnetic spectrum (i.e. Red edge and near-IR) that is more sensitive to the chlorophyll content of plants. Based on this information, weeds can be located in aerial images using semantic segmentation techniques.

Semantic segmentation is the task of assigning a label with semantic meaning to each pixel in an image. Deep Learning (DL) based approaches for semantic segmentation have been achieving the state-of-the-art during the last years due to its potential to learn representative features exploiting contextual information. In addition, DL approaches have been employed for agricultural applications to monitor crops, detect pests, or even to estimate agricultural production.

(Sa et al., 2018b) studied the influence of the dataset employed by the comparison of the results of models trained upon only R, G, B bands, and R, G, B, NIR, CIR, and NDVI. The authors concluded that having information that best describes the problem will help to train models with better inference results. (Sa et al., 2018b) and (Hinzmann et al., 2018) show how to work with a dataset with images of high-resolution and coarse resolution, proposing the use of SegNet architecture and Inception modules.

The remainder parts of this work are organized as follows. Section 2 describes the theoretical fundamentals of the architectures employed in this work. Section 3 explains the methodology used in our study, Section 4 shows the experimental protocol followed in our experiments as well as the results obtained for each of the architectures. Finally, Section 5 summarizes the conclusions extracted from all experiments that were carried out.

2. THEORETICAL FOUNDATIONS

2.1 SegNet

Proposed by the University of Cambridge, SegNet is a deep encoder-decoder architecture (see Figure 1) for multi-class dense (pixel-wise) semantic segmentation. The encoder stage contains 13 convolutional layers taken from the VGG16 architecture (Simonyan, Zisserman, 2014). These layers consist of a sequence of convolutional and pooling layers with the aim of extracting characteristics while reducing the size of the feature map (Badrinarayanan et al., 2017). The decoder stage reuses pooling indices obtained in the max-pooling layers, removing the need for learning in the upsample stage. For this reason, SegNet is considered more efficient than other networks as it requires less computational resources.

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-W12-2020-551-2020 | © Authors 2020. CC BY 4.0 License.
Primary publication at IEEE Xplore: https://doi.org/10.1109/LAGIRS48042.2020.9165562

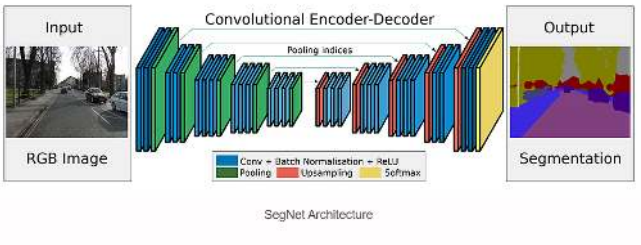

Figure 1. SegNet architecture (taken from (Badrinarayanan et al., 2017)).

2.2 U-Net

Proposed in 2015 by the University of Freiburg in Germany, the U-Net is a fully convolutional network with an encoder-decoder architecture with skip connections (see Figure 2). These skip connections are employed during the pooling operation where pooling indices are saved for being used later. In the encoder stage, the feature maps are extracted, using blocks of convolutional layers with no-padding and max-pooling layers with stride 2 for downsampling. The decoder stage is symmetric to the expanding stage. This returns precise localization using transposed convolutions. This architecture only contains convolutional layers without any dense layer, accepting images of any size (Ronneberger et al., 2015, Lin et al., 2017).

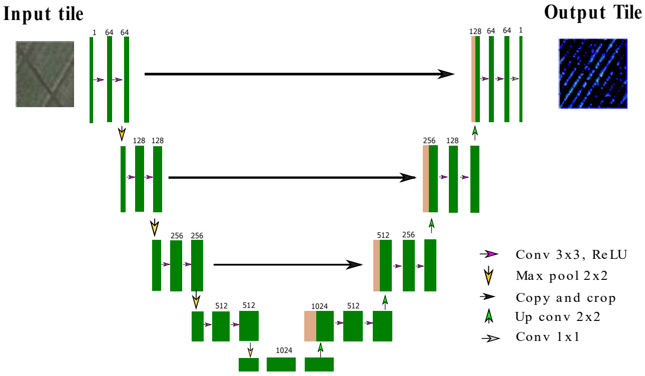

Figure 2. U-net architecture (modified from (Hu et al., 2018)).

2.3 DeepLabv3

Presented by Google, DeepLabv3 architecture has stood out for its innovation in the convolution process by considering dilated convolution, also known as atrous convolution.

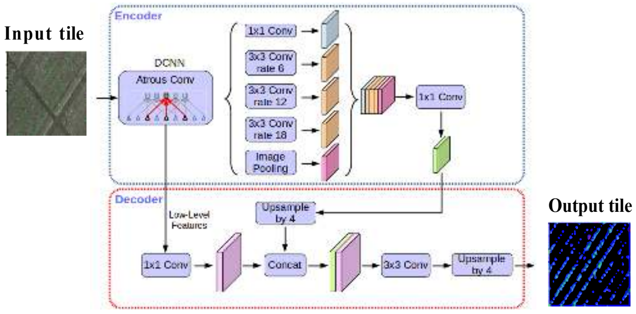

Figure 3. Deeplab arquitecture (taken from (Chen et al., 2018)).

Then, DeepLabv3 captures contextual information at multiple scales using parallel atrous convolutions with different dilation rates (Chen et al., 2018). Fig 3 shows Deeplab v3 architecture, describing the processes considered in the encoder-decoder stage.

3. METHODOLOGY

The methodology employed in this work for weed detection is summarized in Figure 4. There are two main phases: training, where the model learns how to recognize the classes of interest, and testing, where the model performs inference over samples never seen during training.

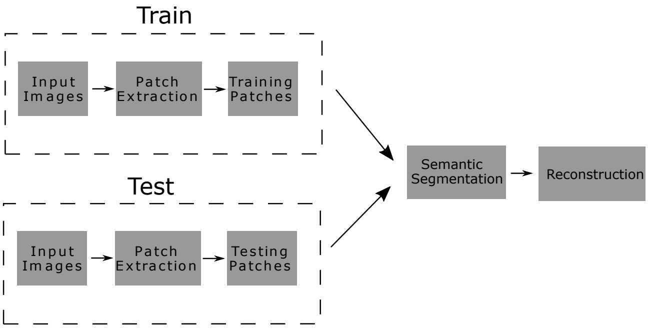

Figure 4. Methodology.

In both phases, training and testing, it is necessary to perform a pre-processing to the input images as they are too big leading to an excessive computational cost as well as a high consumption of resources. For this reason, patches are extracted from each image and its reference with certain overlapping, also known as *stride* (see Figure 5).

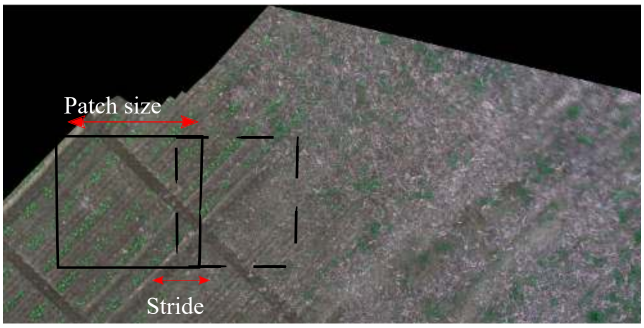

Figure 5. Patch extraction with certain size and overlapping, also known as *stride*.

Once the patches are extracted, the semantic segmentation model is trained upon tuples of patches and their references. This step can be executed many times varying model's hyper-parameters in pursuance of improving model's accuracy. Then, during testing, inference, with the trained model, is performed over images that were not employed during training. Finally, all outputs are joint to reconstruct the entire image as it was divided into patches. For this purpose, it is necessary to save the key of each patch to facilitate its location during the reconstruction step.

4. EXPERIMENTAL ANALYSIS

4.1 Dataset

The dataset corresponds to a sugar beet agricultural field. It comprises a total of 5 aerial images (see detailed description in

Table 1) captured by a drone (Sa et al., 2018b). The dataset also contains tiles extracted from the images, with a size of 480 × 360.

Camera	RedEdge-M				
Image Name	000	001	002	003	004
Width	5995	4867	6403	5470	4319
Height	5854	5574	6405	5995	4506
Area Covered (ha)	0.312	0.1108	0.2096	0.1303	0.1307
GSD (cm)	1.04	0.94	0.96	0.99	1.07
Tile Resolution (row/col) pixels	360 × 480				
# tiles	221	176	252	204	117
Protocol 1	train	train	train	**test**	train
Protocol 2	train	train	train	**test**	**test**
# channels	5				
Crop	Sugar beet				

Table 1. Description of the dataset employed in the experiments. Modified from (Sa et al., 2018b).

This dataset provides R, G, B and NIR (Near-Infrared) bands, as well as CIR (Color-Infrared) and NDVI (Normalized Difference Vegetation Index) images. CIR image is composed by stacking the R, G, and NIR bands. NDVI is computed following the Equation 1.

$$NDVI = \frac{NIR - R}{NIR + R} \tag{1}$$

where R and NIR refer to the Red and Near-Infrared bands respectively.

4.2 Experimental Protocol

The aforementioned methodology in Section 3 have been applied for semantic segmentation using the selected dataset according to the following protocols:

Protocol 1

This protocol is similar to the one employed by (Sa et al., 2018b). For training, patches extracted from images $\{000, 001, 002, 004\}$ and its corresponding references were employed (see Table 1 for more details). Then, inference was performed over patches extracted from image $\{003\}$.

Protocol 2

This protocol was considered to simulate an scenario with less labeled samples than in Protocol 1. In this case, only images $\{000, 001, 002\}$ were employed for training and the remainder ones, $\{003, 004\}$ for inference.

4.3 Experimental Setup

For Protocol 1, patches of size 480 × 360 without overlapping were employed to train all architectures: SegNet, U-Net and DeepLabV3. For Protocol 2, patches of size 512 × 512 with an overlapping of 30% were extracted from the original images in the dataset to train only DeepLabV3.

A set of experiments were carried out varying the hyperparameters of each architecture during training. The best parameter setup obtained in our experiments was the following: Adam optimizer with a learning rate of 10^{-3} and weight decay of 10^{-5} during 50 epochs and early stop to break after 8 epochs without improvement. For validation purposes, 25% of the training patches were selected. During training, we monitor the *mean IoU* (intersection over union) over the validation set. All models are trained on data augmented using random rotations of 30°, vertical and horizontal flips. As the dataset is highly unbalanced, weights proportional to class frequencies at pixel level were considered in the loss function to penalize errors in classes with less samples.

4.4 Quantitative Assessment

The results obtained in our experiments are evaluated in terms of F1-score and AUC (area under the ROC curve).

F1-score is defined as the harmonic mean between *Precision* and *Recall* (see Equation 2) and varies from 0 to 1. Precision and Recall are computed as in Equation 3 and Equation 4, where true positives (*TP*) are results in which the model correctly infers the reference, false positive (*FP*) are those results incorrectly predicted and false negative (*FN*) are results in which the model incorrectly predicts a another class.

The metric AUC describes the relation between the true positives rate and the false positives rate.

$$F1score = \frac{Precision \times Recall}{Precision + Recall} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

4.5 Results

Table 2 and Table 3 summarize the results obtained by both protocols in terms of AUC and F1-score metrics, respectively.

As in (Sa et al., 2018b), the Protocol 1 takes patches of 480 × 360 without stride. Table 2 shows the metric AUC obtained for all architectures, and table 3 the F1-score.

				AUC		
Models	Channels	Size	Protocol	Weed	Crop	Background
SegNet	7	480 × 360	1	0.72	0.81	0.85
DeepLabv3	7	480 × 360	1	**0.83**	**0.89**	**0.92**
U-Net	7	480 × 360	1	0.66	0.77	0.72
DeepLabv3	7	512 × 512	2	0.67	0.79	0.80

Table 2. AUC score per class obtained in the experiments for both protocols.

				F1-score		
Models	Channels	Size	Protocol	Weed	Crop	Background
SegNet	7	480 × 360	1	0.56	0.74	0.85
DeepLabv3	7	480 × 360	1	**0.78**	**0.81**	**0.92**
U-Net	7	480 × 360	1	0.62	0.70	0.97
DeepLabv3	7	512 × 512	2	0.59	0.73	0.86

Table 3. F1-score per class obtained in the experiments for both protocols.

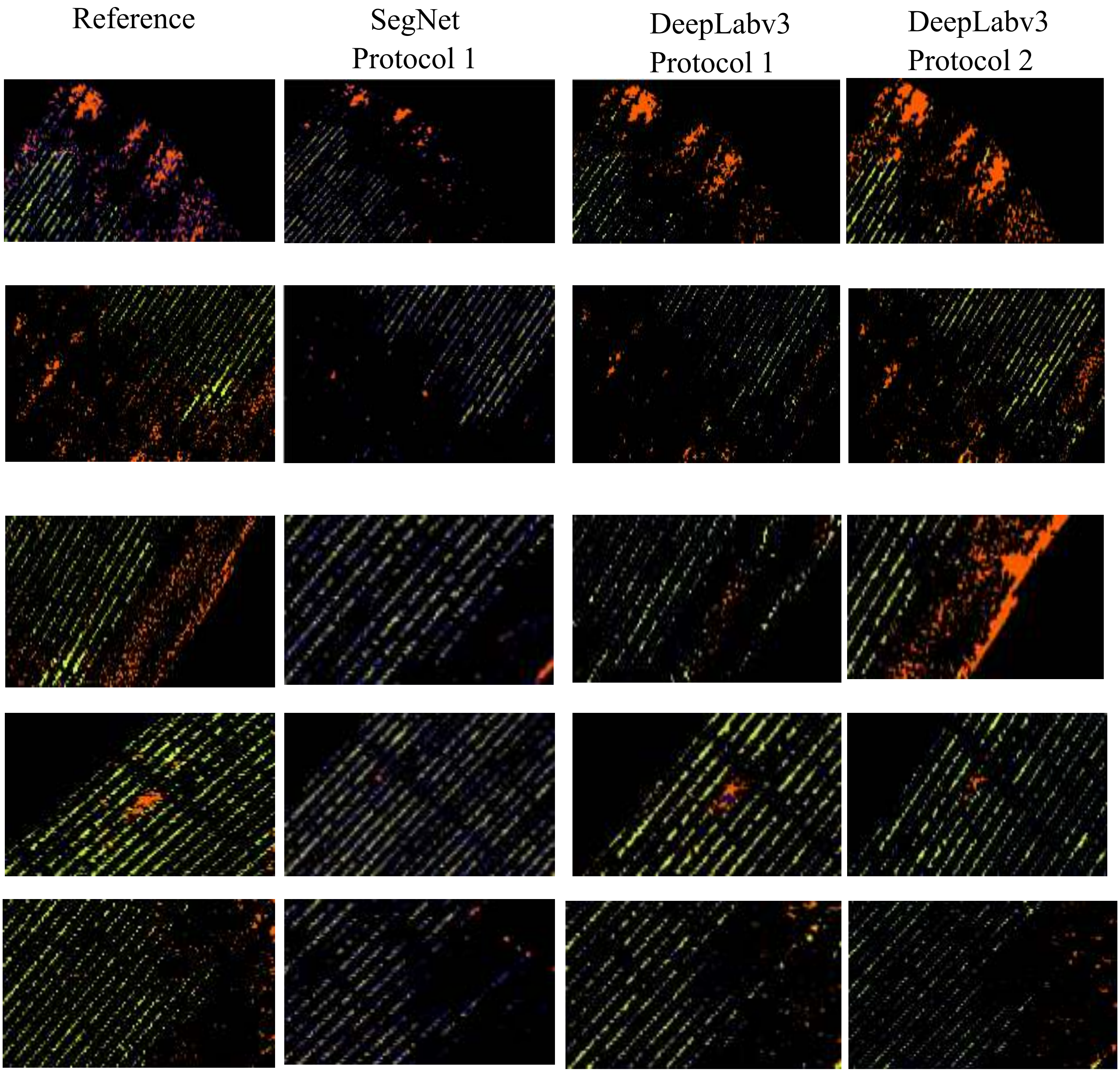

Figure 6. Snips of the classification Maps obtained in the experiments. From left to right: Reference, SegNet (Protocol 1), DeepLabv3 (Protocol 1) and DeepLabv3 (Protocol 2).

In both tables, the highest accuracies are in bold. DeepLabv3 model achieved the highest accuracy because it considers Atrous convolution, having multi-scale feature maps. However, U-Net has a good response in critical areas but classify erroneously more pixels of crops. Figure 6 illustrates this behaviour.

U-Net presents a good response compared to SegNet, because the U-Net architecture takes weights in the corners, identifying very small object with short separations. Consequently, when considering this, it was evident that the model obtained by training the U-Net with the R, G, B channels presented a low performance compared to the U-Net trained upon all 7 channels. The main reason is related to the usage of NDVI channel, which contains information of great relevance regarding the crop, describing a reflectance index for each component in the agricultural field.

Weed is represented in small regions of our data, being represented by small amounts of pixels. DeepLabv3 architecture was able to correctly classify it due to its multiscale atrous convolution representations.

From Protocol 1, it was concluded that DeepLabv3 showed a better behaviour, for this reason it was considered to increase the context of our dataset with the aim of improving the model. In Protocol 2, patches of 512 × 512 with an stride of 30 % were extracted, considering much more spatial context. In this sense, the model obtained presented a considerable improvement, having a better performance in critical areas.

Fig 6 shows the classification maps obtained for all models following Protocols 1 and 2. Notice the better results obtained by DeepLabv3 architecture, observing a better performance in some scenarios.

In Protocol 2 the distribution of the agricultural changes with respect to training and test was changed, because considering only one image caused a bias regarding the behavior of the

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-W12-2020-551-2020 | © Authors 2020. CC BY 4.0 License.
Primary publication at IEEE Xplore: https://doi.org/10.1109/LAGIRS48042.2020.9165562

model against a real situation. It was evident in Table 3, The results from Protocol 2 presented a behavior a bit different from the one following Protocol 1 for the image $\{003\}$, especially in areas where weed was the major class.

5. CONCLUSIONS

A comparison between three different architectures for Semantic Segmentation have been performed in this work. The architectures studied in this work are: SegNet, U-Net and DeepLabv3. DeepLabv3 achieved the highest accuracies with values of up to 0.89 and 0.81 in terms of AUC and F1-score, respectively.

It was observed that the spatial context factor allows to obtain better models. Patches with bigger sizes will affect computationally in training, so it should be considered a patch size that provides an optimal context being computationally approachable.

The difference in the computational resource was evident, using the architectures, SegNet, U-Net and DeepLabv3, the latter being the one that needs the most computational resources. For this reason it is to be considered that it is up to the computational resources to select one or another model. Thus, U-Net architecture using patch size 480×360 was the one that most fit in the relation of computational resource vs results obtained by the model.

It was considered that it is better to have more samples regarding the problem during training. This is because the agricultural fields may have different morphologies, during the cultivation stage, having large regions where the crop is not present.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge Intel Semiconductors Brasil for recognizing ICA Lab at PUC-Rio University as Intel AI Innovation Center and for supplying a Sky Lake, 96 Intel(R) Xeon(R) Platinum 8169 processors.

References

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481-2495.

Bu, J., Sun, R., Bai, H., Xu, R., Xie, F., Zhang, Y., Ochieng, W. Y., 2017. Integrated method for the UAV navigation sensor anomaly detection. *IET Radar, Sonar Navigation*, 11(5), 847-853.

Chen, L., 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(4), 834-848.

Hinzmann, T., Schönberger, J. L., Pollefeys, M., Siegwart, R., 2018. Mapping on the fly: real-time 3d dense reconstruction, digital surface map and incremental orthomosaic generation for unmanned aerial vehicles. *Field and Service Robotics*, Springer, 383–396.

Hu, K., Liu, C., Yu, X., Zhang, J., He, Y., Zhu, H., 2018. A 2.5 d cancer segmentation for mri images based on u-net. *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, IEEE, 6–10.

Lim, S., Kim, S., Park, S., Kim, D., 2018. Development of application for forest insect classification using cnn. *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 1128–1131.

Lin, B. S., Michael, K., Kalra, S., Tizhoosh, H. R., 2017. Skin lesion segmentation: U-nets versus clustering. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7.

Pignatti, S., Casa, R., Harfouche, A., Huang, W., Palombo, A., Pascucci, S., 2019. Maize crop and weeds species detection by using uav vnir hyperpectral data. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 7235–7238.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv*, abs/1505.04597.

Sa, I., Chen, Z., Popović, M., Khanna, R., Liebisch, F., Nieto, J., Siegwart, R., 2018a. weedNet: Dense Semantic Weed Classification Using Multispectral Images and MAV for Smart Farming. *IEEE Robotics and Automation Letters*, 3(1), 588-595.

Sa, I., Popović, M., Khanna, R., Chen, Z., Lottes, P., Liebisch, F., Nieto, J., Stachniss, C., Walter, A., Siegwart, R., 2018b. Weedmap: a large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sensing*, 10(9), 1423.

Shakhatreh, H., Sawalmeh, A. H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., Othman, N. S., Khreishah, A., Guizani, M., 2019. Unmanned Aerial Vehicles (UAVs): A Survey on Civil Applications and Key Research Challenges. *IEEE Access*, 7, 48572-48634.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

Revised February 2020

https://doi.org/10.5194/isprs-archives-XLII-3-W12-2020-551-2020 | © Authors 2020. CC BY 4.0 License.
Primary publication at IEEE Xplore: https://doi.org/10.1109/LAGIRS48042.2020.9165562