ISPRS Journal of Photogrammetry and Remote Sensing 160 (2020) 97–106

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery

Lucas Prado Osco[a,*], Mauro dos Santos de Arruda[b], José Marcato Junior[a], Neemias Buceli da Silva[b], Ana Paula Marques Ramos[d], Érika Akemi Saito Moryia[e], Nilton Nobuhiro Imai[e], Danillo Roberto Pereira[f], José Eduardo Creste[c], Edson Takashi Matsubara[b], Jonathan Li[g], Wesley Nunes Gonçalves[a,b]

[a] Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Brazil
[b] Faculty of Computer Science, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil
[c] Faculty of Agronomy, University of Western São Paulo, Presidente Prudente, São Paulo, Brazil
[d] Faculty of Engineering and Architecture, University of Western São Paulo, Presidente Prudente, São Paulo, Brazil
[e] Department of Cartographic Science, São Paulo State University, Mailbox: 19060-900, Presidente Prudente, SP, Brazil
[f] Faculty of Computer Science, University of Western São Paulo, Presidente Prudente, São Paulo, Brazil
[g] Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Deep learning
Multispectral image
UAV-borne sensor
Object detection
Citrus tree counting
Orchard

ABSTRACT

Visual inspection has been a common practice to determine the number of plants in orchards, which is a labor-intensive and time-consuming task. Deep learning algorithms have demonstrated great potential for counting plants on unmanned aerial vehicle (UAV)-borne sensor imagery. This paper presents a convolutional neural network (CNN) approach to address the challenge of estimating the number of citrus trees in highly dense orchards from UAV multispectral images. The method estimates a dense map with the confidence that a plant occurs in each pixel. A flight was conducted over an orchard of Valencia-orange trees planted in linear fashion, using a multispectral camera with four bands in green, red, red-edge and near-infrared. The approach was assessed considering the individual bands and their combinations. A total of 37,353 trees were adopted in point feature to evaluate the method. A variation of σ (0.5; 1.0 and 1.5) was used to generate different ground truth confidence maps. Different stages (T) were also used to refine the confidence map predicted. To evaluate the robustness of our method, we compared it with two state-of-the-art object detection CNN methods (Faster R-CNN and RetinaNet). The results show better performance with the combination of green, red and near-infrared bands, achieving a Mean Absolute Error (MAE), Mean Square Error (MSE), R^2 and Normalized Root-Mean-Squared Error (NRMSE) of 2.28, 9.82, 0.96 and 0.05, respectively. This band combination, when adopting $\sigma = 1$ and a stage ($T = 8$), resulted in an R^2, MAE, Precision, Recall and F1 of 0.97, 2.05, 0.95, 0.96 and 0.95, respectively. Our method outperforms significantly object detection methods for counting and geolocation. It was concluded that our CNN approach developed to estimate the number and geolocation of citrus trees in high-density orchards is satisfactory and is an effective strategy to replace the traditional visual inspection method to determine the number of plants in orchards trees.

1. Introduction

Unmanned aerial vehicle (UAV) platforms can deliver ultra-high spatial resolution images, offer versatility in adverse weather conditions and permit a flexible revisit time (Varela et al., 2018). In precision agriculture, remote sensing data provided by UAV-borne sensors has assisted farmers in the management of their fields (Deng et al., 2018; Hunt and Daughtry, 2018). However, different factors (e.g., plant and ground characteristics, environmental factors) can contribute to the complexity of an image used for plant field analysis (Leiva et al., 2017). In addition, different analytical methods have been employed to evaluate this complexity better.

The use of Deep-Learning (DL) algorithms have increased in remote sensing applications (Zhang et al., 2011a,b; Alshehhi et al., 2017; Ball et al., 2017; Liu et al., 2018; Liu and Abd-Elrahman, 2018; Paoletti et al., 2018; Ma et al., 2019). DL algorithms based on convolutional

* Corresponding author.
E-mail address: pradoosco@gmail.com (L.P. Osco).

https://doi.org/10.1016/j.isprsjprs.2019.12.010
Received 8 August 2019; Received in revised form 7 November 2019; Accepted 11 December 2019
0924-2716/ © 2019 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

neural network (CNN) have presented a high performance for different types of application in image data from agricultural fields (Kamilaris and Prenafeta-Boldú, 2018; Wu et al., 2019). These applications involve the analysis of wheat spikes (Hasan et al., 2018), wheat-ear density estimation (Madec et al., 2019), rice seedlings in the field (Wu et al., 2019) and the counting of fruits (Chen et al., 2017), plants (Djerriri et al., 2018) and trees (Jiang et al., 2017; Li et al., 2017) in crop fields.

Information as to the number of plants in a crop field is essential for farmers because it helps them estimate productivity, evaluate the density of their plantations and errors occurring during the seedling process (Ampatzidis and Partel, 2019). However, counting plants is a labor-intensive and time-consuming task (Leiva et al., 2017). To address this issue, recent researches have investigated the potential of the CNN approach applied to images obtained from UAV-borne sensors (Djerriri et al., 2018; Onish and Ise, 2018; Salamí et al., 2019). One type of agricultural activity that relies heavily on plant counting data is the tree type (Li et al., 2017).

Different techniques in UAV-borne sensor imagery have been implemented to identify and count trees (Goldbergs et al., 2018). Until recently, the delineation of different tree rows from UAV data was consistently used for this task (Jakubowski, et al., 2013; Tao et al., 2015; Verma et al., 2016). Also, automatic detection and delineation methods are being used for trees in agricultural fields, such as citrus plantations (Ozdarici-Ok, 2015). Recently, the implementation of CNN in UAV image produced high precision results, up to 99.9% (Ampatzidis and Partel, 2019) and 94.59% (Csillik et al., 2018),

Although studies have given high accuracy in counting citrus trees using CNN in UAV multispectral images, the current methodology (Ampatzidis and Partel, 2019; Csillik et al., 2018) is based on object detection CNNs. These CNNs use rectangles to detect each plant individually, but their detection and performance decrease as the image becomes crowded and the plant size decreases (Kang et al., 2019). In such cases, the boundaries of individual plants may not be sufficiently visible to detect a rectangle, which may increase the difficulty of discriminating individual plants. Up to the time of writing, the performance of CNN to count citrus trees considering a high-density orchard is still unknown.

This paper addresses the mentioned gap and presents a CNN approach to cope with the challenge of estimating the number of citrus trees in highly dense orchards from UAV multispectral images. Our method not only provides the counting but also the geolocation of each tree, similar to object detection methods. The rest of this paper is organized as follows: Section 2 provides a literature review of tree detection and crown delineation; Section 3 shows the study area and materials used; Section 4 details the proposed method; Section 5 presents and discusses the experimental results, and; Section 6 concludes the paper.

2. Related works

Automated tree detection in computer vision presents different challenges since its performance can be affected by sensor characteristics and tree complexity (Ozdarici-Ok, 2015). Sensors used in this task involve UAV-based and satellite systems (Jiang et al., 2017; Varela et al., 2018; Ozdarici-Ok, 2015; Zhang et al., 2016b) such as, synthetic aperture radar (SAR) (Ndikumana et al., 2018; Ho Tong Minh et al., 2018), light detection and ranging (LiDAR) (Tao et al., 2015; Li et al., 2016; Hartling et al., 2019), and optical imagery (Surový et al., 2018; Li et al., 2016). Regarding tree complexity, the most common challenges are crown-type differences, shadow complexity, background effects, and spectral heterogeneity, which variate according to vegetation characteristics, planting-method, and landscape conditions (Özcan et al., 2017). This is problematic since no computer vision technique can be universally applied, and different types of approaches must be tested to address specific issues.

In relation to tree delineation literature, studies are generally separated into two groups: tree detection and crown delineation (Özcan et al., 2017). For tree detection, the size of the tree and the spatial resolution of the image are the most important aspects (Larsen et al., 2011; Nevalainen et al., 2017). In tree crown delineation, different approaches such as valley following, watershed segmentation, and region growing are used for boundary extraction (Mathews and Jensen, 2013). In recent years, studies have provided an extensive literature review on these delineation methods, indicating innovative techniques emerging in remote sensing analysis (Ozdarici-Ok, 2015; Özcan et al., 2017).

The traditional techniques used for analyzing images include regression analysis, vegetation indices, linear polarizations, wavelet-based filtering and machine learning (ML) such as support vector machine (SVM), k-means, artificial neural networks (ANN), among others (Ghamisi et al., 2017; Ball et al., 2017; Kamilaris and Prenafeta-Boldú, 2018; Index et al., 2019). As one type of ML techniques, Deep Learning (DL) is recently gaining attention in both environmental and computer vision applications (LeCun et al., 2015; Guo et al., 2016). Similar to ANN; DL uses a deeper neural network with a data hierarchical representation in various convolutions (Ghamisi et al., 2017; Badrinarayanan et al., 2017). This results in a larger leaning capability and improvement in its performance and precision regarding different applications.

A DL algorithm consists of different components that depend on the network architectures (Ball et al., 2017). These components consist of convolutions, fully connected layers, memory cells, gates, pooling layers, activation functions, encode/decode and others (Lecun et al., 2015); while most common architectures are recurrent neural networks (RNN), unsupervised pre-trained networks (UPN), and convolutional neural network (CNN) (Kamilaris and Prenafeta-Boldú, 2018). For image and pattern recognition, CNN has presented better performance overall and is currently been implemented in different remote sensing approaches (Zhang et al., 2016a). A recent review study indicated that CNNs appeared in numerous review papers, representing 42% of the DL techniques used in solving agricultural problems (Kamilaris and Prenafeta-Boldú, 2018).

Although there are different applications of CNN in remote sensing, they can be basically divided into three types: spectral information extraction (Chen et al., 2017; Ghamisi et al., 2017); spatial information extraction (Zhang et al., 2016a) and; spectral-spatial information extraction (Zhang et al., 2017; Li et al., 2017). The latter presents an advantage since both spatial and spectral combined information can significantly improve its accuracy (Paoletti et al, 2018). Likewise, feature learning is one of the main advantages of CNN, but an adequate number of datasets must be available to describe the problem (Dijkstra et al., 2019). For tree detection, the number of features used for test and validation of a method variate accordingly to the characteristics of the study (Safonova et al., 2019; Ampatzidis and Partel, 2019; Dijkstra et al., 2019; Hartling et al., 2019; Csillik et al., 2018; Fan et al., 2018; Özcan et al., 2017).

In citrus tree counting, CNN has been the subject of recent studies (Ampatzidis and Partel, 2019; Csillik et al., 2018). These studies either implemented a simple CNN with a refinement algorithm based on superpixels (Csillik et al., 2018) or a region-based CNN detection algorithm (YOLOv3) (Ampatzidis and Partel, 2019. Both methods utilized an object detection approach, which performs well when canopies have a minimum distance between them (Ampatzidis and Partel, 2019). Other object-based approaches like Faster R-CNN (Ren et al., 2015) and RetinaNet (Lin et al., 2017) can also be used in tree detection. A previous study showed these methods potential to discriminated tree species (Santos et al., 2019). However, until this moment, both methods were not tested in high-density citrus-tree. The object detection approach relies on the bright pixels being recognized as the tree, while dark pixels (shadows) represents the boundary of the tree-canopy (Özcan et al., 2017). In a high-density orchard, this type of approach is expected to be less consistent and even problematic, thus decreasing its

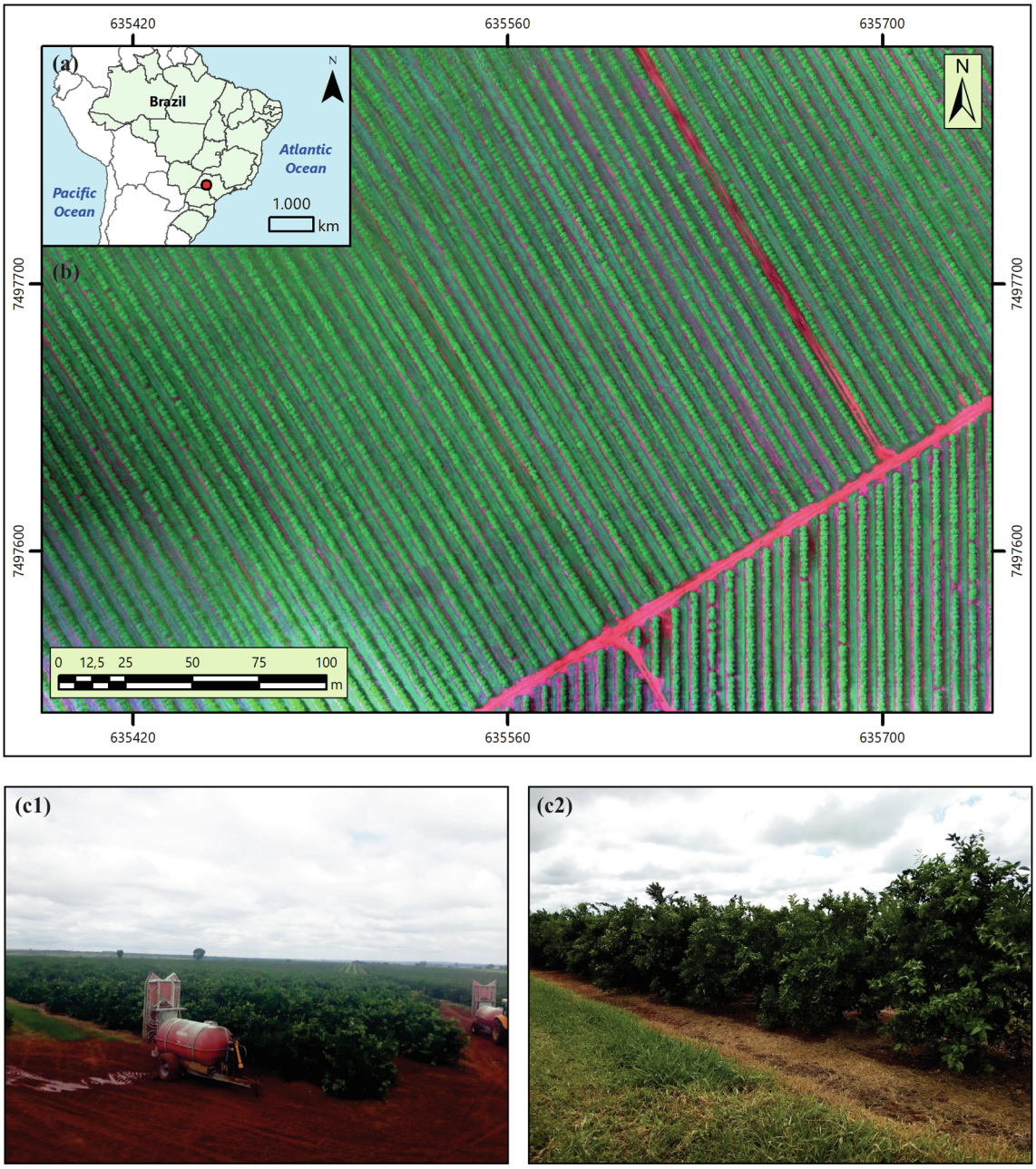

Fig. 1. Characteristics of the study area: (a) location map; (b) band combination displaying a portion of the evaluated area; (c1, c2) examples of planting lines of our studied site.

performance (Ampatzidis and Partel, 2019).

To overcome these issues, this study presents a new CNN approach for counting citrus trees in multispectral images obtained from a UAV-borne sensor. Its details are described in the following section. Unlike previous researches which have estimated a rectangle for each plant, the present method estimates a dense map with the confidence that a plant occurs at each pixel, which is more suitable for situations with high plant density. In the experiments, the use of individual spectral bands or the combination of them was also tested to ascertain which is more suitable for the proposed task. To evaluate the robustness of our method, we compared it against object detection CNN methods like Faster R-CNN and RetinaNet. This analysis conducted here may contribute to optimize the counting of citrus plants while at the same time indicating the importance of evaluating different spectral regions in a high-density orchard.

3. Materials and studied area

Fig. 1 shows our studied area with planting lines of a Valencia-orange tree orchard (Citrumelo Swingle rootstock), located in a property in Ubirajara, SP, Brazil. The area has approximately 70 ha, with Valencia-orange trees planted at a 7 × 1.9 m spacing, with around 752 plants per ha. The UAV flight took place on March 22, 2018, and the trees were in their vegetative state. The trees were approximately 5 years old and about 3 m high, reaching their maturity and production stages.

The images were acquired with a Parrot Sequoia camera (©Parrot-Drones SAS, USA) onboard the eBee SenseFly UAV (©SenseFly, Parrot-Group, USA), which operates in the four spectral bands of green, red, red-edge, and near-infrared (NIR), respectively. A total of 37,353 trees were manually identified in the orthophoto, which was generated using 2,389 images, acquired in the study area. Details describing the

Table 1
Parrot Sequoia camera and eBee SenseFly flight details.

Spectral band	Wavelength	Bandwidth	Spectral resolution	10 bits	Flight high	120 m
Green	550 nm	40	**GSD - Spatial resolution**	12.9 cm	**Flight time**	01:30P.M..
Red	660 nm	40	**HFOV**	70.6°	**Weather**	cloudy/partially-cloudy
Red-edge	735 nm	10	**VFOV**	52.6°	**Precipitation**	0 mm
Near-infrared	790 nm	40	**DFOC**	89.6°	**Wind**	at 1 to 2 m/s

Ground Sample Distance (GSD) Horizontal Field of View (HFOV); Vertical Field of View (VFOV); Displayed Field of View (DFOC).

cameras and flight conditions are presented in Table 1.

The orthorectification was performed with Pix4DMapper software using 9 ground control points (GCPs) surveyed with dual-frequency GNSS (Global Navigation Satellite System) Leica Plus GS15 receiver, in RTK (Real-Time Kinematic) mode. The images were radiometrically corrected using the radiance values of a calibrated reflectance plate, recorded with the camera prior to the flight. An orthorectified surface reflectance image was generated, and the tree locations were generated as point features using the photointerpretation technique.

4. Method

Our approach takes a UAV multispectral image as input and produces the location of each plant. An image has w × h pixels for each of the d bands. The problem of plant counting was modeled as a 2D confidence map estimation problem (Cao et al., 2017). The map is a 2D representation of the confidence that a particular plant occurs in each pixel. The proposed approach uses CNN to estimate the 2D confidence map. We use the ground truth confidence map by placing a 2D Gaussian kernel at each plant location (manually labeled) to train the CNN. Sections 3.1 and 3.2 presents a detailed description of this process.

Given the confidence map, predicted and refined by CNN, the location of each plant is obtained from the peaks (local maximum), as described in Section 4.3. If n plants occur in the image, there should be a peak in the 2D confidence map corresponding to each plant. The steps of the proposed approach are described in the following sections.

4.1. Generation of 2D confidence maps

Given the locations $L = \{l_1, l_2, \ldots, l_n\} \mid l_k \in R^2$ of n plants in an image, the ground truth confidence map C is obtained by placing a 2D Gaussian kernel at each plant location. To obtain C, a confidence map C_k is first calculated for each plant $k \in [0, n]$. The value of each location $p \in R^2$ in C_k is defined by

$$C_k(p) = \exp\left(-\frac{|p-l_k|_2^2}{\sigma^2}\right) \quad (1)$$

where σ is the important parameter controlling the spread of the peak. Ideally, σ is proportional to the size of the tree canopy. The ground truth confidence map C is obtained by aggregating the individual maps via a maximum operator

$$C(p) = \max_k C_k(p) \quad (2)$$

Fig. 2 illustrates the confidence map for two images and three values of σ. The first column shows the images and locations of plants in red dots. The next three columns present the confidence maps for σ = 0.5, 1.0, 1.5, respectively. The ground truth confidence map is used to train a CNN.

4.2. Confidence map estimation

Our approach uses CNN to learn a regression function that receives an image as input and returns a prediction of the confidence map as shown in Fig. 3. The initial part of the CNN (Fig. 3a) is based on the VGG16 (Simonyan and Zisserman, 2015). The first two convolutional layers have 64 filters of size 3 × 3, and they are followed by a 2 × 2 max-pooling layer. The third and fourth convolutional layers have 128 3 × 3 filters, which are also followed by a 2 × 2 max-pooling layer. Finally, the last two convolutional layers have 256 filters of size 3 × 3. All convolutional layers use rectified linear units (ReLU) as the activation function. In this work, the first part receives an image with 256 × 256 pixels with d bands and produces a feature map F with a dimension of 64 × 64 due to the max-pooling layers.

The feature map F generated by the first part of the CNN is given as input to T stages that estimate the confidence map. At the first stage (Fig. 3b), a series of convolutional layers Ω generate the confidence map $\hat{C}^1 = \Omega(F)$. Ω is a sequence of five convolutional layers: three layers with 128 filters of size 3 × 3, one layer with 512 filters of size 1 × 1, and one layer with a single filter that corresponds to the confidence map.

In a subsequent stage t (Fig. 3c), the prediction of the previous stage $\hat{C}^{t22121}$ and the feature map F are concatenated and used to produce a refined confidence map $\hat{C}^t = \Psi(F, \hat{C}^{t-1})$. Ψ is a sequence of seven convolutional layers: five layers with 128 filters of size 7 × 7 and two layers with filters of size 1 × 1. The stages refine the predictions of the confidence map over successive steps, $t \in \{1, \ldots, T\}$.

To train the CNN, the loss function (Equation (3)) is applied at the end of each stage. This intermediate supervision addresses the vanishing gradient problem as shown in (Cao et al., 2017). Since the size of the predicted confidence map is smaller than the image size, the ground truth confidence map is generated with the output size of the CNN, which in this work was 64 × 64 pixels.

$$f^t = \sum_p \left\| C(p) - \widehat{C}^t(p) \right\|_2^2 \quad (3)$$

where C is the ground truth confidence map generated. Finally, the overall loss function is given by

$$f = \sum_{t=1}^{T} f^t \quad (4)$$

4.3. Plant localization from confidence map

The location of the plants is obtained from the peaks (local maximum) of the predicted confidence map of the last stage $\hat{C}^T$, which is the refined confidence map. A location $p = (x_p, y_p)$ is the local maximum if $\hat{C}^T(p) > \hat{C}^T(v)$ for all neighbors v. The 4-connected pixels were considered as neighbors to every location $p = (x_p, y_p)$, i.e., $v = (x_{p \pm 1}, y_p)$ or $(x_p, y_{p \pm 1})$.

To avoid noise, the peaks need to be separated by at least δ pixels. This prevents two plants from being detected very close to each other. Also, a peak must have confidence greater than a threshold τ. After preliminary experiments, $\delta = 3$ and $\tau = 0.2$ were used. Fig. 4 shows an example of the confidence map, where the width and height are the image dimensions and the blue peaks represent the regions with local maximum confidence.

4.4. Experimental setup

The orthorectified surface reflectance image was split into 562 patches of 256 × 256 non-overlapping pixels (with approximately

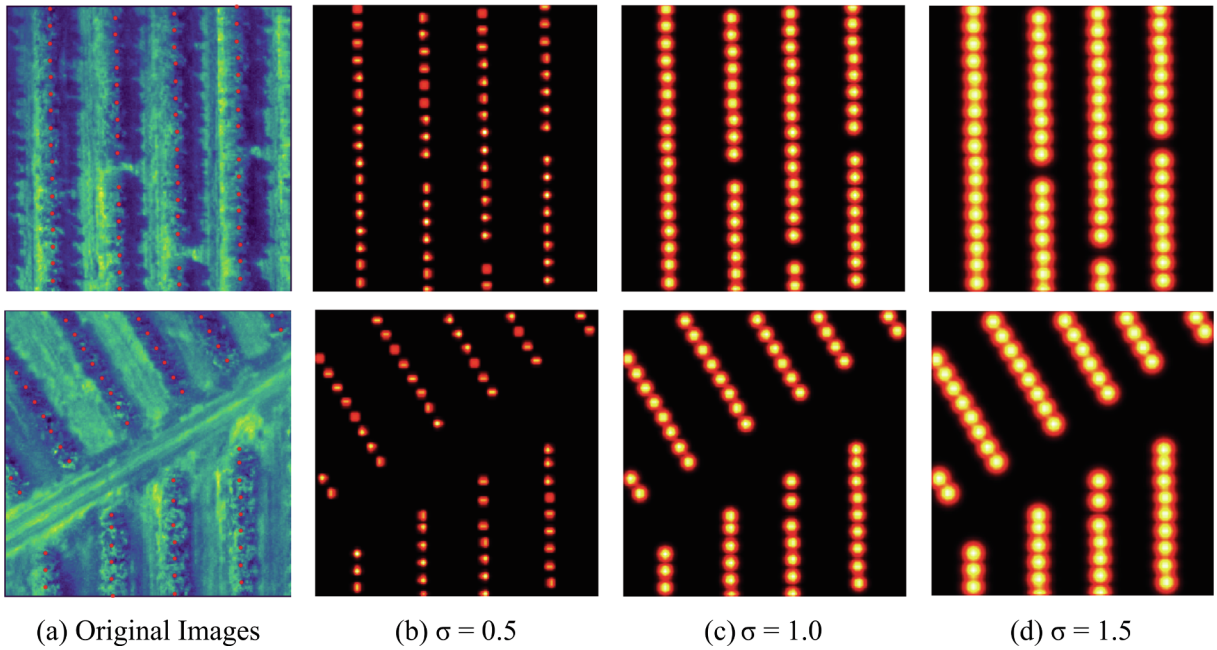

Fig. 2. Example of two images and their corresponding confidence maps for different values of σ.

33 × 33 meters). To evaluate the proposed approach, the patches were randomly divided into training, validation and testing sets made up of 80% (448 patches), 10% (56 patches), and 10% (56 patches), respectively. For training, the stochastic gradient descent (SGD) optimizer was used with a momentum of 0.9. Hyperparameter tuning was performed on the learning rate and the number of epochs, using the validation set to reduce the risk of overfitting. After a minimal hyperparameter tuning, the learning rate was 0.01 and the number of epochs was 300.

Instead of training the proposed approach from scratch, the weights of the first part were initialized with pre-trained weights in ImageNet. Although weights are trained in RGB images from ImageNet, it was found out that the transfer learning assists the training of the proposed approach. When the multispectral image had more than three channels, an additional dimension with random weights in the first layer was included.

In the experiments, regression metrics are reported measuring the agreement between the number of annotated and predicted plants. The metrics were mean absolute error (MAE), mean squared error (MSE), coefficient of determination (R^2), and normalized root-mean-squared error (NRMSE). Given the number of annotated y_j and predicted $\hat{y}_j$ plants for patch j, MAE calculates the average of the absolute errors, $MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$. Similarly, MSE estimates the average of the squares of the errors, $\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$.NRMSE represents the square root of the normalized MSE. This metric facilitates the comparison between methods that work at different scales.

Finally, the coefficient of determination (R^2) estimates the correlation between the number of annotated and predicted plants. To assess the quality of plant detection, we also used classification metrics such as precision, recall, and F1 calculated according to $p = \frac{tp}{tp+fp}$, $r = \frac{tp}{tp+fn}$, and $F1 = 2 \times \frac{p \times r}{p+r}$respectively. We defined a true positive (tp) if the predicted and annotated position of the plant is at less than a maximum distance d. False-positive (fp) and false negative (fn) are calculated similarly using the distance d. In this work, the distance d was defined as the size of the tree canopy (120 cm). We compared our method to two state-of-the-art object detection methods, RetinaNet and Faster R-CNN.

Training and testing were performed using a desktop computer with Intel(R) Xeon(R) CPU E3-1270@3.80 GHz, 64 GB memory, and NVIDIA Titan V graphics card (5120 Compute Unified Device Architecture - CUDA cores and 12 GB graphics memory). The methods were implemented using Keras-Tensorflow on the Ubuntu 18.04 operating system. The computational cost for the different number of stages (T) considering this desktop had already been assessed.

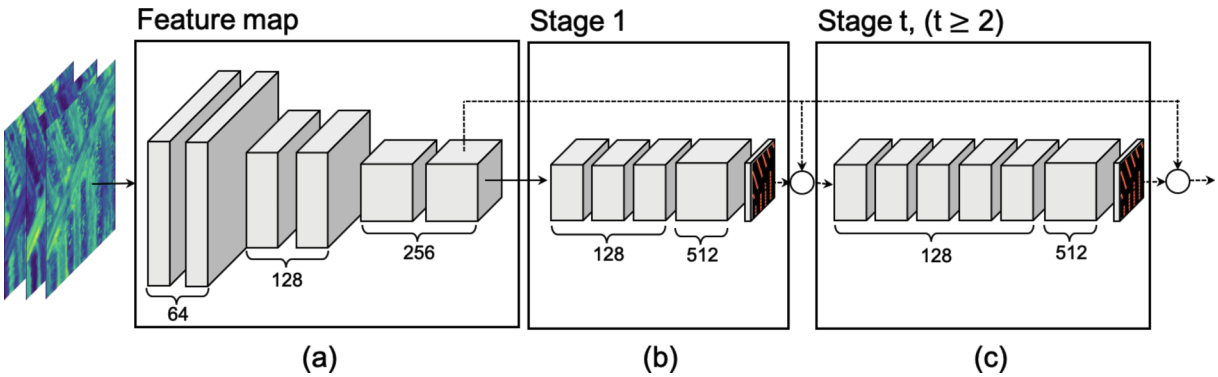

Fig. 3. CNN used for confidence map prediction. It consists of an initial part (a) to extract a feature map of the input image. This feature map is used as input to the first stage (b). The concatenation of the feature map and the prediction map of the previous stage is used as input for the remaining stages.

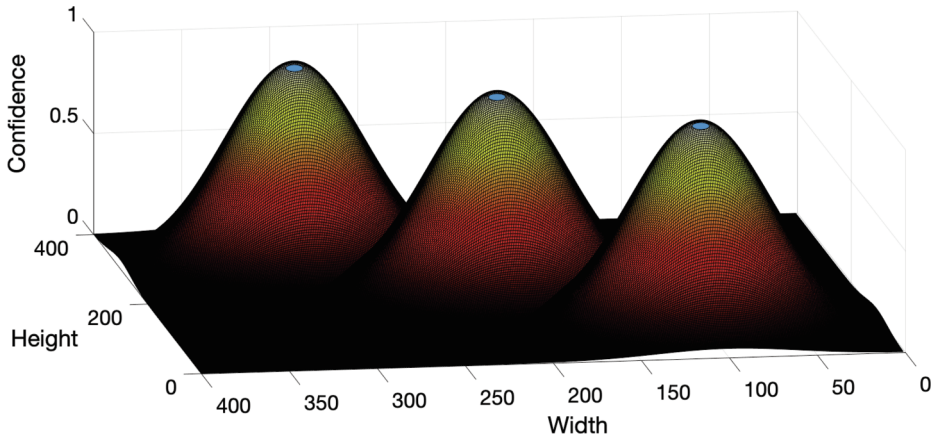

Fig. 4. Example of the confidence map in three dimensions.

5. Results and discussion

5.1. Analysis of the proposed method parameters

Table 2 presents the results for different bands and combinations among them. The objective is to evaluate which bands are most appropriate for plant counting using the proposed approach. These results were obtained using T = 6 stages and σ = 1.0. Even considering only one spectral band (e.g., green), the proposed approach already presents satisfactory results, with MAE, MSE, R^2, and NRMSE of 2.51, 10.72, 0.96, and 0.039 respectively. However, a performance increase was obtained when combining the green, red and NIR bands, giving an NRMSE of 0.038.

It can also be seen that using the Red-edge band did not imply good results compared to the other bands. We observed that the Red-edge band does not have sufficient contrast regarding other targets. Red-edge parameters such as curve slope and reflectance can be used to differentiate illuminated from shaded canopies (Index et al., 2019), and its usage is commonly known in remote sensing applications. However, the evaluated region (735 ± 10 nm) in this study presented a high similarity between other vegetation targets.

The spectral response from the citrus plants in comparison to other types of land cover (bare soil, shallow grassland, and dense grassland) in the study area is displayed in Fig. 5. By collecting different samples (one hundred for each land cover type), it could be seen that the orange-trees and dense grassland presented similar surface reflectance at the Red-edge region. This may be indicative of the reduced CNN performance for this band. In general, similar studies implemented common RGB cameras in their analysis (Weinstein et al., 2019; Csillik et al., 2018; Fan et al., 2018; Varela et al., 2018; Ampatzidis and Partel, 2019), so this type of problem was not perceptible. But the CNN struggle in the Red-edge band in this study case is an important finding since it directs towards adversity in using this band for the proposed task.

Table 2
Results obtained with different bands and combinations.

Bands	MAE	MSE	R^2	NRMSE
Green	2.51	10.72	0.96	0.039
Red	2.74	13.08	0.95	0.046
Red-edge	3.65	40.56	0.85	0.077
NIR	2.98	18.11	0.93	0.052
Green, Red	2.40	13.32	0.95	0.046
Green, Red-edge	2.67	15.68	0.94	0.050
Green, NIR	2.93	16.09	0.94	0.057
Red, Red-edge	2.37	15.18	0.94	0.050
Red, NIR	2.82	17.35	0.93	0.052
Red-edge, NIR	2.96	17.74	0.93	0.052
Green, Red, Red-edge	2.65	15.67	0.94	0.050
Green, Red, NIR	**2.28**	**9.82**	**0.96**	**0.038**
Green, Red-edge, NIR	2.89	20.09	0.92	0.057
Red, Red-edge, NIR	2.68	14.44	0.95	0.047
Green, Red, Red-edge, NIR	2.56	13.47	0.95	0.046

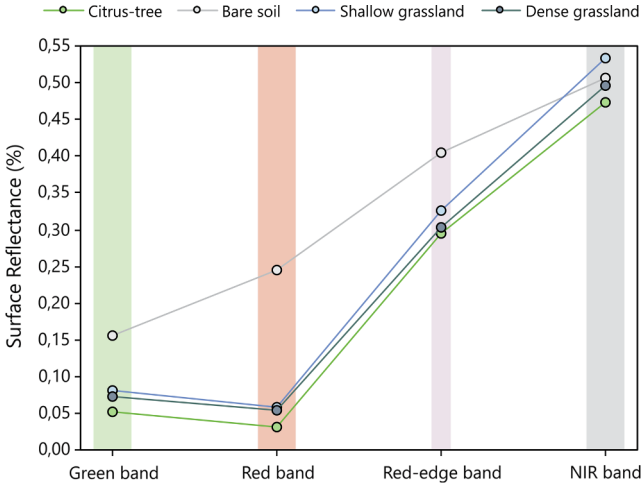

Fig. 5. Spectral behavior of different types of land cover commonly present in the study area.

When increasing to two bands, the use of the Green and Red bands obtained the best result, although it did not surpass the results obtained by the Green band alone. On the other hand, using the Green, Red and NIR bands obtained the best result. These bands achieved MAE, MSE, R^2, and NRMSE of 2.28, 9.82, 0.96 and 0.05, respectively. Considering the four bands as input images, the results were satisfactory although it did not surpass the best result because of the inclusion of the Red-edge band, which does not help in counting the plants. σ, which is responsible for generating the ground truth confidence maps used in the training of the proposed approach, was also evaluated. In these experiments, the green, red and NIR bands that achieved the best results among all bands in the previous experiment were used. σ has a great influence on the results, as can be seen from Tables 3 and 4. For small σ in relation to the tree canopy, results were low as the confidence map does not cover the plants properly (see Fig. 2b). Instead, σ = 1.5 (large values) generates ground truth confidence maps whose peaks are close and can be confused. The best result was obtained for σ = 1.0, which, in this case, is better fitted to the size of the tree canopy.

Finally, the number of stages that refine the confidence map predicted by the proposed approach was evaluated. As expected, the results improve as the number of stages is increased. This shows that the refinement of the confidence map helps in counting the plants. The

Table 3
Evaluation of the σ responsible for generating ground truth Confidence maps to train the proposed approach.

σ	MAE	MSE	R^2	NRMSE
0.5	5.11	58.86	0.87	0.098
1.0	**2.28**	**9.82**	**0.96**	**0.038**
1.5	3.56	25.63	0.90	0.064

Table 4
Evaluation of the number of stages t used to refine the Confidence map predicted by the proposed approach.

Stages (T)	MAE	MSE	R^2	NRMSE
1	3.61	21.05	0.92	0.057
2	2.86	17.39	0.93	0.052
4	2.56	14.42	0.95	0.047
6	2.28	9.82	0.96	0.038
8	**2.05**	**8.75**	**0.97**	**0.036**
10	2.21	11.79	0.96	0.043

proposed approach achieved its best result with eight stages (T = 8).

The results show that the proposed approach provided accurate results for counting plants and can be used to automate this task. This performance approximates from the accuracy obtained in lesser difficult conditions, such as a high-spaced citrus plantation (Ampatzidis and Partel, 2019; Csillik et al., 2018). A visually similar density condition was evaluated in a different crop type (Fan et al., 2018), which achieve 93% accuracy on tobacco plant detection using CNN. One advantage of this approach is that it was conducted multispectral imagery, while previous studies used LIDAR and hyperspectral data (Wu et al., 2016; Wu and Prasad, 2018). Regardless, it's possible that the presented approach accuracy could be improved if three-high is inserted in a CNN layer.

5.2. Qualitative results

To analyze the results qualitatively, a region around the annotated locations was considered to visualize the proximity of the prediction and the center of the plants. Fig. 6 shows the results using the best configuration (three bands, σ = 1.0, and T = 8). The predicted locations are represented by red dots in this figure and the plant regions are represented by yellow circles whose center is the location annotated by the specialist. It can be seen that the proposed approach can correctly predict most plant locations, with a 2.05 trees error per image, so that they are aligned with the annotated locations and within the plant region.

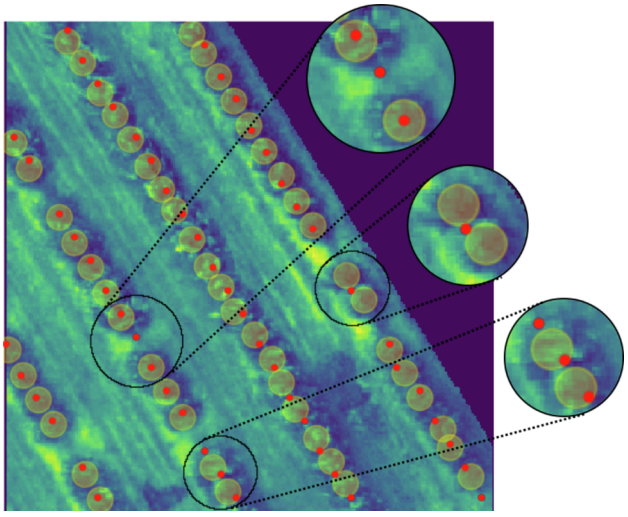

Fig. 7. Examples of prediction errors in our approach.

The results show that planting lines are also identified without the need for any annotation or additional procedure. Identifying planting lines is also an important feature in remote sensing of agricultural fields since it can easily detect missing trees and help optimize crop management (Dian Bah et al., 2018; Oliveira et al., 2018). Remote sensing approaches were already conducted in canola fields (Hassanein et al., 2019), tomato crops (Ramesh et al., 2016), vineyards (Puletti et al., 2014) and others, but none has been found for citrus trees orchards. Nonetheless, some difficulties were observed considering the characteristics of the area investigated here. Fig. 7 shows examples of the main challenges faced by our approach.

It can be seen that far-center predictions occur in short planting lines (2–4 plants) or when much of the plant canopy is occluded. However, even in images where these cases occur, the proposed approach is capable of predicting the location of the vast majority of plants. Even with a fixed plantation line size, the area had some missing trees that were previously removed due to health conditions. But, different plantation lines with spaced tree locations were identified by the

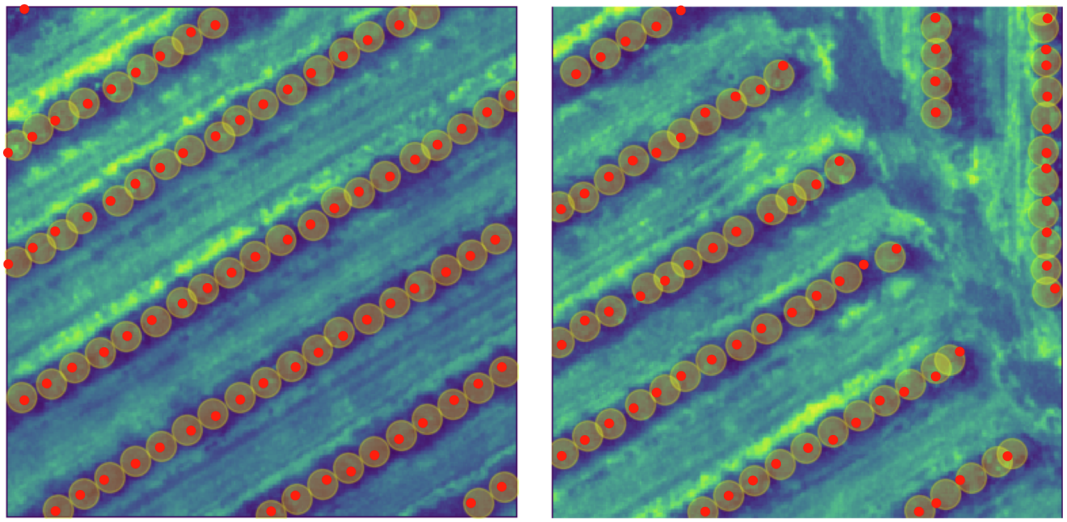

Fig. 6. Comparison of predicted locations (red dots) and plant regions in two images. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

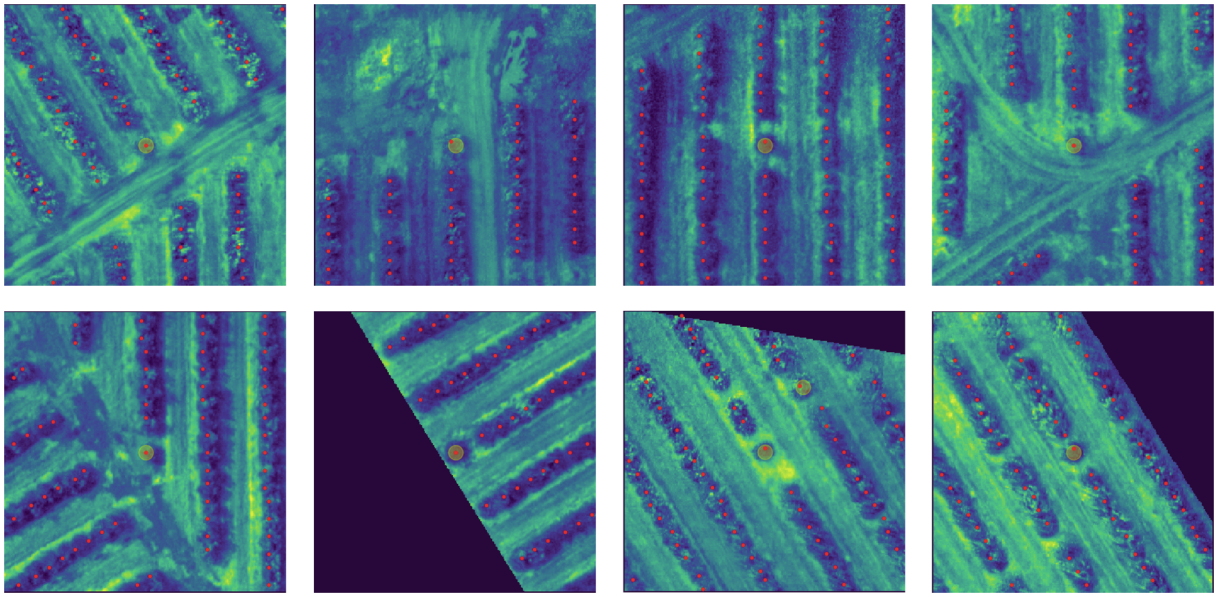

Fig. 8. Examples of spaced trees correctly identified with the proposed approach.

CNN method without difficulty (Fig. 8). This indicates that our approach is also suitable for estimating isolated trees with different plant spacing.

5.3. Comparison with object detection methods

The proposed approach was compared with recent object detection methods such as Faster R-CNN and RetinaNet. To train the object detection methods, we used the plant position (x, y) as the center of the rectangle. The size of the rectangle corresponds to the size of the plant canopy (240 cm). We considered Green, Red, and NIR bands for this comparison. Similarly, an inverse process was used during the testing stage, obtaining the plant position from the center point of the rectangle predicted by the RetinaNet and Faster R-CNN methods.

Table 5 shows the results obtained by all methods using MAE, Precision, Recall, and F1 metrics. We can see that the proposed approach achieved better results for all metrics with 0.95 and 0.96 for Precision and Recall, respectively. In addition, the proposed approach achieved an MAE of 2.05 while RetinaNet and Faster R-CNN provided values of 30.87 and 37.85, respectively. RetinaNet and Faster R-CNN achieved only 0.74 and 0.54 for the F1 score, against 0.95 of the proposed approach. These results indicate that the proposed approach can predict citrus trees with high precision, having a very low number of false detections. The results of object detection methods are consistent with recent works for other high-density object detection applications. Goldman et al. (2019) and Hsieh et al. (2017) showed that these methods do not present satisfactory results when the rectangles have a high intersection, that is, a high density of objects.

Fig. 9 shows the visual results of the predictions generated by the three methods in two images. We can see that our approach has few errors in detecting plants. Faster R-CNN is the most misleading method, failing to identify plants in the images, while RetinaNet predicts more plants than those in the image, generating many false predictions. Note that the Precision reflects this behavior, being lower for RetinaNet than for Faster R-CNN since the number of false positives is higher for RetinaNet.

Table 5
Comparison of the proposed approach with recent object detection methods.

Methods	MAE	Precision	Recall	F1
RetinaNet	30.87	0.62	0.92	0.74
Faster R-CNN	37.85	0.86	0.39	0.54
Proposed approach	**2.05**	**0.95**	**0.96**	**0.95**

5.4. Computational cost

Table 6 presents the computational cost of the proposed approach using different values for the number of stages, which is the main parameter influencing the size of the CNN. This table presents the average time in seconds to process an image with 256 × 256 pixels and three bands, in addition to the estimated number of images per second (FPS) that the proposed approach is capable of processing. The number of bands does not change the cost significantly since the only change is one more dimension in the first layer.

Still, one observation that must be noted is that, by increasing the number of stages, the computational cost also increases. Using one stage, the proposed approach is capable of processing approximately 258 images per second, with a cost of 0.0039 s per image. Considering the best result that was obtained with eight stages (Table 4), the proposed approach is able to process approximately 25 images per second. The speed/accuracy trade-off can be considered in the choice of the number of stages. If an application needs to run in real-time with more than 30 images per second, then four or six stages is a good alternative.

6. Concluding remarks

This paper presented a CNN approach to estimate the number and location of citrus trees from UAV multispectral imagery. Our results archived 0.97 in R^2 and 0.036 trees in NRMSE. The combination of the spectral bands green, red and near-infrared produced better performance than the use of individual spectral bands. Our method also demonstrated reasonable computational cost for embedded real-time applications. One of the advantages of our approach is in estimating a dense map to detect individual trees in high-density plantations, rather than the object-detection approach using rectangles to represent trees. The comparison against object-based methods returned a higher Precision (0.95) and lower MAE (2.05) for our method. This approach is recommended for the counting of citrus trees and we hoped that new studies for different crops will utilize and evaluate it.

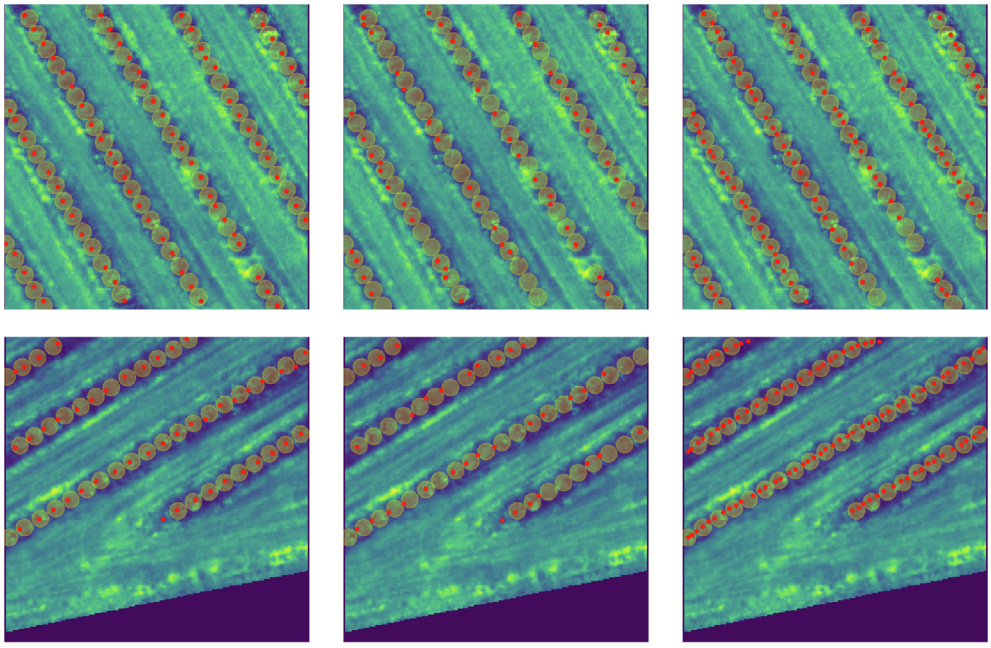

Fig. 9. Examples of the predictions generated by the three methods: (a) Our approach, (b) Faster R-CNN and (c) RetinaNet.

Table 6
Evaluation of the computational cost of the proposed approach for the different number of stages.

Stages (T)	Times (s)	FPS
1	0.0039 (± 0.0002)	258.26
2	0.0092 (± 0.0006)	108.30
4	0.0215 (± 0.0008)	46.49
6	0.0330 (± 0.0010)	30.31
8	**0.0401 (± 0.0012)**	**24.93**
10	0.0498 (± 0.0015)	20.10

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by CNPq (p: 433783/2018-4 and 304173/2016-9), CAPES Print (p: 88881.311850/2018-01) and Fundect (p: 59/300.066/2015). The authors acknowledge the donation of a Titan X and a Titan V by NVIDIA.

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.
Ampatzidis, Y., Partel, V., 2019. UAV-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence. Remote Sensing 11 (4), 410.
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. A comprehensive survey of deep learning in remote sensing: Theories, tools and challenges for the community. J. Appl. Remote Sens. 11 (4), 1–64.
Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., 2017. Realtime multi-person 2D pose estimation using part affinity fields. CVPR 2017, 1302–1310.
Chen, S.W., Shivakumar, S.S., Dcunha, S., Das, J., Okon, E., Qu, C., Kumar, V., 2017. Counting apples and oranges with deep learning: a data-driven approach. IEEE Rob. 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.
Deng, L., Mao, Z., Li, X., Hu, Z., Duan, F., Yan, Y., 2018. UAV-based multispectral remote sensing for precision agriculture: A comparison between different cameras. ISPRS J. Photogramm. Remote Sens. 146, 124–136.
Dian Bah, M., Hafiane, A., Canals, R., 2018. Deep learning with unsupervised data labeling for weed detection in line crops in UAV images. Remote Sensing 10 (11), 1690.
Dijkstra, K., van de Loosdrecht, J., Schomaker, L.R.B., Wiering, M.A., 2019. Centroidnet: a deep neural network for joint object localization and counting. In: Brefeld, U. (Ed.), Machine Learning and Knowledge Discovery in Databases, pp. 585–601.
Djerriri, K., Ghabi, M., Karoui, M.S., Adjoudj, R., 2018. Palm trees counting in remote sensing imagery using regression convolutional neural network. IGARSS 2018, 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 Obs. Remote Sens. 11 (3), 876–887.
Ghamisi, P., Plaza, J., Chen, Y., Li, J., Plaza, A.J., 2017. Advanced spectral classifiers for hyperspectral images: A review. IEEE Geosci. Remote Sens. Mag. 5 (1), 8–32.
Goldbergs, G., Maier, S.W., Levick, S.R., Edwards, A., 2018. Efficiency of individual tree detection approaches based on light-weight and low-cost UAS imagery in Australian Savannas. Remote Sensing 10 (2), 161.
Goldman, E., Herzig, R., Eisenschtat, A., Goldberger Hassner, J.T., 2019. Precise Detection in Densely Packed Scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5227–5236.
Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S., 2016. Deep learning for visual understanding: a review. Neurocomputing 187, 27–48.
Hartling, S., Sagan, V., Sidike, P., Maimaitijiang, M., Carron, J., 2019. Urban tree species classification using a worldview-2/3 and LiDAR data fusion approach and deep learning. Sensors 19 (6), 1–23.
Hasan, M.M., Chopin, J.P., Laga, H., Miklavcic, S.J., 2018. Detection and analysis of wheat spikes using convolutional neural networks. Plant Methods 14 (1), 1–13.
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.
Ho Tong Minh, D., Ienco, D., Gaetano, R., Lalande, N., Ndikumana, E., Osman, F., Maurel, P., 2018. Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR Sentinel-1. IEEE Geosci. Remote Sens. Lett. 15 (3), 465–468.
Hsieh, M.R., Lin, Y.L., Hsu, W.H., 2017. Drone-based object counting by spatially regularized regional proposal network. In: In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4145–4153.
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.
Index, S., Xu, N., Tian, J., Tian, Q., Xu, K., Tang, S., 2019. Analysis of Vegetation red edge with different illuminated/shaded canopy proportions and to construct normalized difference canopy. Remote Sensing 11 (10), 1192 1–16.
Jakubowski, M.K., Li, W., Guo, Q., Kelly, M., 2013. Delineating individual trees from lidar data: a comparison of vector- and raster-based segmentation approaches. Remote Sensing 5 (9), 4163–4186.

Jiang, H., Chen, S., Li, D., Wang, C., Yang, J., 2017. Papaya tree detection with UAV images using a GPU-accelerated scale-space filtering method. Remote Sensing 9 (7), 721.
Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: a survey. Comput. Electron. Agric. 147, 70–90.
Kang, D., Ma, Z., Chan, A.B., 2019. Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. IEEE Trans. Circuits Syst. Video Technol. 29 (5), 1408–1422.
Larsen, M., Eriksson, M., Descombes, X., Perrin, G., Brandtberg, T., Gougeon, F.A., 2011. Comparison of six individual tree crown detection algorithms evaluated under varying forest conditions. Int. J. Remote Sens. 32 (20), 5827–5852.
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, D., Guo, H., Wang, C., Li, W., Chen, H., Zuo, Z., 2016. Individual tree delineation in windbreaks using airborne-laser-scanning data and unmanned aerial vehicle stereo images. IEEE Geosci. Remote Sens. Lett. 13 (9), 1330–1334.
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 Sensing 9 (1), 22.
Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.
Liu, T., Abd-Elrahman, A., Morton, J., Wilhelm, V.L., 2018. Comparing fully convolutional networks, random forest, support vector machine, and patch-based deep convolutional neural networks for object-based wetland mapping using images from small unmanned aircraft system. GI Sci. Remote Sens. 55 (2), 243–264.
Liu, T., Abd-Elrahman, A., 2018. Deep convolutional neural network training enrichment using multi-view object-based analysis of Unmanned Aerial systems imagery for wetlands classification. ISPRS J. Photogramm. Remote Sens. 139, 154–170.
Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177.
Madec, S., Jin, X., Lu, H., De Solan, B., Liu, S., Duyme, F., Baret, F., 2019. Ear density estimation from high resolution RGB imagery using deep learning technique. Agric. For. Meteorol. 264, 225–234.
Mathews, A.J., Jensen, J.L.R., 2013. Visualizing and quantifying vineyard canopy LAI using an unmanned aerial vehicle (UAV) collected high density structure from motion point cloud. Remote Sensing 5 (5), 2164–2183.
Ndikumana, E., Minh, D.H.T., Baghdadi, N., Courault, D., Hossard, L., 2018. Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue. France. Remote Sensing 10 (8), 1217.
Nevalainen, O., Honkavaara, E., Tuominen, S., Viljanen, N., Hakala, T., Yu, X., Tommaselli, A.M.G., 2017. Individual tree detection and classification with UAV-Based photogrammetric point clouds and hyperspectral imaging. Remote Sensing 9 (3), 185.
Oliveira, H.C., Guizilini, V.C., Nunes, I.P., Souza, J.R., 2018. Failure detection in row crops from UAV Images using morphological operators. IEEE Geosci. Remote Sens. Lett. 15 (7), 991–995.
Özcan, A.H., Hisar, D., Sayar, Y., Ünsalan, C., 2017. Tree crown detection and delineation in satellite images using probabilistic voting. Remote Sens. Lett. 8 (8), 761–770.
Ozdarici-Ok, A., 2015. Automatic detection and delineation of citrus trees from VHR satellite imagery. Int. J. Remote Sens. 36 (17), 4275–4296.
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.
Puletti, N., Perria, R., Storchi, P., 2014. Unsupervised classification of very high remotely sensed images for grapevine rows detection. Eur. J. Remote Sens. 47 (1), 45–54.
Ramesh, K.N., Chandrika, N., Omkar, S.N., Meenavathi, M.B., Rekha, V., 2016. Detection of Rows in Agricultural Crop Images Acquired by Remote Sensing from a UAV. Int. J. Image, Graph. Signal Proce. 8 (11), 25–31.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Proce. Syst. 91–99.
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 Sensing 11 (6), 643.
Salamí, E., Gallardo, A., Skorobogatov, G., Barrado, C., 2019. On-the-fly olive tree counting using a UAS and cloud services. Remote Sensing 11 (3), 316.
Santos, A., Marcato Junior, J., Araujo, M.S., Martini, D.R., Tetila, E.C., Siqueira, H.L., Aoki, C., Eltner, A., Matsubara, E.T., Pistori, H., Feitosa, R., Liesenberg, V., Gonçalves, W.N., 2019. Assessment of CNN-based methods for individual tree detection on images captured by RGB cameras attached to UAVs. Sensors 19 (16), 3595.
Simonyan, K., Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. 1–14. Retrieved from http://arxiv.org/abs/1409.1556.
Surový, P., Almeida Ribeiro, N., Panagiotidis, D., 2018. Estimation of positions and heights from UAV-sensed imagery in tree plantations in agrosilvopastoral systems. Int. J. Remote Sens. 39 (14), 4786–4800.
Tao, S., Wu, F., Guo, Q., Wang, Y., Li, W., Xue, B., Fang, J., 2015. Segmenting tree crowns from terrestrial and mobile LiDAR data by exploring ecological theories. ISPRS J. Photogramm. Remote Sens. 110, 66–76.
Varela, S., Dhodda, P.R., Hsu, W.H., Prasad, P.V.V., Assefa, Y., Peralta, N.R., Griffin, T., Sharda, A., Ferguson, A., Ciampitti, I.A., 2018. Early-season stand count determination in corn via integration of imagery from unmanned aerial systems (UAS) and supervised learning techniques. Remote Sensing 10 (2), 343.
Verma, N.K., Lamb, D.W., Reid, N., Wilson, B., 2016. Comparison of canopy volume measurements of scattered eucalypt farm trees derived from high spatial resolution imagery and LiDAR. Remote Sensing 8 (5), 388.
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 Sensing 11 (11), 1309.
Wu, B., Yu, B., Wu, Q., Huang, Y., Chen, Z., Wu, J., 2016. Individual tree crown delineation using localized contour tree method and airborne LiDAR data in coniferous forests. Int. J. Appl. Earth Observ. Geoinf. 52, 82–94.
Wu, H., Prasad, S., 2018. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. IEEE Trans. Image Process. 27 (3), 1259–1270.
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 Sensing 11 (6), 691.
Zhang, H., Li, Y., Zhang, Y., Shen, Q., 2017. Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. Remote Sensing Letters 8 (5), 438–447.
Zhang, L., Zhang, L., Kumar, V., 2016a. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4 (2), 22–40.
Zhang, P., Gong, M., Su, L., Liu, J., Li, Z., 2016b. Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 116, 24–41.