SEMANTIC SEGMENTATION OF CITRUS-ORCHARD USING DEEP NEURAL NETWORKS AND MULTISPECTRAL UAV-BASED IMAGERY

Lucas Prado Osco [1,2,*], Keiller Nogueira [3], Ana Paula Marques Ramos [1], Mayara Maezano Faita Pinheiro [1], Danielle Elis Garcia Furuya [1], Wesley Nunes Gonçalves [2,4], José Marcato Junior [2], Jefersson Alex dos Santos [5].

[1] Faculty of Engineering and Architecture and Urbanism; Graduate Program of Environment and Regional Development, University of Western São Paulo (UNOESTE), Presidente Prudente, São Paulo, Brazil; E-mail: lucasosco@unoeste.br; anaramos@unoeste.br; mayarafaita@gmail.com; daniellegarciafuruya@gmail.com

[2] Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul (UFMS), Campo Grande, Mato Grosso do Sul, Brazil; E-mail: jose.marcato@ufms.br

[3] Computing Science and Mathematics Division, University of Stirling, Stirling, FK9 4LA, Scotland, UK; E-mail: keiller.nogueira@stir.ac.uk

[4] Faculty of Computer Science, Federal University of Mato Grosso do Sul (UFMS), Campo Grande, Mato Grosso do Sul, Brazil; E-mail: wesley.goncalves@ufms.br

[5] Department of Computer Science, Federal University of Mato Grosso do Sul (UFMG), Belo Horizonte, Brazil; E-mail: jefersson@dcc.ufmg.br

* Corresponding author: lucasosco@unoeste.br

Abstract: Accurately mapping farmlands is important for precision agriculture practices. Unmanned Aerial Vehicles (UAV) embedded with multispectral cameras are commonly used to map vegetation in these areas. However, separating plantation fields from the remaining objects in a multispectral scene is a difficult task for most traditional algorithms. In this manner, deep learning methods that perform semantic segmentation could help improve the overall outcome. To the best of our knowledge, in the agricultural context, it is yet unknown the performance of deep networks to semantic segmentation in UAV-based multispectral imagery; especially in arboreous vegetation types like citrus-orchards. Here, we evaluate state-of-the-art deep learning methods to semantic segment citrus-trees in multispectral images. For this purpose, we used a multispectral camera that operates at the green (530-570 nm), red (640-680 nm), red-edge (730-740 nm), and also near-infrared (770-810 nm) spectral regions. We evaluated the performance of the five state-of-the-art pixelwise methods: FCN, U-Net, SegNet, DeepLabV3+, and DDCN. Our results indicate that the evaluated methods performed similarly in the proposed task, returning F1-Scores between 94.00% (FCN and U-Net) and 94.42% (DDCN). We also determined the inference time needed per area, and although the DDCN method was slower, based on a qualitative analysis, it performed better in highly shadow-affected areas. We conclude that the semantic segmentation of citrus orchards is highly achievable with deep neural networks. The state-of-the-art deep learning methods investigated here proved to be equally suitable to solve this task, providing fast solutions with inference time varying from 0.98 to 4.36 minutes per hectare. This approach could be incorporated into similar research, and contribute to decision-making and accurate mapping of the plantation fields.

Keywords: precision agriculture; convolutional neural network; remote sensing; thematic map.

1. INTRODUCTION

Effective farming decisions require accurate mapping of agricultural fields. Many techniques were employed for attending this task in the past years, with the majority of them associated with remote sensing approaches (Hunt and Daughtry, 2018; Weiss et al., 2020). In the agricultural context,

remote sensing data is important to monitor nutrient content (Delloye et al., 2018; Osco et al., 2019a), detect water-stress effects (Krishna et al., 2019; Osco et al., 2019b), identify leaf damage (Safonova et al., 2019), predict chlorophyll content (Kalacska et al., 2015; Shah et al., 2019), yield estimation (Chen et al., 2017; Hunt et al., 2019; Jin et al., 2019; Sun et al., 2019), among others. Most of these tasks were performed at leaf and/or canopy level, with data collected by different types of sensors at proximal, terrestrial, aerial and orbital platforms (Surový et al., 2018; Ozdarici-Ok, 2015; Paoletti et al., 2018; Osco et al., 2020a). Unmanned Aerial Vehicle (UAV) platforms gained more lately attention in many application areas, including precision agriculture mainly because of its relatively low-cost and high capacity to map areas with very high spatial-resolution. UAV-based images are largely used to substitute the visual inspection of agricultural landscapes since some practices are often labeled as labor-intensive, biased, and time-consuming (Leiva et al., 2017).

The aforementioned examples were achieved with a combination of different methodologies, which consist of regression/correlation analysis, spectral indices, morphological operations, spectral classification, etc. But another type of technique applied to data associated with remote sensing comes from artificial intelligence. Recently, deep learning is quickly gaining momentum as a method for image processing and data analysis (Goodfellow et al., 2016). Deep learning is a type of machine learning technique that is constructed as a deeper type of artificial neural network that allows hierarchical data representation (LeCun et al., 2015; Ghamisi et al., 2017; Badrinarayanan et al., 2017). A deep neural network can be constructed with different kinds of layers, which tends to improve its performance and returns a larger learning capability than most common networks or other types of learners (LeCun et al., 2015; Ball et al., 2017). Although known for a high demand for computational power and high requirement for labeled data, deep neural networks have achieved impressive performances in many tasks, such as image classification (Krizhevsky et al., 2012; Nogueira et al., 2017), semantic segmentation (Badrinarayanan et al., 2017; Nogueira et al., 2019a), object detection (Ren et al., 2015; Nogueira et al., 2019b; Santos et al., 2020; Osco et al., 2020b) and others.

Different components constitute the architecture of a deep neural network, and among the most frequently adopted architectures, Convolutional Neural Networks (CNN) have presented a better performance, in general, for image and pattern recognition (Alshehhi et al., 2017). As for agricultural studies, approximately 42% of the deep learning architectures implemented were based on CNN (Kamilaris and Prenafeta-Boldú, 2018). The most common components of a CNN architecture are convolution and deconvolution layers, pooling and max-pooling layers, fully-connected layers, activation functions, and others (Goodfellow et al., 2016). Regarding remote sensing, data extraction methods consider the spectral (Ghamisi et al., 2017), spatial (Li et al., 2017) and spectral-spatial information (Zhang et al., 2017). Approaches that consider both spectral and spatial information in their model can improve estimates significantly (Zhang et al., 2017). This has been the most common strategy when dealing with vegetation analysis and deep networks in the last few years (Djerriri et al.,

2018; Li et al., 2017; Csillik et al., 2018; Safonova et al., 2019; Weinstein et al., 2019; Osco et al., 2020b).

Deep learning-based methods are a fairly new concept in agricultural practices involving the usage of remote sensing imagery. Some of which include classification, localization, and object detection tasks (Djerriri et al., 2018; Li et al., 2017; Hassanein et al., 2019; Wu et al., 2019; Csillik et al., 2018; Fan et al., 2018; Safonova et al., 2019; Osco et al., 2020), while others are related to image segmentation, like semantic or instance (Nogueira et al., 2016; Kamilaris and Prenafeta-Boldú, 2018). The deep learning method chosen to deal with a specific task is linked to the scene, data and also the target's characteristics. When counting plants in a high-density object detection scenario, we discovered, in previous research in a citrus-orchard (Osco et al., 2020b), that a dense map refined in a multi-stage type of CNN architecture was better than state-of-the-art bounding-box methods. Still, the idea of separating the detected vegetation from the remaining objects in the image is more appropriate with a segmentation process. Semantic segmentation is able to assign a class-label to every pixel of an image. Deep learning-based approaches designed to tackle this task receive, as input, an image, that may be composed of a given number of bands, and return, as output, another image, generally with the same size of the input data with each pixel associated with one class. This outcome, commonly called a thematic map, may help in the full understanding of the scene which, in turn, may assist several applications, including disaster relief (Nogueira et al., 2018), urban planning (Vakalopoulou et al., 2015; Nogueira et al., 2019a) and others.

In agriculture-related problems, most semantic segmentation processes with deep neural networks use RGB (Red-Green-Blue) images or include a combination with other information to help solve a specific issue. For orange-fruit detection and segmentation, a Mask R-CNN architecture (He et al., 2017) was proposed in a combination with RGB and RGB + HSI (Hue-Saturation-Intensity) images (Ganesh et al., 2019). The SegNet architecture (Badrinarayanan et al., 2017) was also compared with FCN (Long et al., 2015) method in an RGB data-set for rice lodging identification (Yang et al., 2020). Another study, implementing RGB and near-infrared (NIR) information from a sensor embedded in a ground-robot, was able to transfer the knowledge from a network trained on a different crop to semantic segment weeds with the SegNet-Basic architecture (Bosilj et a., 2019). SegNet was also used to segment out trunks, branches, and trellis wires in apple-tree canopies in RGB image (Majeed et al., 2020). An FCN (Long et al., 2015), in conjunction with UAV RGB-based imagery, was used for winter-wheat ear segmentation (Ma et al., 2020). Nonetheless, up to the time of writing, no literature information regarding the semantic segmentation of images in a non-RGB domain capture with UAV based systems, specifically related to agriculture datasets, was found.

In arboreous vegetation types, the automatic delineation in images often requires information related to the spectral heterogeneity, shadow complexity and background effects (Nevalainen et al., 2017). These techniques mostly rely on the spectral divergence between the trees and non-tree types of a pixel. Normally, brighter pixels are associated with the vegetation while darker pixels are viewed as their boundary (Özcan et al., 2017). In agricultural fields, an important culture worldwide is citrus,

and recent deep learning-based methods have been proposed to assist precision farming in different orchards. A study (Ampatzidis and Partel, 2019) investigated the performance of the YOLOv3 network to detect and count citrus-trees, archiving high precision and accuracies values. A variation of a CNN with a refinement algorithm based on superpixels (Csillik et al., 2018) was used in an object detection approach to also count the trees. As mentioned, one of our previous studies (Osco et al., 2020) was conducted in a citrus-orchard, and we proposed an object detection approach that, differently from the previous methods, performed better in high-density plantations. Although the individual detection of a tree is important for many agriculture practices, this high-density is a reality in many citrus-orchards. This may be a problem for most state-of-the-art deep learning methods of object detection (Ampatzidis and Partel, 2019). In this manner, a semantic segmentation, while performing a different type of approach (i.e. not used to count trees), could be used to properly map these highly-dense areas and estimate vegetation-cover and used as in input for plantation-lines extractions.

Separating and extracting plantation fields of a remote sensing image provides important information on plantation cover-area and location. The semantic segmentation of plantation fields by deep networks is a novel and improved computational manner to accurately separate the vegetation from the remaining objects in an image-scene. As a benefit, it should provide an accurate mapping of the area while demanding low effort from the human counterpart. Novel deep learning-based methods are being constantly proposed, and their robustness assessment should be performed in several applications. We intend, here, to fill a part of this gap in the agriculture context. To the best of our knowledge, no study investigated the performance of deep learning-based methods to fulfill this task in UAV-based multispectral imagery, especially for citrus-orchards. In this paper, we evaluate the performance of five state-of-the-art deep neural networks to semantic segment citrus-trees in multispectral images. The rest of this paper is organized as follows: Section 2 provides highly-detailed info of the methods applied; Section 3 exposes the obtained results while also discusses their implications. Finally, Section 4 concludes this study.

2. MATERIALS AND METHOD

Our workflow was divided into five main stages (Figure 1). We began collecting our data (a) in a Citrus-orchard area where the UAV flight was performed. Later, we processed our image data (b) and generated an orthomosaic. We then processed our Citrus data by labeling it in a Geographical Information System (GIS) environment and split it into groups training and testing subsets (c). We proceed to perform the semantic segmentation process (d) in a computational environment, selecting five state-of-the-art deep neural networks for the proposed task. Finally, we performed the evaluation (e) of the tested methods and compared it against each other. Details regarding these processes are described in the following subsections.

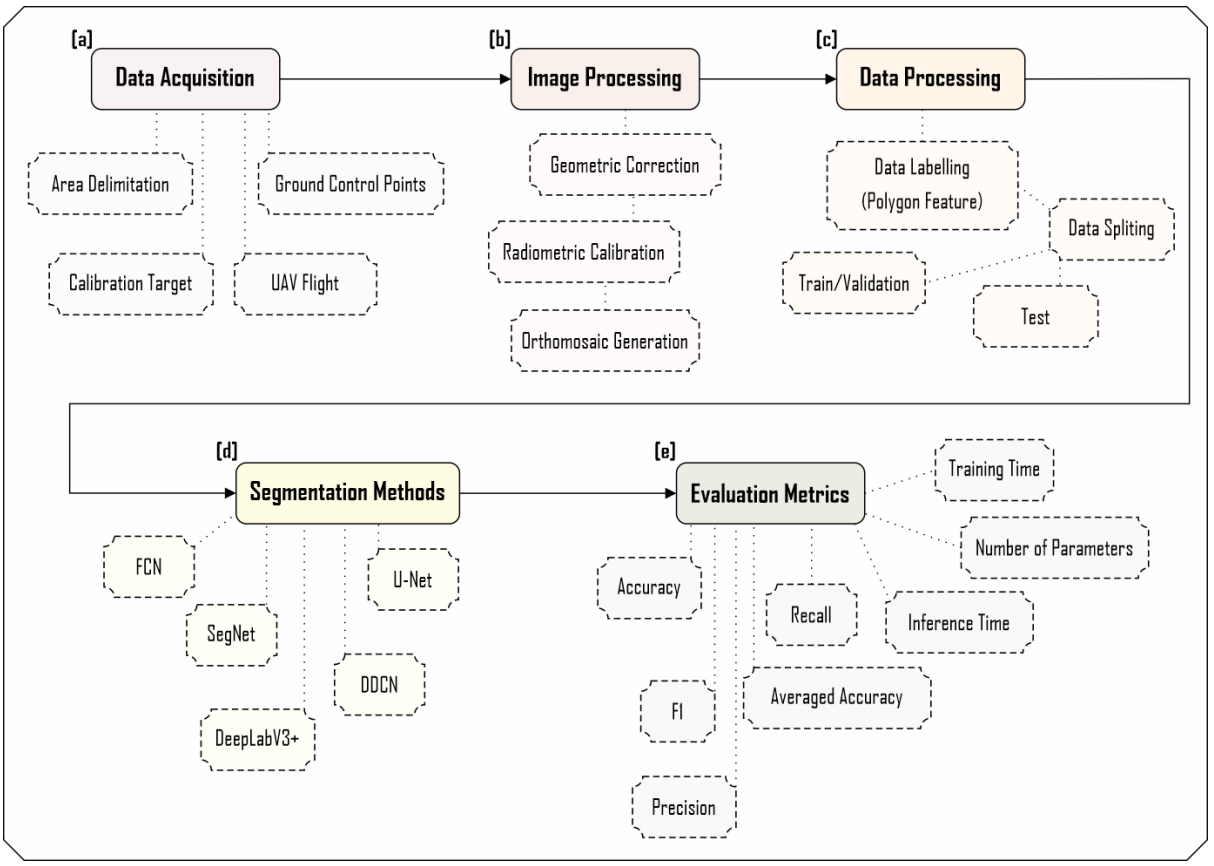

Figure 1. Workflow summarizing the fundamental steps of the conducted approach.

2.1 Data Acquisition and Image Processing

To compose our experiment's data-set, we used a Valencia-orange orchard (planted over a Citrumelo Swingle rootstock), located in the interior of the municipality of Ubirajara, São Paulo, Brazil. During our survey, the Citrus trees were in their vegetative stage, around 5-years old, with 3-meter averaged high ground-related. The labeled area (Figure 2) is around 70.4 ha, with high-density trees at 7x1.9 meters in-line spacing, resulting in a total of 750 trees per hectare. Aside from trees, the UAV flight also registered dirt-roads between the plantation plots, streets for locomotion, different densities of grasses, buildings, and other objects. Months prior to the flight, the orchard's soil was fertilized with 250 kg/ha of Nitrogen in the form of Urea, 125 kg/ha of Phosphorus excreted, expressed as P_2O_5, and 167 kg/ha of Potassium Oxide (K_2O). The field is predominantly composed of red-yellow podzolic soil, situated in a C_WA Köppen subtropical climate type unit.

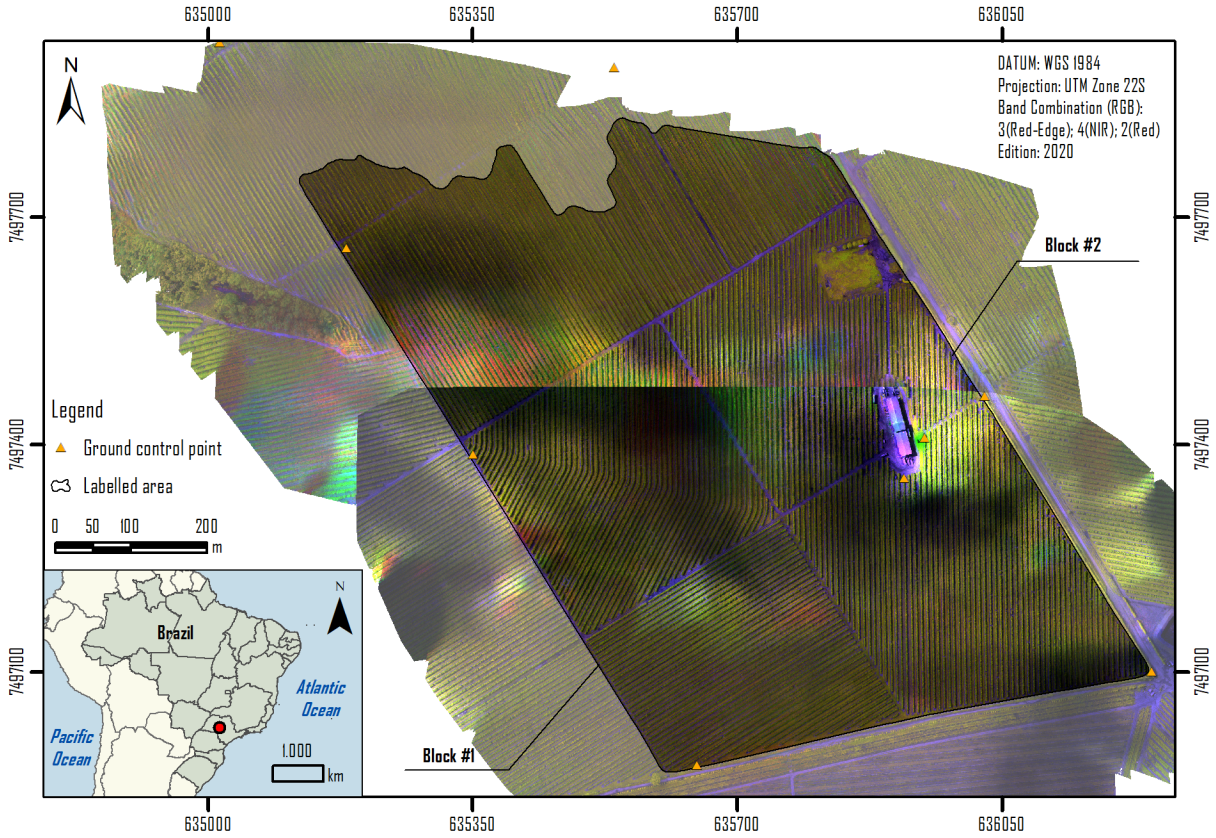

Figure 2. Citrus-orchard area used to evaluate state-of-the-art deep neural network methods.

We conducted two flights with an eBee SenseFly UAV platform shipped with a Parrot Sequoia camera to perform the imaging of the area. Parrot Sequoia camera records images in the following spectral regions: green (530-570 nm), red (640-680 nm), red-edge (730-740 nm) and near-infrared (770-810 nm). The flight took place at the end of the summer (south-hemisphere), March 22, 2018, in partially cloudy conditions, 29 °C and 0 mm precipitation, with light air breeze at 1 to 2 m/s from the northeast (Yr, 2018). We adopted a flight altitude of 120 meters-high in relation to the terrain. At this height, Parrot Sequoia can generate images with a GSD (Ground Sample Distance) equivalent to 12.9 cm. We program two flight routes to register areas beyond the boundaries of the area of interest. The first flight occurred at 13:30, while the second followed around 14:15 (local time). Alongside both flights, we registered a total of 9 control points (Figure 2) throughout the area to support the phototriangulation process. Each point was surveyed with the GNSS (Global Navigation Satellite System) Leica Plus GS15, in RTK (Real-Time Kinematic) mode, remaining in operation between 11:00 and 17:00 (local time). The RTK position per point report registered an accuracy of 0.003 m.

From the conducted flights, we processed two different blocks of images (Figure 2), whose first block was formed by 1,183 images and the second one was formed by 1,206 images. We used the Pix4Dmapper software for the calibration processes. We optimized the interior and exterior parameters and created a sparse-dense cloud. The 9 control points were given as reference locations to optimize the Structure-From-Motion (SfM) method. Dense point-clouds were generated based on

the MVS (Multi-View Stereo) method. The RMSE (Root Mean Square Error) of this process was about 0.129 meters. We generated the DSM (Digital Surface Model) of both blocks. For the Parrot Sequoia radiometric calibration method, we registered, prior to both flights, a calibration plaque inherent to this sensor. We converted the Digital Number (DN) values to surface reflectance using the calibration parameters described in the Parrot Sequoia manual. An orthorectified surface reflectance image was generated for each band at each block (I and II). Both image blocks were used to create a unique mosaic. The orthomosaic was composed of 2,389 scenes altogether.

2.3 Semantic Segmentation Methods And Experimental Setup

Five state-of-the-art methods were investigated in this study: (i) Fully Convolutional Network (FCN) (Long et al., 2015), (ii) U-Net (Ronneberger et al., 2015), (iii) SegNet (Badrinarayanan et al., 2017), (iv) Dynamic Dilated Convolution Network (DDCN) (Nogueira et al., 2019a), and (v) DeepLabV3+ (Chen et al., 2018).

2.3.1 Fully Convolutional Network (FCN)

Fully Convolutional Network (FCN) (Long et al., 2015) was one of the primary deep learning-based techniques proposed to perform semantic segmentation. This network extracts features and generates an initial coarse classification map using a set of convolutional layers that, due to their internal configuration, outputs a spatially reduced (when compared to the original input) outcome. In order to restore the original resolution and output the thematic map, this approach employs deconvolution layers (Zeiler and Fergus, 2014) that learn how to upsample the initial classification map and produce the final dense prediction.

The FCN architecture experimented in this work is presented in Figure 3. This network has 6 convolutional layers (in which the first three are followed by a pooling layer) responsible to extract the features and generate an initial coarse classification. This outcome is further processed by 3 deconvolution layers (Zeiler and Fergus, 2014), which are responsible to spatially upsample the prediction producing the final dense output, with the same height and width of the input image. The input of the last two deconvolutions is, in fact, the output of the previous layer combined (via element-wise addition) with the prediction generated by an extra convolutional layer that receives, as input, the features extracted by one of the pooling layer (as presented in Figure 3). This concept of combining features from multiple layers to produce the final prediction is advantageous, as it allows the model to exploit low-, mid- and high-level information, captured from the input data, to generate the final outcome.

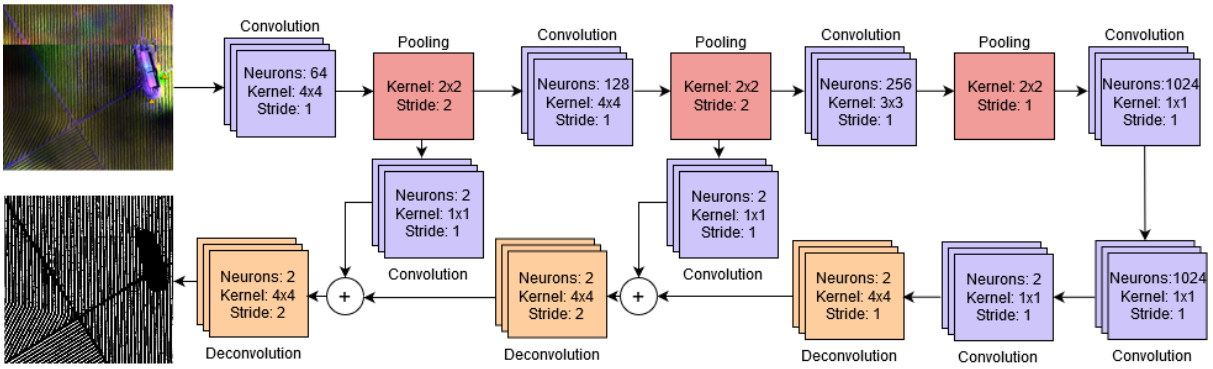

Figure 3. Fully Convolutional Network architecture (Long et al., 2015).

2.3.2 U-Net

U-Net (Ronneberger et al., 2015) was one of the first networks to propose encoder-decoder architectures to perform semantic segmentation. In this design, the encoder is usually composed of several convolution and pooling layers, and responsible to extract the features and generate an initial coarse prediction map. The decoder, commonly composed of convolution, deconvolution (Zeiler and Fergus, 2014) and/or unpooling layers (Goodfellow et al., 2016), is responsible to further process the initial prediction map, increasing its spatial resolution gradually and generating the final prediction. Note that, normally, the decoder can be seen as a mirrored/symmetrical version of the encoder, with the same number of layers but replacing some of the operations with their counterparts (i.e., convolution with deconvolution, pooling with unpooling, etc).

The U-Net architecture exploited in this work is presented in Figure 4. The encoder of this network is composed of two blocks, each composed of two convolutions and one pooling layer, plus a final convolutional layer. It receives the input image and outputs coarse feature maps four times smaller. The decoder is composed of a single convolutional layer followed by two blocks, each consisting of deconvolution and two convolution layers. This part receives the coarse feature map (produced by the encoder) and outputs the final fine prediction image. Two interesting aspects of this architecture should be highlighted. The first one is that the downsampling process performed by the pooling operations in the encoder is reverted using deconvolutional layers in the decoder, i.e., the pooling layers in the encoder are replaced by deconvolutions in the decoder phase. These are the only layers capable of changing the spatial resolution of the data. The second one is that after each deconvolutional layer, there is an operation that concatenates the features produced by this layer with the ones extracted from the convolution before a pooling layer (as presented in Figure 4). Similarly to the FCN (Long et al., 2015), this is performed so that the model is able to exploit multi-level features to improve the final prediction.

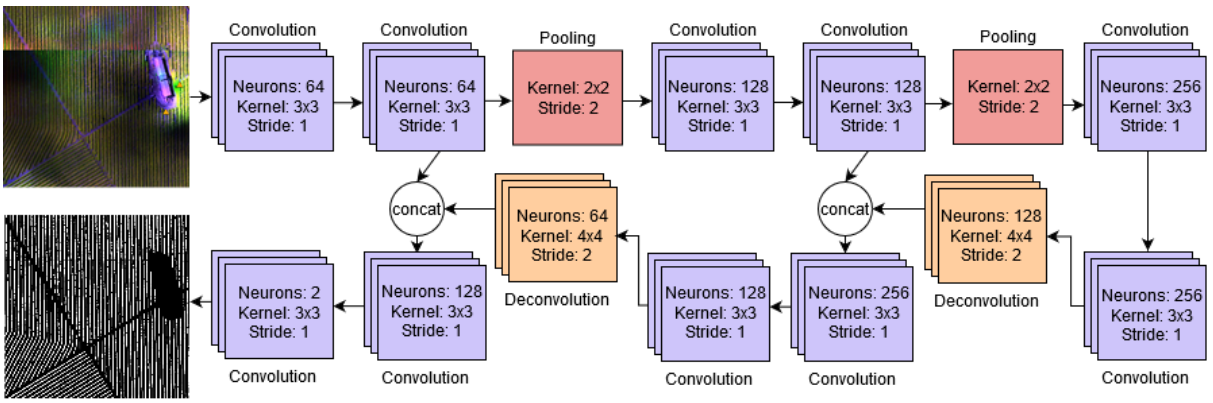

Figure 4. UNet architecture (Ronneberger et al., 2015).

2.3.3 SegNet

SegNet (Badrinarayanan et al., 2017) is another type of encoder-decoder network proposed specifically for semantic segmentation. However, differently from the previous model, this network employs unpooling operations, instead of deconvolution layers, in the decoder to increase the spatial resolution of the coarse map generated by the encoder. Figure 5 presents the SegNet architecture exploited in this work. The encoder of this network is composed of three blocks, each one composed of two convolutions and one pooling layer. It receives the input image and outputs a coarse feature map six times smaller. The decoder also has three blocks, each one composed of an unpooling and two convolution layers. This part receives the coarse feature map (produced by the encoder) and outputs the final fine prediction image. In the exploited architecture, each pooling layer of the encoder is directly replaced by an unpooling layer in the decoder. These are the only operations able to change the spatial resolution of the input, with the former one reducing it and the latter one restoring it.

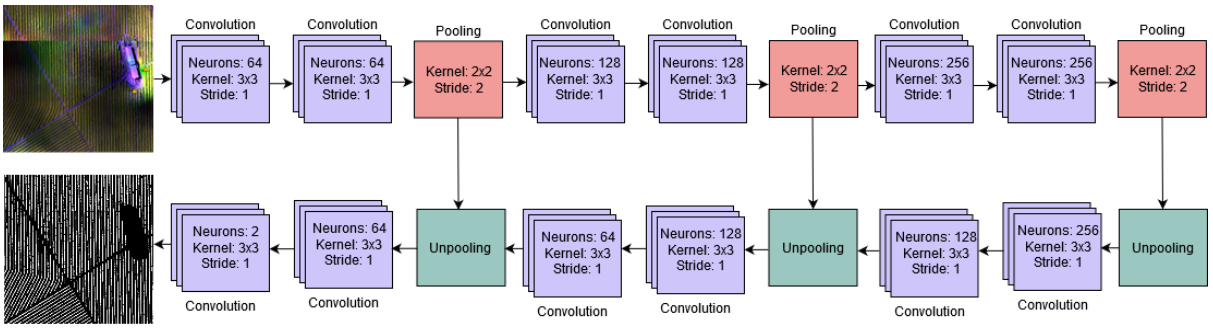

Figure 5. SegNet architecture (Badrinarayanan et al., 2017).

2.3.4 DeepLabV3+

More recently, researchers observed that smoother predictions could be produced if the input image was not considerably downsampled. However, the conservation of the data resolution over the network would imply the model's inability to efficiently explore the receptive field concept (i.e., the

input area of influence on the output) (Goodfellow et al., 2016). This is an important drawback given that when outputting dense predictions, it is critical for each output pixel to have a big receptive field, such that no important information is left out when making the prediction (Luo et al., 2016). To overcome this, dilated convolutions (Yu and Koltun, 2015) were introduced. Such layers are capable of increasing the receptive field without downsampling the input. The DeepLab networks (Chen et al., 2014, Chen et al., 2017a, Chen et al., 2017b, Chen et al., 2017b) were one of the first to exploit the benefits of dilated convolutional layers (Yu and Koltun, 2015). Such models propose to use some initial convolutional layers to moderately reduce the input resolution, which is then kept constant by the final (dilated and standard) convolutions.

Specifically, in this work, we evaluated the latest version of the DeepLab networks, i.e., the DeepLabV3+ (Chen et al., 2018), whose architecture is presented in Figure 6. This network starts with three blocks, each composed of two convolutions followed by one pooling layer, responsible to learn an initial representation. Such features are further processed by a special module, called Atrous Spatial Pyramid Pooling, composed of several parallel dilated convolution layers (Yu and Koltun, 2015) that process the same input features using distinct dilation rates, thus allowing the model to capture multi-scale information. This representation is concatenated with low-level features extracted from the first pooling and then processed by one extra convolutional layer. Finally, three convolutional layers further process the concatenated features, which are then upsampled by a bilinear interpolation to produce the final dense prediction map.

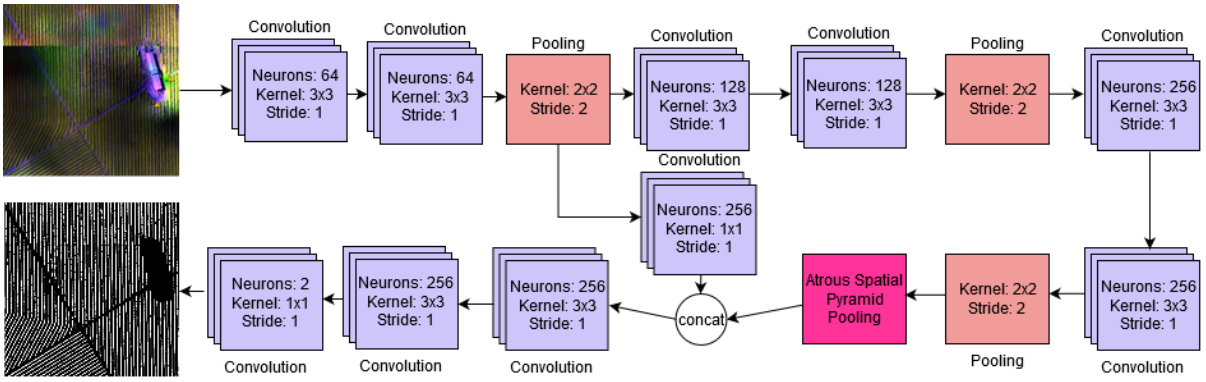

Figure 6. DeepLabV3+ architecture (Chen et al., 2018).

2.3.5 Dynamic Dilated Convolutional Network (DDCN)

Dynamic Dilated Convolutional Network (DDCN) (Nogueira et al., 2019a) takes the previous concept of preserving the input image resolution to the extreme. Specifically, this approach proposes a novel multi-scale training strategy that uses dynamically-generated input images to converge a dilated model that never downsamples the input data. Technically, this technique receives as input the original images and a probability distribution over the possible input sizes, i.e., over the sizes that might be used to generate the input patches. In each iteration of the training procedure, a size is randomly selected from this distribution and is then used to create a totally new batch. By processing

these batches, each composed of several images with one specific pre-selected size, the model is capable of capturing multi-scale information. Furthermore, in the prediction step, the algorithm selects, based on scores accumulated during the training phase for each evaluated input size, the best resolution. Then, the technique processes the testing images using batches composed of images with the best-evaluated size.

Although several models (including FCNs, U-Nets, and SegNet) could be used with the aforementioned training strategy, they might have complications if the input size is too small to the point that it does not allow the creation of the coarse map and, consequently, of the final thematic map. To overcome this, Nogueira et al. (2019a) proposed to use a network composed entirely of dilated convolutions (Yu and Koltun, 2015) that never reduces the input image. This fully dilated model fits perfectly into the proposed multi-scale training strategy as it is capable of processing inputs of any size. Following this concept, Figure 7 presents the fully dilated network architecture tested in this work. It has a total of 8 dilated blocks, each one composed of a dilated convolution and a pooling layer, followed by a standard convolutional layer responsible for the final prediction. It is important to observe that, although pooling layers are employed, they do not reduce the spatial resolution of the input due to a specific configuration of stride and padding.

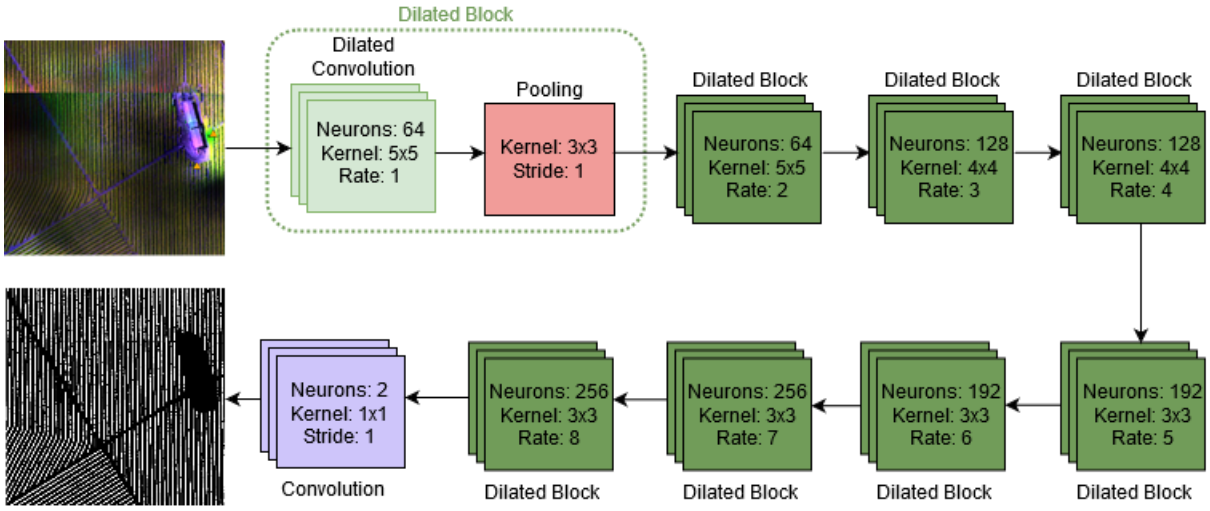

Figure 7. DDCN architecture (Nogueira et al., 2019a).

2.3.6 Protocol

All models were trained using the same training/test protocol. Precisely, the original data, collected over the citrus orchard (Figure 2), was divided into training and test sets, as presented in Figure 8. The former set, consisting of approximately ⅔ of the total amount of labeled pixels, was employed to converge the networks whereas the latter one, composed of the remaining labeled pixels (⅓), was used to evaluate the models. Obtained results are reported in terms of accuracy, averaged accuracy, F1-Score, precision and recall values based on the performance on the test set.

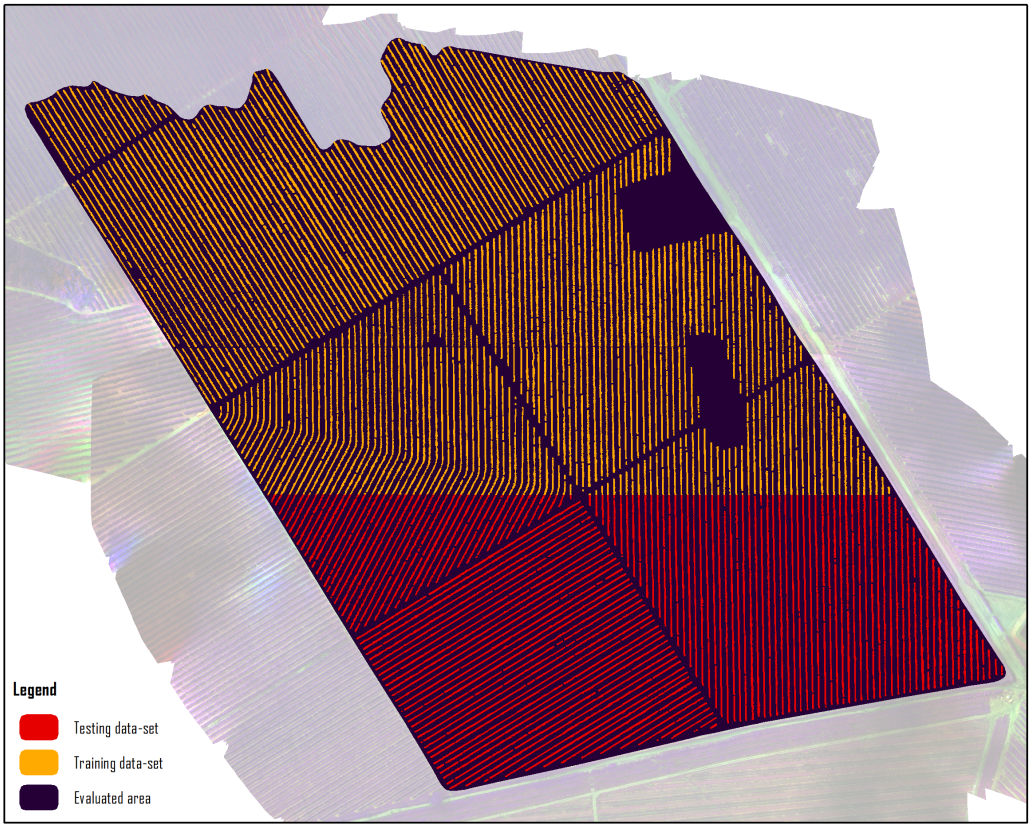

Figure 8. Training (yellow-copper color) and test (red-rust color) sets in proportions of ⅔ (66.67%) and ⅓ (33.34%), respectively.

Aside from this, it is important to emphasize that all aforementioned networks were trained from scratch. All of them employed input patches of 32x32 pixels, except the DDCN (Nogueira et al., 2019a) that used a uniform distribution that allows the method to select an input resolution from 2 possibilities: 32x32, and 64x64. Note that other input patch sizes were evaluated but they did not produce significant improvement in the results only increasing the training time. During training, all the approaches used the same set of hyperparameters, which was defined based on previous convergence analyses. Specifically, the learning rate, weight decay, momentum, and the number of iterations were 0.01, 0.005, 0.9, and 200,000, respectively. After every 50,000 iterations, the learning rate was reduced following an exponential decay with parameter 0.5.

All deep learning-based models exploited in this work were implemented using TensorFlow, a Python framework conceived to allow efficient exploitation of deep learning with Graphics Processing Units (GPUs). The code will be made publicly available after acceptance of the work. All experiments conducted were performed on 64-bit Intel i7-8700K@3.70GHz CPU workstation, with 64 GB memory, and NVIDIA® GTX 1080 GPU with 12Gb of memory, under a 10.0 CUDA version. Debian 4.19.98-1 version was used as the operating system.

3. RESULTS AND DISCUSSION

The evaluated deep learning methods returned similar accuracies in the proposed approach, ranging from 94.88% to 95.46%. (Table 1). Although a slight difference aside indicates that the DDCN performed better in a quantitative point-of-view, it is accurate to say that all of the five state-of-the-art networks are capable of segmenting citrus-trees satisfactorily in the multispectral imagery data-set. This information is important, as of until the moment, no agricultural field segmentation was evaluated with this kind of spatial-spectral data. The fact that these deep neural networks are able to highly separate vegetated covered-area from other targets while maintaining the original resolution of the image input is an important characteristic. Multispectral images are largely used for monitoring vegetation health, as in multiple precision farming applications (Citation, Year). By accurately mapping the plantation area with a few false-positives, it is possible to extract more accurate information solely from the culture itself.

Table 1. Evaluation metrics obtained for the experimented methods.

Pixelwise Methods	Accuracy (%)	Avr. Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
FCN (Long et al., 2015)	94.88	94.28	94.00	95.83	96.77
U-Net (Ronneberger et al., 2015)	94.96	93.59	94.00	97.10	95.71
SegNet (Badrinarayanan et al., 2017)	94.96	94.13	94.06	96.25	96.49
DeepLabV3+ (Chen et al., 2018)	95.15	94.49	94.31	96.20	96.81
DDCN (Nogueira et al., 2019a)	95.46	94.28	94.42	97.04	96.64

Another information obtained from our experiment was the processing time needed to perform the training and inference (Table 2). Most of the evaluated methods returned proximal inference time for both GPU and CPU tests, with the exception of the DDCN method (Nogueira et al., 2019a), which took around four times the amount of time needed to perform the same task. However, an estimation of this inference time per area demonstrates how rapidly these neural networks can segment trees in the given data-set once they are trained. This information is crucial for precision agriculture since this response could be incorporated into decision-making regarding area-size and priority. Regardless, it should be noted that the times informed here are considering the system used to train these methods (see Section 2.3.6). Despite that, this information is rarely considered when performing this task, and future research could benefit from the intention exposed here.

Table 2. Information obtained with the experimented methods.

Pixelwise Methods	Number of Parameters (in millions)	Training Time (GPU hours)	Inference Time (GPU and CPU min.)	Inference Time per Area (GPU and CPU min./ha)

FCN (Long et al., 2015)	3.83	21.5	5.4 and 6.0	0.23 and 0.25
U-Net (Ronneberger et al., 2015)	1.86	21.3	5.4 and 6.0	0.23 and 0.25
SegNet (Badrinarayanan et al., 2017)	2.32	21.4	5.4 and 6.0	0.23 and 0.25
DeepLabV3+ (Chen et al., 2018)	5.16	21.7	5.4 and 6.0	0.23 and 0.25
DDCN (Nogueira et al., 2019a)	2.08	64.2	22.8 and 24.0	0.97 and 1.02

Our tested area was composed of ⅓ of the total of the experimental field site (Figure 2). As previously stated, this experimental area is around 70.4 ha, of which our test-data is around 23.5 ha. During the image labeling process, we verified that our ground-truth data (i.e. citrus-trees labeled as polygon features) occupied approximately 9.2 ha and 5.5 ha in both training and testing data-sets, respectively. The data-sets were composed of different plantation line orientations/directions (Figure 8) and the same target, objects, and challenges. Figure 9 displays the resulting segmentation of the five state-of-the-art methods evaluated in this study.

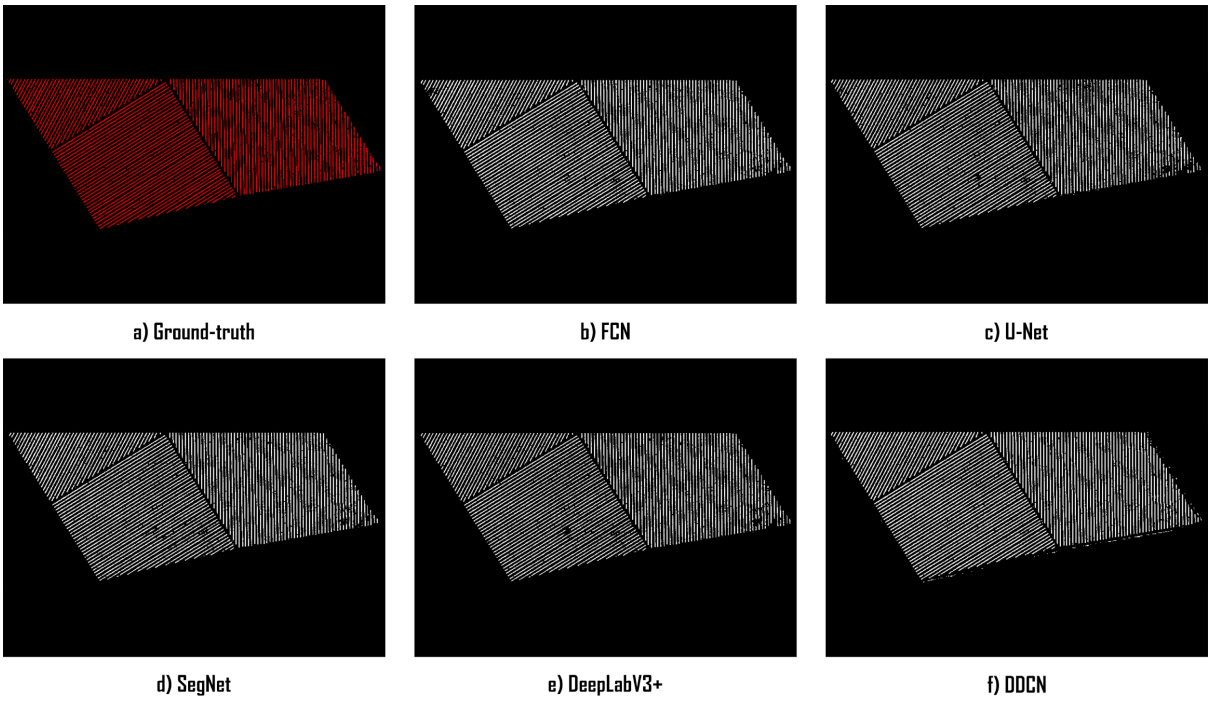

Figure 9. Ground-truth and visual results of the evaluated methods. From top-to-left: Ground-truth, FCN (Long et al., 2015), U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), DeepLabV3+(Chen et al., 2018), DDCN (Nogueira et al., 2019a).

Given the size of the area evaluated, flight time and atmospheric conditions, it produced a heterogeneous data-set with different lumination conditions (see Section 2.2). Figure 10 demonstrates a comparison with a false-color combination and the segmented result from the quantitatively overall

best method (DDCN). In the top-row of Figure 10, it is noticeable that the segmentation performance was satisfactory, although differences in illumination geometry and in the orientation of the plantation-line should oppose a hindrance for the network to handle. However, in the bottom-row of Figure 10, it is notable that the segmentation worsened. This condition could be explained by two major factors in the data-set: highly shadow-affected areas, and; highly dense-grassland areas.

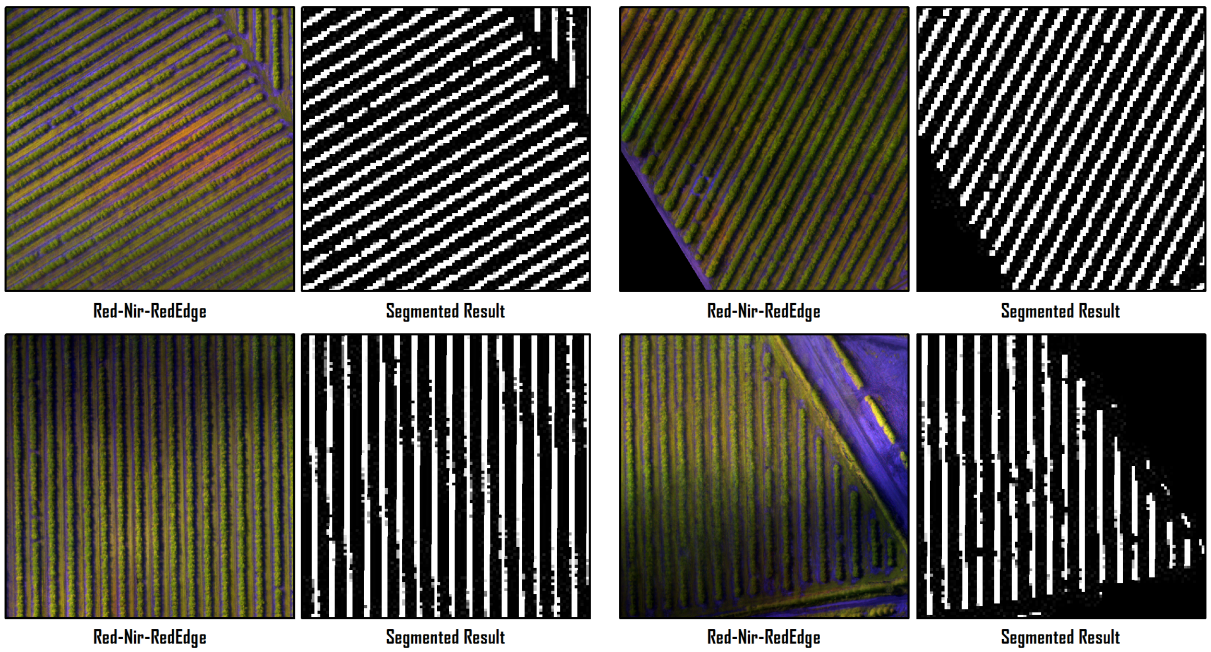

Figure 10. Challenges faced by the investigated approach (false-color combination compared with the DDCN method).

As stated, a qualitative evaluation of the results indicated that most of the problems faced by the investigated deep networks are related to shadowy areas. In the previous comparison (Figure 10) we observed only the DDCN result. As an example, Figure 11 highlights all the five methods in one particular highly shadow-affected area. All networks, with the exception of the DDCN (Nogueira et al., 2019a), performed quite poorly here. The DDCN method captures multi-scale information and selects the best resolution based on scores accumulated during the training phase. This difference aside may have helped this network to map more accurately these shadowy areas than the other methods, even though it returned a slight amount of accuracy in quantitative terms (Table 1). However, as for recommending the application of this architecture over the others, the processing and inference times are something to be taken into consideration (Table 2).

The spectral similarity among objects in a scene is known to be a potential problem for most image processes (Citation, Year). In the orchard data-set, most of the plantation-streets are protected by a living-cover (i.e. grassland). This practice, in Brazil citrus-plantations, has become customary in commercial sites since it assists in water infiltration, environmental control, and soil-erosion protection. However, for remote sensing imagery, this type of additional protection in land produces

high spectral similarity with the citrus culture. This characteristic was mostly discernible with the Red-edge and the Near-infrared regions (Figure 12). Interestingly, this was not particularly a problem in plantation-gaps, where a line was interrupted by a cut tree. This is because not many of these areas were filled with highly-dense grassland, as single-trees were previously chopped because of greening infestations.

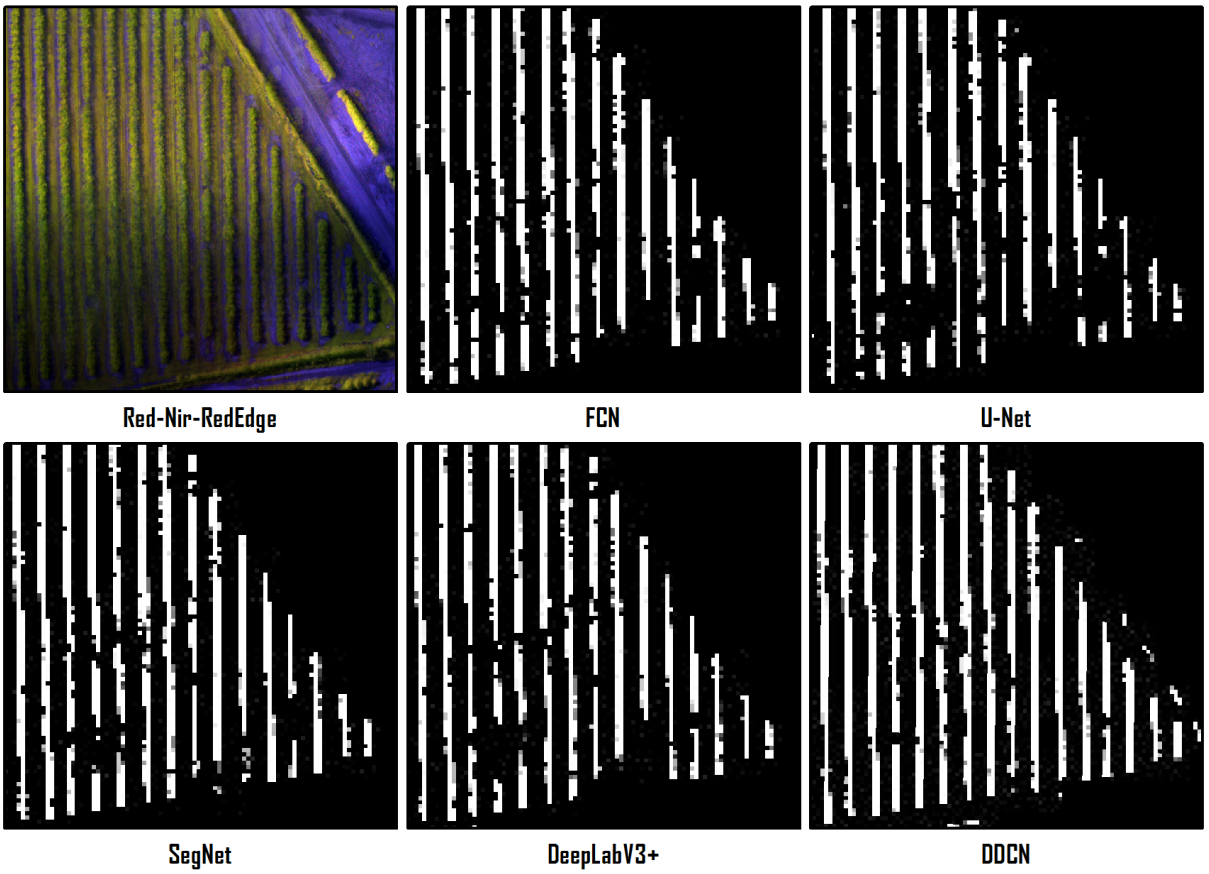

Figure 11. Highly shadow-affected area and segmentation results from the tested methods.

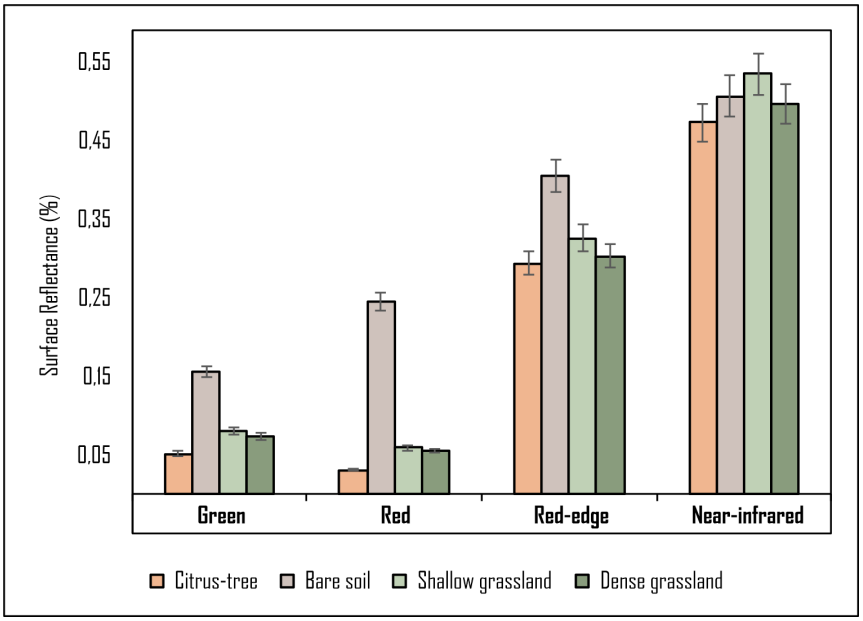

Figure 12. The spectral behavior of the objects most commonly present in the scene. This graphic was constructed with the selection of multiple pixels of a minimum of one-hundred samples per object.

Although the aforementioned challenges could impose somewhat of a problem for most of the image processing methods, state-of-the-art deep neural networks returned highly accurate results (i.e. above 94%). The remaining percentage not classified in our test data-set could be explained by the qualitative approach presented. But indeed, CNNs are also commonly known for producing errors in image-boundaries (Citation, Year), which could too explain the cap reached here. The importance of this approach, as stated, comes with the robustness assessment of these methods in a challenging multispectral data-set, in the agricultural context with UAV-based imagery. From the practical view, the significance of segmenting almost pixel-exclusively the plantation itself, while knowing the inference time needed per area, is a valuable advantage. For precision agriculture practices, the addition of the deep learning methods demonstrated and its subsequent results could auxiliary decision-making.

4. CONCLUSION

We conducted experiments in order to evaluate five deep neural networks to semantic segment citrus-trees in UAV-based multispectral images. Our study demonstrated that the semantic segmentation is highly appropriate for separating and extracting plantation fields of remote sensing imagery in the evaluated conditions. Our data indicate that the investigated methods performed similarly in the proposed task, returning accuracies between 94.88% (FCN) and 95.46% (DDCN). Based on a qualitative analysis, the DDCN method performed better in highly shadow-affected areas. We conclude that the semantic segmentation of citrus orchards is highly achievable with state-of-the-art deep networks. The deep learning methods investigated provided fast solutions to segment the plantation cover-area, with an inference time varying from 0.98 to 4.36 minutes per

hectare. The framework presented here may help other studies while providing primary information for exploring these methods in said context. Our approach could be also incorporated into similar agricultural areas and contribute to decision-making and accurate mapping of plantation fields.

Funding: This research was funded by CNPq (p: 303559/2019-5, 433783/2018-4 and 304173/2016-9), CAPES Print (p: 88881.311850/2018-01) and Fundect (p: 59/300.066/2015 and 59/300.095/2015).

Acknowledgments: The authors acknowledge the support of UFMS (Federal University of Mato Grosso do Sul) and CAPES (Finance code 001).

Conflicts of 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.

REFERENCES

Alshehhi, R., Marpu, P. R., Woon, W. L., Mura, M. D., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. **ISPRS J. Photogramm. Remote Sens.**,130, 139–149.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. **IEEE Trans. Pattern Anal. Mach. Intell.**, 39(12), 2481–2495.

Ball, J. E., Anderson, D. T., Chan, C. S. 2017. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. **J. Appl. Remote Sens.**, 11(04), 042609.

Bosilj P, Aptoula E, Duckett T, Cielniak G. Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. **J. F. Robot.** 2020;37(1):7–19.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. **arXiv** preprint arXiv:1412.7062.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **IEEE Trans. Pattern Anal. Mach. Intell.**, 40(4), 834-848.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. 2017b. Rethinking atrous convolution for semantic image segmentation. **arXiv** preprint arXiv:1706.05587.

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. **In** Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

Chen, S. W., Shivakumar, S. S., Dcunha, S., Das, J., Okon, E., Qu, C., … Kumar, V., 2017c. Counting apples and oranges with deep learning: A data-driven approach. **IEEE Robot. Autom**. Lett., 2(2), 781–788.

Csillik, O., Cherbini, J., Johnson, R., Lyons, A., Kelly, M., 2018. Identification of citrus trees from unmanned aerial vehicle imagery using convolutional neural networks. **Drones**, 2(4), 39-55.

Delloye, C., Weiss, M., Defourny, P., 2018. Retrieval of the canopy chlorophyll content from Sentinel-2 spectral bands to estimate nitrogen uptake in intensive winter wheat cropping systems. **Remote Sens. Environ**., 216, 245–261.

Djerriri, K., Ghabi, M., Karoui, M. S., Adjoudj, R., 2018. Palm trees counting in remote sensing imagery using regression convolutional neural network. **Proc. IGARSS**, 2627–2630.

Fan, Z., Lu, J., Gong, M., Xie, H., Goodman, E. D., 2018. Automatic tobacco plant detection in UAV images via deep neural networks. **IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.**, 11(3), 876–887.

Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. 2018. **Comput. Electron. Agric.** 145, 311–318.

Ganesh P, Volle K, Burks TF, Mehta SS. Deep Orange: Mask R-CNN based Orange Detection and Segmentation. **IFAC-PapersOnLine**. 2019;52(30):70–5.

Ghamisi, P., Plaza, J., Chen, Y., Li, J., Plaza, A. J., 2017. Advanced spectral classifiers for hyperspectral images: A review. **IEEE Geosci. Remote Sens. M.**, 5(1), 8–32.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. **Deep learning**. MIT press, 2016.

Hassanein, M., Khedr, M., El-Sheimy, N., 2019. Crop row detection procedure using low-cost UAV imagery system. **ISPRS Archives**, 42(2/W13), 349–356.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. 2017. Mask r-cnn. **In** Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Hunt, E. R., Daughtry, C. S. T., 2018. What good are unmanned aircraft systems for agricultural remote sensing and precision agriculture? **Int. J. Remote Sens.**, 39(15–16), 5345–5376.

Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S., 2019. High resolution wheat yield mapping using Sentinel-2. **Remote Sens. Environ.**, 233, 111410.

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., & Lobell, D. B., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. **Remote Sens. Environ**., 228, 115–128.

Long, J., Shelhamer, E., & Darrell, T. 2015. Fully convolutional networks for semantic segmentation. **In** Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

Luo, W., Li, Y., Urtasun, R., & Zemel, R. 2016. Understanding the effective receptive field in deep convolutional neural networks. **In** Advances in neural information processing systems (pp. 4898-4906).

Kamilaris, A., & Prenafeta-Boldú, F. X., 2018. Deep learning in agriculture: A survey. **Comput. Electron. Agric.**, 147, 70–90.

Kang, H.; Chen, C. Fast implementation of real-time fruit detection in apple orchards using deep learning. 2020. **Comput. Electron. Agric.**, 168, 105108.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. **In** Advances in neural information processing systems (pp. 1097-1105).

Lecun, Y., Bengio, Y., & Hinton, G., 2015. Deep learning. **Nature**, 521(7553), 436–444.

Leiva, J. N., Robbins, J., Saraswat, D., She, Y., & Ehsani, R., 2017. Evaluating remotely sensed plant count accuracy with differing unmanned aircraft system altitudes, physical canopy separations, and ground covers. **J. Appl. Remote Sens.**,11(3), 036003.

Li, W., Fu, H., Yu, L., & Cracknell, A., 2017. Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. **Remote Sens.**, 9(1), 22-35.

Ma J, Li Y, Du K, Zheng F, Zhang L, Gong Z, et al. Segmenting ears of winter wheat at flowering stage using digital images and deep learning. **Comput. Electron. Agric**. 2020;168(December 2019):105159.

Majeed Y, Zhang J, Zhang X, Fu L, Karkee M, Zhang Q, et al. Deep learning based segmentation for automated training of apple trees on trellis wires. **Comput. Electron. Agric**. 2020;170(August 2019):105277.

Nogueira, K., Dalla Mura, M., Chanussot, J., Schwartz, W. R., & dos Santos, J. A. 2016. Learning to semantically segment high-resolution remote sensing images. **In** 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 3566-3571). IEEE.

Nogueira, K., Penatti, O. A., & Dos Santos, J. A. 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. **Pattern Recognition**, 61, 539-556.

Nogueira, K., Fadel, S. G., Dourado, Í. C., Werneck, R. D. O., Muñoz, J. A., Penatti, O. A., ... & Torres, R. D. S. 2018. Exploiting ConvNet diversity for flooding identification. **IEEE Geoscience and Remote Sensing Letters**, 15(9), 1446-1450.

Nogueira, K., Dalla Mura, M., Chanussot, J., Schwartz, W. R., & dos Santos, J. A. 2019a. Dynamic multicontext segmentation of remote sensing images based on convolutional networks. **IEEE Transactions on Geoscience and Remote Sensing**, 57(10), 7503-7520.

Nogueira, K., Cesar, C., Gama, P. H., Machado, G. L., & dos Santos, J. A. 2019b. A Tool for Bridge Detection in Major Infrastructure Works Using Satellite Images. **In** 2019 XV Workshop de Visão Computacional (WVC) (pp. 72-77). IEEE.

Ozdarici-Ok, A., 2015. Automatic detection and delineation of citrus trees from VHR satellite imagery. **Int. J. Remote Sens.**, 36(17), 4275–4296.

Osco, L. P., Paula, A., Ramos, M., Pereira, D. R., Akemi, É., Moriya, S., … Matsubara, E. T., 2019a. Predicting canopy nitrogen content in citrus-trees using random forest algorithm associated to spectral vegetation indices from UAV-imagery. **Remote Sens.**, 2019, 11(24), 2925-2942;

Osco, L. P., Arruda, M. S., Junior, J. M., da Silva, N. B., Ramos, A. P. M., Moriya, É. A. S., ... Li, J., 2020b. A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery. **ISPRS J. Photogramm. Remote Sens.**, 160, 97-106.

Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A., 2018. A new deep convolutional neural network for fast hyperspectral image classification. **ISPRS J. Photogramm. Remote Sens.**, 145, 120–147.

Ren, S., He, K., Girshick, R., & Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. **In** Advances in neural information processing systems (pp. 91-99).

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 (pp. 234-241). Springer, Cham.

Safonova, A., Tabik, S., Alcaraz-Segura, D., Rubtsov, A., Maglinets, Y., & Herrera, F., 2019. Detection of fir trees (Abies sibirica) damaged by the bark beetle in unmanned aerial vehicle images with deep learning. **Remote Sens.**, 11(6), 643-462.

Sun, J., Di, L., Sun, Z., Shen, Y., & Lai, Z. 2019. County-level soybean yield prediction using deep CNN-LSTM model. **Sensors**, 19(20), 1–21.

Too, E.C.; Yujian, L.; Njuki, S.; Yingchun, L. 2019. A comparative study of fine-tuning deep learning models for plant disease identification. **Comput. Electron. Agric.**, 161, 272–279.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., & Paragios, N. 2015. Building detection in very high resolution multispectral data with deep learning features. **In** 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 1873-1876). IEEE.

Weinstein, B. G., Marconi, S., Bohlman, S., Zare, A., & White, E., 2019. Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks. **Remote Sens.**, 11(11), 1309-1322.

Weiss, M., Jacob, F., & Duveiller, G. 2020. Remote sensing for agricultural applications: A meta-review. **Remote Sens. Environ.**, 236(November 2019), 111402.

Wu, J., Yang, G., Yang, X., Xu, B., Han, L., & Zhu, Y., 2019. Automatic counting of in situ rice seedlings from UAV images based on a deep fully convolutional neural network. **Remote Sens.**, 11(6), 691-710.

Yang M Der, Tseng HH, Hsu YC, Tsai HP. 2020. Semantic segmentation using deep learning with vegetation indices for rice lodging identification in multi-date UAV visible images. **Remote Sens**.12(4).

Yr. Norwegian Meteorological Institute. **Norwegian Broadcasting Corporation**. Editor: Jensen, I. S. Available at: < https://www.yr.no/ >. Accessed in: 12 Jul. 2019.

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

Zeiler, M. D., & Fergus, R. 2014. Visualizing and understanding convolutional networks. **In** European conference on computer vision (pp. 818-833). Springer, Cham.

Zhang, H., Li, Y., Zhang, Y., & Shen, Q., 2017. Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. **Remote Sens. Lett.**, 8(5), 438–447.

Zhong, Y.; Zhao, M. Research on deep learning in apple leaf disease recognition. 2020. **Comput. Electron. Agric**., 168, 105146.