Contents lists available at ScienceDirect

Expert Systems With Applications

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

Counting and locating high-density objects using convolutional neural network

Mauro dos Santos de Arruda [a], Lucas Prado Osco [b,e], Plabiany Rodrigo Acosta [a], Diogo Nunes Gonçalves [a], José Marcato Junior [b], Ana Paula Marques Ramos [c,d], Edson Takashi Matsubara [a], Zhipeng Luo [f], Jonathan Li [g], Jonathan de Andrade Silva [a], Wesley Nunes Gonçalves [a,b,*]

[a] Faculty of Computer Science, Federal University of Mato Grosso do Sul, Av. Costa e Silva, Campo, Grande 79070-900, Brazil
[b] Faculty of Engineering, Architecture, and Urbanism and Geography, Federal University of Mato Grosso do Sul, Av. Costa e Silva, Campo, Grande 79070-900, Brazil
[c] Environment and Regional Development Program, University of Western São Paulo, Rodovia Raposo Tavares, 572 km Limoeiro, Presidente, Prudente 19067-175, Brazil
[d] Agronomy Program, University of Western São Paulo, Rodovia Raposo Tavares, 572 km Limoeiro, Presidente, Prudente 19067-175, Brazil
[e] Faculty of Engineering and Architecture and Urbanism, University of Western São Paulo, R. José Bongiovani, 700-Cidade Universitária, Presidente, Prudente 19050-920, Brazil
[f] Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen FJ 361005, China
[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
Object counting
Tree counting
Car counting

ABSTRACT

This paper presents a Convolutional Neural Network (CNN) approach for counting and locating objects in high-density imagery. To the best of our knowledge, this is the first object counting and locating method based on a feature map enhancement combined with a multi-sigma refinement of the confidence map. The proposed method was evaluated in two counting datasets: trees and cars. For the tree dataset, our method returned a mean absolute error (MAE) of 2.05, a root-mean-squared error (RMSE) of 2.87 and a coefficient of determination (R^2) of 0.986. For the car dataset (CARPK and PUCPR+), our method was superior to state-of-the-art methods. In the these datasets, our approach achieved an MAE of 4.45 and 3.16, an RMSE of 6.18 and 4.39, and an R^2 of 0.975 and 0.999, respectively. We conclude that the proposed method is suitable for dealing with high object-density, returning a state-of-the-art performance for counting and locating objects.

1. Introduction

Neural networks have been widely used in different applications, including ground source heat pump performance prediction (Esen et al., 2017, 2008a, 2008b, 2008c, 2009), plastic waste (Kokoulin et al., 2018), classifying cardiac arrhythmias (Castillo et al., 2012) and monitoring wildlife (d. S. de Arruda et al., 2018). In particular, the task of counting and locating objects in images have been the attention of several approaches (Sindagi & Patel, 2018). These methods help to control and count people (Idrees et al., 2018), support car detection (Hsieh et al., 2017), and even count bacterial colonies (Ferrari et al., 2017). As expected, the majority of these methods are based on the well-known object detection task, including the recent methods based on convolutional neural networks (Faster R-CNN (Ren et al., 2017), Mask-RCNN (He et al., 2020), RetinaNet (Lin et al., 2020)), multi-scale variants (Multi-Scale Structures (Ohn-Bar & Trivedi, 2017), Multi-scale deep feature learning network (Ma et al., 2020), Gated CNN (Yuan et al., 2019)) and ensembles of models (Xu et al., 2020). Many object detection methods consider a bounding box (bbox) around the targeted objects and can provide both location (center of the bbox) and counting (number of bboxs). Recent contribution to this matter is from Hsieh et al. (2017), where authors proposed, simultaneously,

[*] Corresponding author at: Faculty of Computer Science, Federal University of Mato Grosso do Sul, Av. Costa e Silva, Campo, Grande 79070-900, Brazil.
E-mail addresses: mauro.arruda@ufms.br (M.d.S. de Arruda), lucasosco@unoeste.br (L.P. Osco), plabiany.acosta@ufms.br (P.R. Acosta), diogo.goncalves@ufms.br (D.N. Gonçalves), jose.marcato@ufms.br (J. Marcato Junior), anaramos@unoeste.br (A.P.M. Ramos), edsontm@facom.ufms.br (E.T. Matsubara), zpluo@stu.xmu.edu.cn (Z. Luo), junli@uwaterloo.ca (J. Li), jonathan.silva@ufms.br (J.d.A. Silva), wesley.goncalves@ufms.br (W.N. Gonçalves).

https://doi.org/10.1016/j.eswa.2022.116555
Received 27 May 2020; Received in revised form 11 May 2021; Accepted 14 January 2022
Available online 5 February 2022

Layout Proposal Networks (LPNs) and spatial kernels to detect objects in videos. These additions helped to improve the object counting and location using an object detection framework. Still, even state-of-the-art methods return bounding boxes that partially overlap multiple objects, which is still a problem since the adjacent object region is detected as a separate object (Goldman et al., 2019).

One of the biggest challenges regarding counting and location of objects in images is the high-object density. Object detection methods are, in general, not adequate for high-density scenes (Goldman et al., 2019). In this scenario, overlapping objects are difficult to analyze due to the size of the instances and the standpoint of the scene. Thus, approaches that model the problem of counting objects with a density estimation has been defined as state-of-the-art solutions, and are providing interesting solutions for dense scenes (Aich & Stavness, 2018; Goldman et al., 2019). In Goldman et al. (2019), the authors proposed a CNN-based detection method, using the bounding box, to cope with densely packed scenes. They considered a layer to estimate a quality score index and used a novel EM (Expectation–Maximization) merging unit to solve the overlap ambiguities with this score. However, handling high-density objects in images is still a concerning issue, both in counting and locating objects.

Another problem regarding object count from detection frameworks is the need of detailed ground-truth labeled data, which is hard to obtain at large-scales (Russakovsky et al., 2015). Acquiring a large-scale annotated data is a time-consuming process. Because of that, approaches based on a lighter weight image label is something that researchers have previously proposed Fiaschi et al. (2012), Zhang et al. (2018). Still, recent studies are implementing point annotations to reduce the supervision task (Aich & Stavness, 2018; Liu et al., 2019). Point annotations are easier to obtain than bounding-boxes, and many counting and locating approaches do not need to rely on them to identify an object (Liu et al., 2019). These types of approaches can rely on context information, and, for most problems, object instances will share a similar color, texture, and shape; meaning that the method will learn how to recognize them even if only using point features (Aich & Stavness, 2018).

Recently, state-of-the-art methods to count objects include the VGG-GAP and VGG-GAP-HR (Aich & Stavness, 2018) approaches, Layout Proposal Networks (LPN) (Hsieh et al., 2017) and Deep IoU CNN (Goldman et al., 2019). These methods were applied in counting and locating cars, crowds, biological cells and products from supermarket shelves, returning impressive performances in high-density scenes. Despite the promising results, scale variations, clutter background, occlusions, and especially high-density of objects are still challenges that hinder methods of providing high-quality predictions. That way, in previous work, we developed an initial model for the location and counting of Citrus-trees in UAV multispectral images (Osco et al., 2020). This initial model significantly surpassed methods for detecting objects such as RetinaNet and Faster-RCNN.

In this paper, we present a method for counting and locating objects based on convolutional neural networks (Simonyan & Zisserman, 2015). The method is based on a density estimation map with the confidence that an object occurs in each pixel, following Aich and Stavness (2018). Unlike previous works that estimate a bounding box for each object, the estimation of a density map allows a better refinement of the occurrence of objects in each pixel of the image. Different from previous work Osco et al. (2020), the proposed method uses a feature map enhancement with a Pyramid Pooling Module (PPM) (Zhao et al., 2017) that allows to incorporate global information at different scales. Consequently, the proposed method incorporates sufficient global context information for a good characterization of objects similarly to Zhang et al. (2019) with its hierarchical context module. Thus, in this paper, we hypothesize that this approach is most suitable for situations of high object density, since it incorporates detection information in each pixel with the density map and improves this learning with regional information provided by the PPM module.

Another potential pitfall of previous methods is the missed detections due to object occlusion and high-density scenes. To compensate for these problems, and produce the correct predictions, we also propose a multi-sigma refinement over the ground-truth to provide hierarchical learning of the object positions. The multi-sigma refinement phase starts from a rough prediction of the object position to a more refined prediction of the center of the object. Our hypothesis is that this refinement allows the method to provide more assertive predictions, decreasing the number of missed detections caused by occlusion and high-density scenes. To incorporate these improvements, we divided the proposed method into four main phases: (1) a feature map generation using a Convolutional Neural Network (CNN); (2) a global context insertion in the feature map using a Pyramid Pooling Module (PPM); (3) a Multi-Sigma Refinement of the confidence maps; and (4) object position estimation through peaks in the confidence map.

To verify the performance of the proposed approach, we performed experiments in three image datasets in two challenging applications. First, we perform a parameter evaluation in a tree counting dataset containing 3,370 images and approximately 232,000 objects. This dataset presents trees with irregular distribution and different growth stages, different from our previous research (Osco et al., 2020). Once the best parameters were defined, we evaluated the generalization of the method in two car-counting benchmarks: CARPK and PUCPR+. For that, we evaluated the proposed method with 13 other state-of-the-art object detection methods.

2. Proposed method

This section describes our method to count and locate objects. This method uses a three-channel image, with $w \times h$ pixels, as input, and processes it with a CNN. The object counting and location is modeled after a 2D confidence map estimation, following the procedures presented in Aich and Stavness (2018).

The confidence map is a 2D representation of the likelihood of an object occurs in each pixel. We improved the confidence map estimation by including global and local information through a Pyramid Pooling Module (PPM) (Zhao et al., 2017). We also proposed a multi-sigma prediction phase to refine the confidence map to a more accurate prediction of the center of the objects.

Fig. 1 illustrates the phases of the proposed method, which are detailed in the following section. Our approach is divided into four main phases: (1) feature map generation with a CNN (Section 2.1); (2) feature map enhancement with the PPM (Section 2.2); (3) multi-sigma refinement of the confidence map (Sections 2.3 and 2.4); and, (4) object position obtention by peaks in the confidence map (Section 2.5).

2.1. Feature map using CNN

The first part of the proposed approach uses a convolutional neural network to extract a feature map from a given input image (Fig. 1(a)). The feature map is used to characterize the input image and allow the confidence map estimation for the object detection task. This feature map extraction module is based on the VGG19 (Simonyan & Zisserman, 2015), where the first two convolutional layers have 64 filters of a 3×3 size and are followed by a maximum pooling layer with a 2×2 window. The last two convolutional layers have 256 filters with a 3×3 size. All convolutional layers use the rectified linear units (ReLU) function, with a stride of 1 and zero-padding, returning an output with the same resolution as the input.

We evaluated two variations of our method for different input images dimensions. The first variation receives an input image with 512×512 resolution and produces a feature map in the final layer with 64×64 resolution. Proportionally, the second variation receives an input images with 1024×1024 pixels, and the output feature map has a resolution of 128×128. Despite the low resolution, this map can describe relevant features extracted from the image.

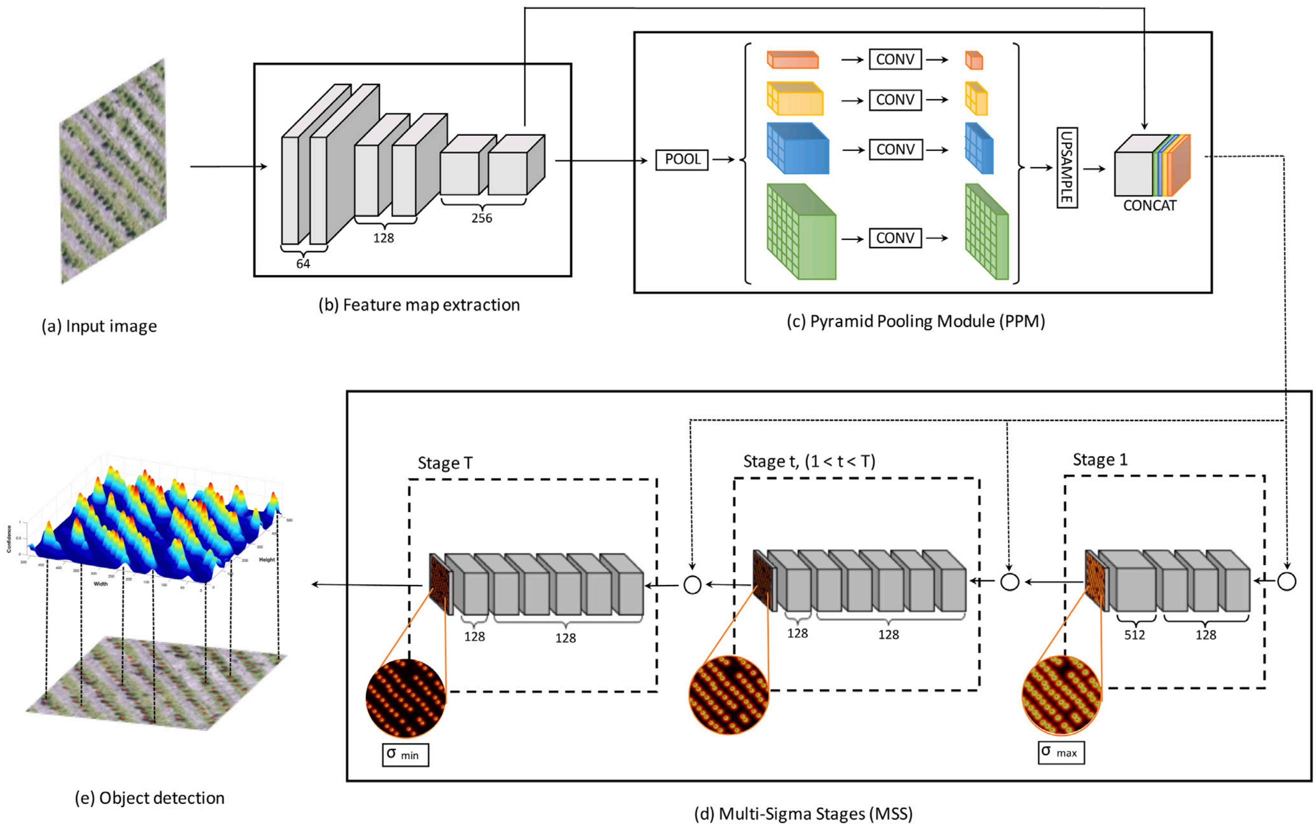


Fig. 1. Our method for the confidence map prediction using the Pyramid Pooling Module (PPM) and the multi-sigma refinement approach. The initial part (b), based on VGG19 (Simonyan & Zisserman, 2015), extracts a feature map from the input image (a). This feature map is used as input for the PPM (c) (Zhao et al., 2017). The resulting volume is then used as input to the first stage of a Multi-Sigma Stages (MSS) phase (d) (Aich & Stavness, 2018). The concatenation of the PPM and the prediction map of the previous stage is used as input for the remaining stages. The T stages apply a standard deviation (σ) for the confidence map peak, starting at maximum-to-minimum so that values are spaced equally.

2.2. Improving feature map with Pyramid Pooling Module

Many CNN cannot incorporate sufficient global context information to ensure a good performance in characterizing high-density objects. To solve this issue, our method adopts a global and subregional context module called PPM (Zhao et al., 2017). This module allows CNN to be invariant to scale since it associates subregional and global information in the feature map. Fig. 1(c) illustrates the PPM that combines the features of four pyramid scales, with resolutions of 1 × 1, 2 × 2, 3 × 3 and 6 × 6, respectively.

The highest general level, shown in orange, applies a global max pooling which creates a 1 × 1 feature map to describe the global image context, such as the number of detected objects in the image. The other levels divide the input map into subregions, forming a grouped representation of the image with their subcontext information, as dense or sparse regions.

The levels of the PPM contain feature maps with various sizes. Because of this, we used a 1 × 1 convolution layer with 512 filters after each level. We upsampled the feature maps to the same size as the input map with bilinear interpolation. Lastly, these feature maps are concatenated with the input map to form an improved description of the image. This step ensures that small object information is not lost in the PPM phase.

Although this module is proposed for semantic segmentation, it has proven to be a robust method for counting objects according to our experiments. The module allowed image information at different scales and its global context to be grouped with the feature map for a better description of the input image, improving the detection performance.

2.3. Multi-sigma refinement

In the multi-sigma refinement phase, the improved feature map obtained by PPM is used as input for the T stages that estimates the confidence map. The first stage (Fig. 1(d)) receives the feature map and generates the confidence map C_1 by using five convolutional layers: three layers with 128 filters with a 3 × 3 size; one layer with 512 filters with a 1 × 1 size; and one layer with a single filter, corresponding to the confidence map.

At a subsequent stage t (Fig. 1(d)), the prediction returned by the previous stage C_{t-1} and the feature map from the PPM process are concatenated. They are used to produce a refined confidence map C_t. The $T-1$ final stages consist of seven convolutional layers: five layers with 128 filters with a 7 × 7 size; and one layers with 128 filters with a 1 × 1 size. The last layers have a sigmoid activation function so that each pixel represents the probability of the occurrence of an object (values between $[0,1]$). The remaining layers have a ReLU activation function. Through the multiple stages, we proposed hierarchical learning of the center of the object. The first stage roughly predicts the position, while the other stages refine this prediction (Fig. 5).

To avoid the vanishing gradient problem during the training phase, we adopted a loss function (Eq. (1)) to be applied at the end of each stage.

$$f_t = \sum_{p} \| \hat{C}_t(p) - C_t(p) \|_2^2, \tag{1}$$

where $\hat{C}_t$ is the ground truth confidence map of the stage t (Section 2.4). The overall loss function is given by:

$$f = \sum_{t=1}^{T} f_t \tag{2}$$

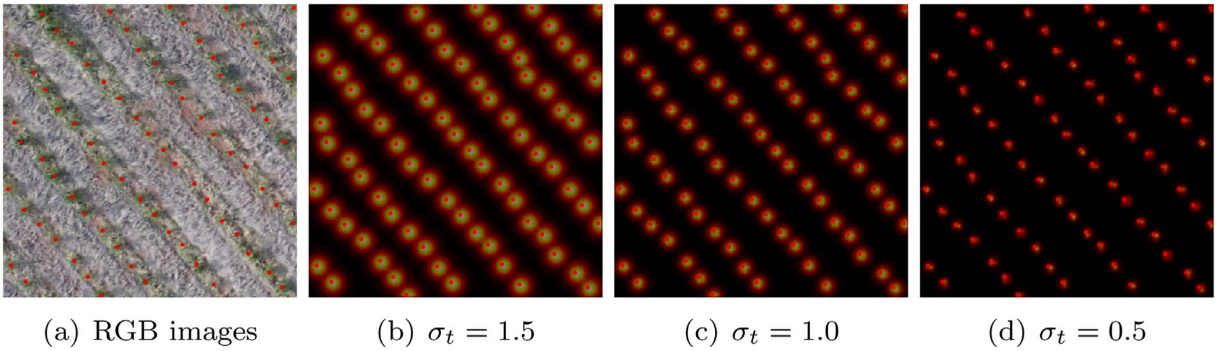


Fig. 2. Example of an RGB image and its corresponding ground-truth confidence maps with different σ_t values.

2.4. Generation of confidence maps

As mentioned in the previous section, to train our method, a confidence map $\hat{C}_t$ is generated as a ground truth for each stage t by using the center of the objects as annotations in the image. The $\hat{C}_t$ is generated by placing a 2D Gaussian kernel at each center of the labeled objects (Aich & Stavness, 2018). The Gaussian kernel has a standard deviation (σ_t) that controls the spread of the confidence map peak, as shown in Fig. 2.

Our approach uses different values of σ_t for each stage t to refine the object center prediction during each stage. The σ_1 of the first stage is set to a maximum value (σ_{max}) while the σ_T of the last stage is set to a minimum value (σ_{min}). The appropriate values of σ_{max} and σ_{min} are evaluated in the experiments. The σ_t for each intermediate stage is equally spaced between $[\sigma_{max}, \sigma_{min}]$. The early stages should return a rough prediction of the center of the objects, and this prediction is refined in the subsequent stages.

Fig. 2 illustrates an example of a ground truth confidence map with three values of σ_t. Fig. 2(a) shows the RGB image and the locations of each objected marked by a red dot. Fig. 2 (b, c, and d) present the ground truth confidence maps for $\sigma_t = 0.5, 1.0$ and 1.5, respectively. In our experiment, the usage of different σ helped refine the confidence map, improving its robustness.

2.5. Object localization from confidence map

Object locations are obtained from the confidence map of the last stage (C_T). We estimate the peaks (local maximum) of the confidence map by analyzing the 4-pixel neighborhood of each given location of p. Thus, $p = (x_p, y_p)$ is a local maximum if $C_T(p) > C_T(v)$ for all the neighbors v, where v is given by $(x_p \pm 1, y_p)$ or $(x_p, y_p \pm 1)$. An example of the object location from the confidence map peaks is shown in Fig. 3.

To avoid noise or low probability of occurrence of the positions p, a peak in the confidence map is considered as an object only if $C_T(p) > \tau$. Besides that, we set a minimum distance δ to allows the method to detect very close objects. After preliminary experiment, we used $\tau = 0.35$ and $\delta = 1$ pixel, that allows the detection of objects from two pixels of distances.

3. Experiments

3.1. Image datasets

To test the robustness of our method, we evaluated it in a new and challenging dataset of eucalyptus tree images. We used this image dataset because there are different tree plantation densities, ranging from extreme cases to more sparsed trees (Fig. 4). This variation in density is a challenge for counting and locating objects. The trees were also at different growth stages. This permitted to evaluate the proposed method in different scales (tree size) and changes in appearance.

The images were captured by an Unmanned Aerial Vehicle (UAV) in a rural property in Mato Grosso do Sul, Brazil, over four different areas of approximately 40 ha each. The eucalyptus trees were planted at different spacing, the densest being at 1.25 m from each other, with an average of 1750 trees per hectare. These trees were at different growth stages, variating between high and canopy areas. The images were acquired with an RGB sensor, which produced a pixel size of 4.15 cm. A total of four orthomosaic were generated from the area of interest. Approximately 232,000 eucalyptus trees were labeled as a point feature by a specialist.

To evaluate the robustness and generability of the proposed approach, we also compared the performance of our method in two well-known image datasets for counting cars: CARPK and PUCPR+ benchmarks (Hsieh et al., 2017). We compare the prediction metrics with state-of-the-art methods such One-Look Regression (Mundhenk et al., 2016), IEP Counting (Stahl et al., 2019), YOLO (Redmon & Farhadi, 2017), YOLO9000 (Redmon & Farhadi, 2017), Faster R-CNN (Ren et al., 2017), RetinaNet (Hsieh et al., 2017; Lin et al., 2020), LPN (Hsieh et al., 2017), VGG-GAP (Aich & Stavness, 2018), VGG-GAP-HR (Aich & Stavness, 2018) and Deep IoU CNN (Goldman et al., 2019).

3.2. Experimental setup

The four orthomosaics were split into 3370 patches with 512 × 512 pixels without overlapping. These patches were randomly divided into training ($n = 2870$), validation ($n = 250$) and testing ($n = 250$) sets. For training the CNN, we applied a Stochastic Gradient Descent optimizer with a momentum of 0.9. To reduce the risk of overfitting, we used the validation set for the hyperparameter tuning on the learning rate and the number of epochs. After minimal hyperparameter tuning, the learning rate was 0.01 and the number of epochs was equal to 100. Instead of training the proposed approach from scratch, we initialized the weights of the first part with pre-trained weights in ImageNet. Six regression metrics, the mean absolute error (MAE) (Chai & Draxler, 2014; Wackerly et al., 2014), root mean squared error (RMSE) (Chai & Draxler, 2014; Wackerly et al., 2014), the coefficient of determination (R^2) (Draper & Smith, 1998), the Precision, Recall, and the F-Measure, were used to measure the performance. Training and testing were performed in 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.

4. Results and discussion

This section presents and discusses the results obtained by the proposed method while comparing it with state-of-the-art methods. First, we demonstrate the influence of different parameters, which includes

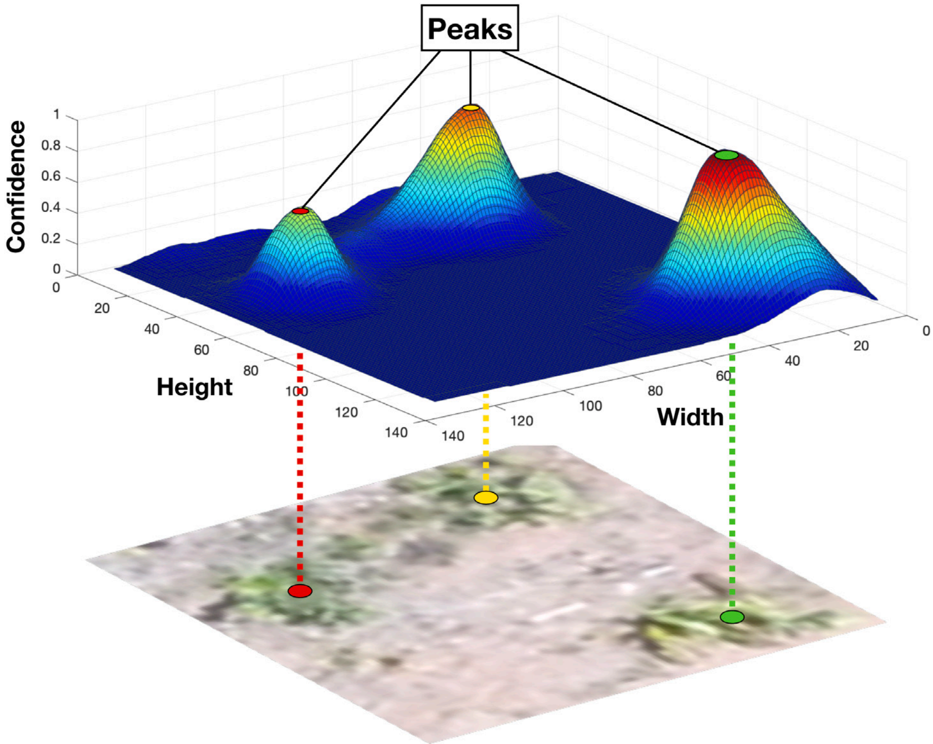


Fig. 3. Example of the localization of eucalyptus trees from a refined confidence map.

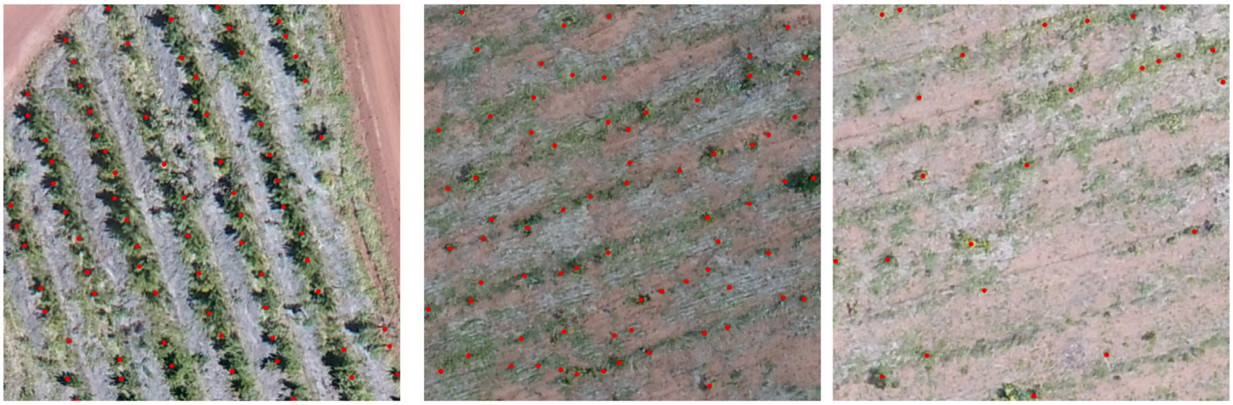


Fig. 4. Examples of the tree dataset. The eucalyptus trees are at different growth stages and plantation densities.

the σ to generate the ground truth confidence maps, the number of stages necessary to refine the prediction, and the usage of PPM (Zhao et al., 2017) to include context information based on multiple scales. Second, we compare the results with a baseline of the proposed method. For this, we used the tree counting dataset and the car counting datasets (CARPK and PUCPR+).

4.1. Parameter analysis

We present the results of the proposed method in the validation set for a different number of stages on the tree counting dataset. These stages are responsible for refining the confidence map. We observed that by using two stages ($T = 2$), the proposed method already returned satisfactory results (Table 1). When increasing to $T = 4$ stages, we obtained the best result, with MAE, RMSE, R^2, Precision, Recall and F-Measure of 2.69, 3.57, 0.977, 0.817, 0.831, and 0.823, respectively. These results indicate the multi-sigma refinement affect the object counting tasks significantly. This is because the confidence map is refined in later stages, increasing the chance of objects be detect in high-density regions. Thus, we verified that the increase in the number of stages is decisive for a good refinement of the predictions. With $T = 6$ or more stages we see that the performance stabilizes and begins to decrease, due to the deepening of the layers.

Table 1
Evaluation of the number of stages (T) on the validation set of the tree counting dataset using $\sigma_{min} = 1$ and $\sigma_{max} = 3$.

Stages (T)	MAE	RMSE	R^2	Precision	Recall	F-Measure
2	2.86	3.82	0.974	0.809	0.825	0.816
4	**2.69**	**3.57**	**0.977**	**0.817**	**0.831**	**0.823**
6	3.48	4.61	0.962	0.805	0.836	0.819
8	2.90	3.79	0.974	0.816	0.823	0.818
10	3.32	4.25	0.967	0.789	0.796	0.790

We evaluated the σ_{min} and σ_{max} responsible for generating the ground truth confidence maps implemented in the T stages. In this experiment, we adopt $T = 4$ stages that achieved the best results from

Table 2
Evaluation of the σ_{max} in the validation set of the tree counting dataset. We adopted the $\sigma_{min} = 1$ and stages $T = 4$.

σ_{max}	MAE	RMSE	R^2	Precision	Recall	F-Measure
2	3.31	4.31	0.966	0.811	0.837	0.822
3	**2.69**	**3.57**	**0.977**	**0.817**	**0.831**	**0.823**
4	3.21	4.24	0.968	0.804	0.816	0.809

Table 3
Evaluation of the σ_{min} in the validation set of the tree counting dataset. We used $\sigma_{max} = 3$ and stages $T = 4$.

σ_{min}	MAE	RMSE	R^2	Precision	Recall	F-Measure
0.5	11.01	13.77	0.658	0.868	0.721	0.783
0.75	2.93	3.89	0.972	0.820	0.831	0.824
1	**2.69**	**3.57**	**0.977**	**0.817**	**0.831**	**0.823**
1.25	3.05	4.01	0.970	0.815	0.822	0.817
1.5	2.94	3.73	0.975	0.818	0.810	0.813

Table 4
Processing time evaluation of the proposed approach for different amounts of T.

Stages (T)	Average Time (s)	Standard deviation
2	0.802	0.022
4	1.426	0.028
6	2.063	0.058
8	2.675	0.059
10	3.373	0.100

the previous experiment. The confidence map from the first stage is generated using σ_{max}, while the last stage uses σ_{min}, and the intermediate stages are constructed from values equally spaced between $[\sigma_{max}, \sigma_{min}]$. A low σ, relative to the object area (e.g., tree canopy) provides a confidence map without correctly covering the object's area. However, a high σ generates a confidence map that, while fully covers the object, may include nearby objects in high-density conditions. These conditions make it difficult to spatially locate objects in the image.

The evaluation for σ_{max} is presented in Table 2. The highest result was obtained with $\sigma_{max} = 3$, which best covers the tree-canopies without overlapping them. Still, we observed that other values for σ_{max} also returned good results. Since σ_{max} is used in the first stage, it does a small influence over the final result, since the confidence map is refined in subsequent stages.

The results for the σ_{min} are summarized in Table 3. The σ_{min} has great influence over the final result since it is responsible for the last confidence map. The overall best result was obtained with a $\sigma_{min} = 1.0$, which achieved a MAE, RMSE, R^2, Precision, Recall and F-Measure of 2.69, 3.57, 0.977, 0.817, 0.831 and 0.823, respectively. This shows that the $\sigma_{min} = 1.0$ is the best fit for the size of the tree canopy. The conducted experiments showed that, with appropriate values of $\sigma_{max} = 3$ and $\sigma_{min} = 1$, high performance for counting trees can be obtained (Table 3).

To verify the potential of our method in real-time processing, we perform a comparison of the processing time performance for different amounts of stages (T). Table 4 shows the processing time of the proposed method for values of $T = 2$, 4, 6, 8 and 10. For this, we used 100 images from the tree test set and extracted the average processing time and standard deviation. We used the values of $\sigma_{min} = 1$ and $\sigma_{max} = 3$ that obtained the best performance in the previous tests. The results showed that the proposed approach can achieve real-time processing. For the best configuration with stages $T = 4$ the approach can deliver an image detection in 1.42 s with a standard deviation of 0.028.

Table 5
Results of the proposed method and its baseline for the tree counting dataset.

Method	MAE	RMSE	R^2	Precision	Recall	F-Measure
Baseline ($\sigma = 0.5$)	11.97	15.10	0.62	0.861	0.709	0.772
Baseline ($\sigma = 1.0$)	2.85	3.72	0.977	0.814	0.833	0.822
Baseline ($\sigma = 2.0$)	3.07	4.37	0.968	0.822	0.805	0.812
Baseline + PPM	2.44	3.38	0.981	0.825	0.836	0.829
Baseline + multi-sigma	2.78	3.64	0.978	0.808	0.833	0.819
Proposed Method	**2.05**	**2.87**	**0.986**	**0.822**	**0.834**	**0.827**

4.2. Tree counting

To analyze the design of the proposed architecture, we compared it with a baseline model that does not include the PPM and the multi-sigma refinement on tree-counting dataset. The overall best result with just the baseline of the CNN was obtained with a $\sigma = 1$, returning an MAE, RMSE, R^2, Precision, Recall, and F-Measure equal to 2.85, 3.72, 0.977, 0.814, 0.833 and 0.822, respectively.

A gain in performance is observable when analyzing the results from the inclusion of the PPM and multi-sigma refinement in the baseline (Table 5). The inclusion of the PPM has no significant improvement for the results, while the baseline with multi-sigma refinement achieves better results. One explanation for this is that multiple stages provide hierarchical learning of the object position, starting from a rough to a more refined prediction of the center of the object. Examples of the confidence map refinement across the stages are shown in Fig. 5. Besides, when we implemented both these two modules, it outperformed all the baselines results. This performance gain can be explained by the sharing of the benefits that the two modules deliver, on the one hand the PPM module delivers subregional and global information in the feature map and the multi-sigma refinement uses this information to refine the objects predictions throughout the stages. The results shows that the combination of these two modules is essential to object counting.

We considered a region around the labeled object position to analyze qualitatively the proximity of the prediction with the center of the object. The results using the best configuration ($\sigma_{min} = 1.0$, $\sigma_{max} = 3.0$, and $T = 4$) is displayed in Fig. 6. The predicted positions are represented by red dots, and the tree-canopies regions are represented by yellow circles whose center is the labeled position. The proposed method can correctly predict most of the tree positions. Another important contribution is that planting-lines are also identified without the need for annotation or additional procedure (Fig. 6(a)). Furthermore, the proposed method can correctly identify trees even outside the planting lines, in a non-regular distribution (Fig. 6(b)).

A comparison of the proposed method with both PPM and multi-sigma refinement against the baseline is displayed in Fig. 7. The baseline fails to detect some trees while returning some false-positives. The proposed method is capable of detecting more difficult true-positives, not detected by the baseline methods, with fewer false-negatives.

Although the proposed method returned a good performance for the tree counting dataset, it also had some challenges (Fig. 8). The "far-from-center" predictions occurred in short planting-lines (Fig. 8 (a)) or in disperse vegetation. This also happened in highly dense areas (Fig. 8 (b)), although in fewer occurrences. Still, the proposed method was capable of predicting the correct position of the majority of trees.

4.3. Density analysis

To verify the performance of the proposed approach for object detection in different types of densities, we divided the tree dataset of 250 images into three density groups: low, medium and high. For this, the images were ordered according to the number of trees annotated, then the three groups were defined based on the quantities of trees in a balanced way. The low corresponds to the images that have up to 52

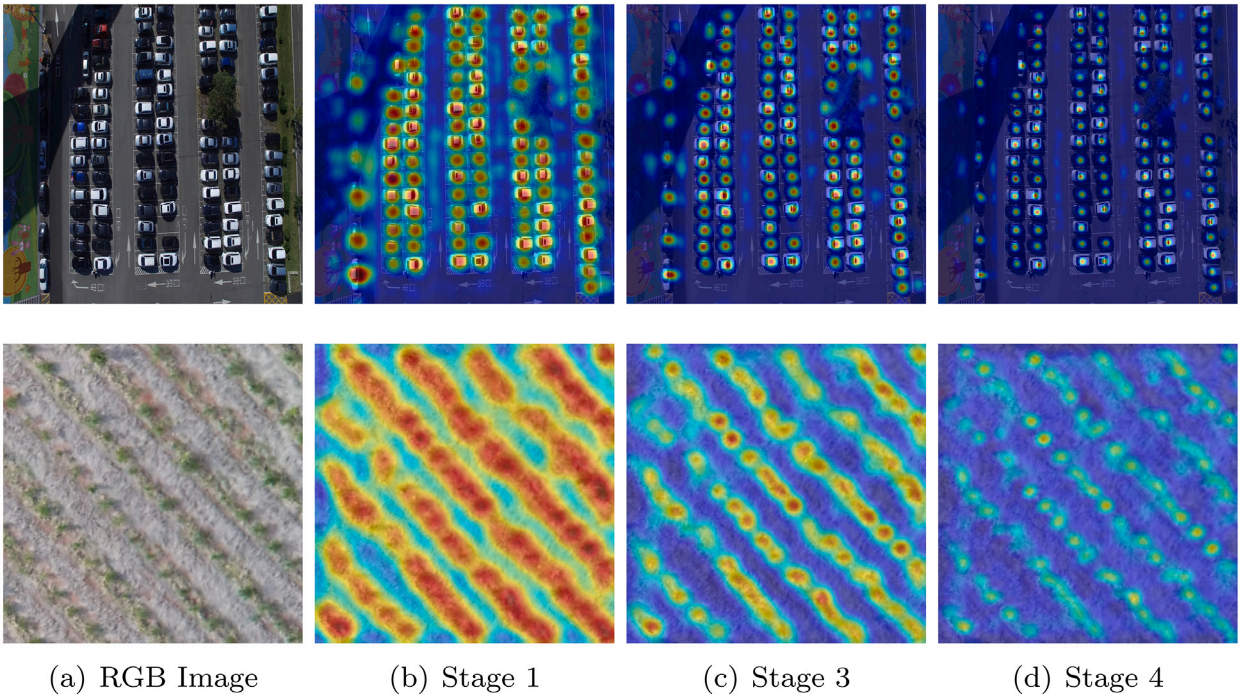


Fig. 5. Example of two images showing the confidence map refinement by our method.

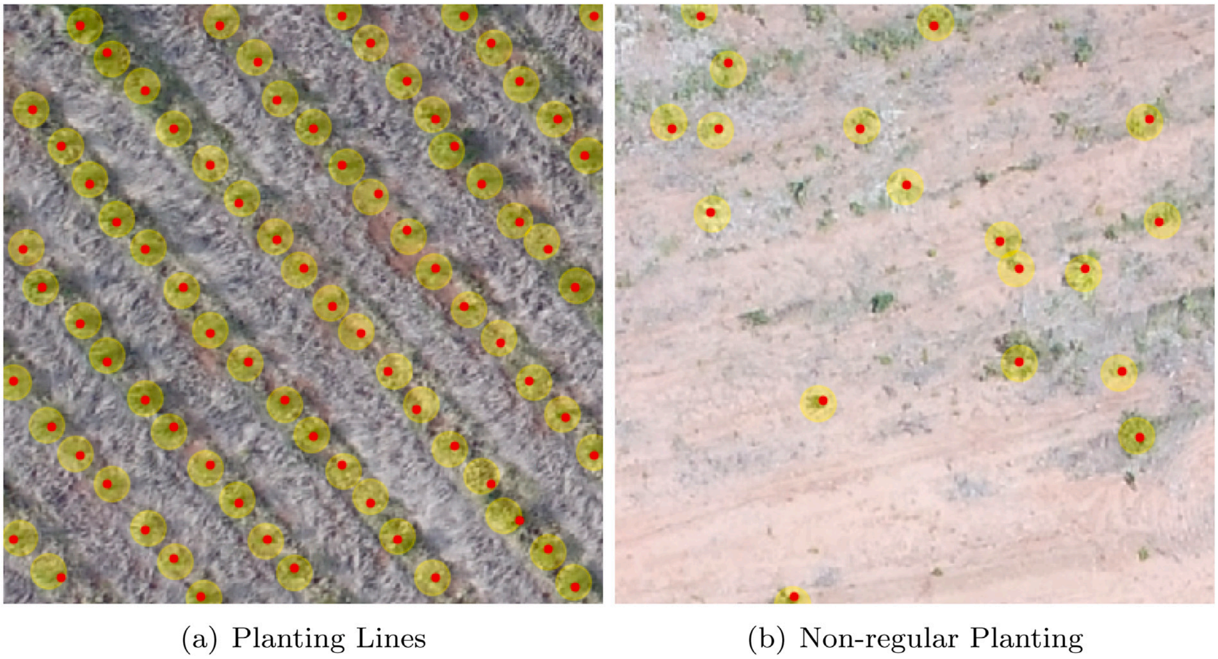


Fig. 6. Comparison of predicted positions (red dots) in two images with different tree density.

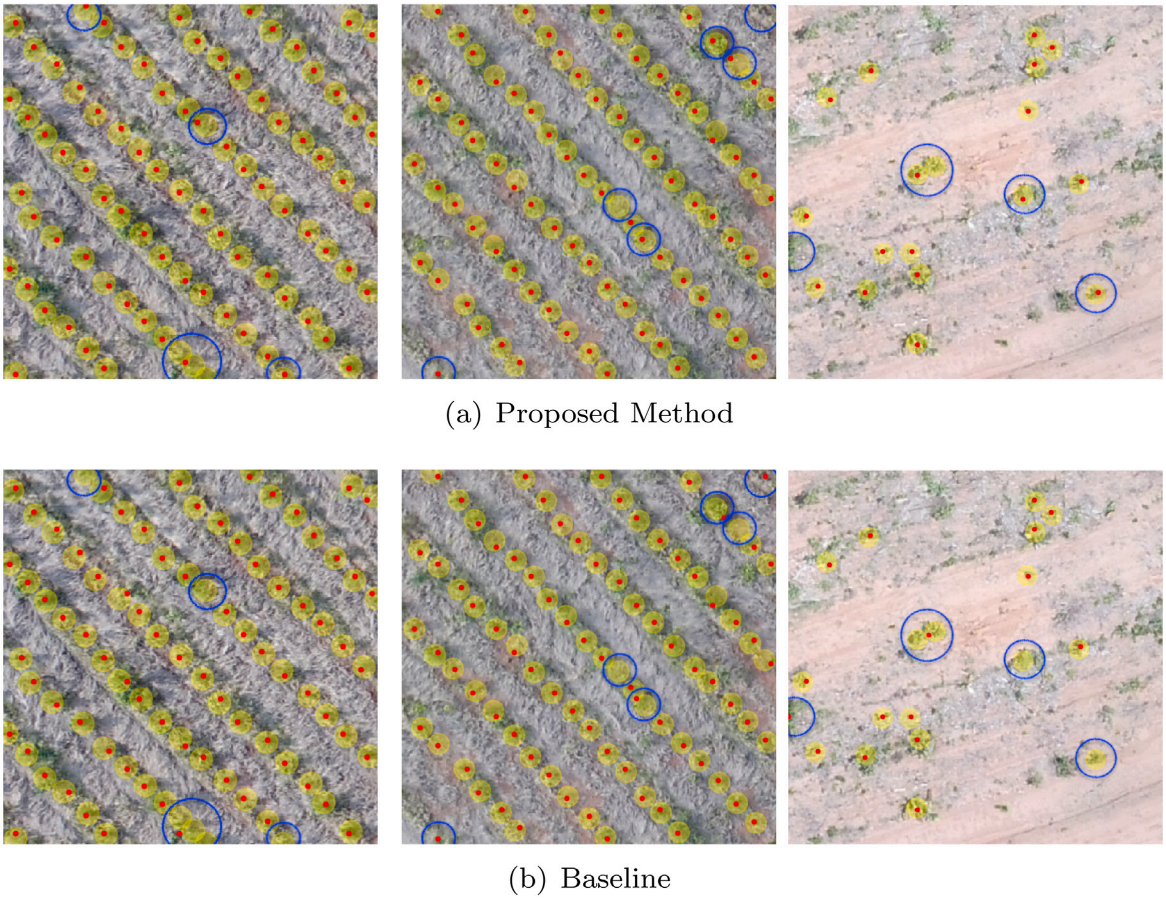
plants, the medium between 53 and 78 plants, and the high above 78 plants. Thus, the sets of low, medium and high test images were left with 83, 90 and 77, respectively.

Table 6 presents the results obtained by the proposed approach at the three density levels. We can see that the approach does equally well at each density level, obtaining better results at the low level achieved an MAE, RMSE, R^2, Precision, Recall, and F-Measure equal to 1.70, 2.34, 0.966, 0.818, 0.846 and 0.829, respectively.

Fig. 9 shows the visual results for plant detection at the three density levels. We can see that the proposed approach is able to correctly detect the centers of the plants, even in irregular plantings (see Fig. 9(a) and (b)). In addition, as shown in Table 6 we can see that at the low level the approach detects the plants positions more easily, since there is not much overlap of the tree-canopies.

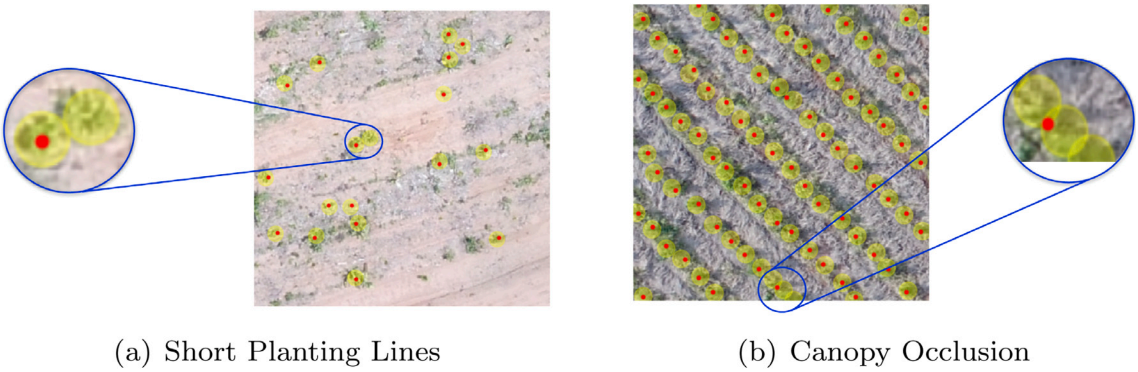
Table 6
Results of the proposed method for different object densities.

Density level	MAE	RMSE	R^2	Precision	Recall	F-Measure
Low	1.70	2.34	0.966	0.818	0.846	0.829
Medium	2.10	2.85	0.865	0.824	0.829	0.826
High	2.38	3.36	0.843	0.823	0.826	0.824


(a) Proposed Method

(b) Baseline

Fig. 7. Comparison of the predicted positions of (a) the proposed method and (b) the baseline. Predicted positions are shown by red dots while tree-canopies are represented by yellow circles. Blue circles show the challenges faced by the methods.


(a) Short Planting Lines

(b) Canopy Occlusion

Fig. 8. Examples of the challenges faced by the proposed method.

4.4. Experiments on cars datasets

To generalize the proposed approach while comparing its robustness against other state-of-the-art methods, we evaluated its performance in two well-known benchmarks: CARPK and PUCPR+ (Hsieh et al., 2017). These benchmarks provide a large-scale aerial dataset for counting cars in parking lots. We adopted the same protocols for the training and testing sets. The images have been resized to 1024 × 1024 pixels since we obtained similar performance when using full-resolution images in our approach.

To perform these experiments, we compare the proposed approach with state-of-the-art methods: One-Look Regression (Mundhenk et al., 2016), IEP Counting (Stahl et al., 2019), YOLO and YOLO9000 (Redmon & Farhadi, 2017), Faster R-CNN (Ren et al., 2017), RetinaNet (Hsieh et al., 2017; Lin et al., 2020), LPN (Hsieh et al., 2017), VGG-GAP and VGG-GAP-HR (Aich & Stavness, 2018), Deep IoU CNN (Goldman et al., 2019), GSP (Aich & Stavness, 2019), Crowd-SDNet (Wang et al., 2021) and GAnet (YuanQiang et al., 2020).

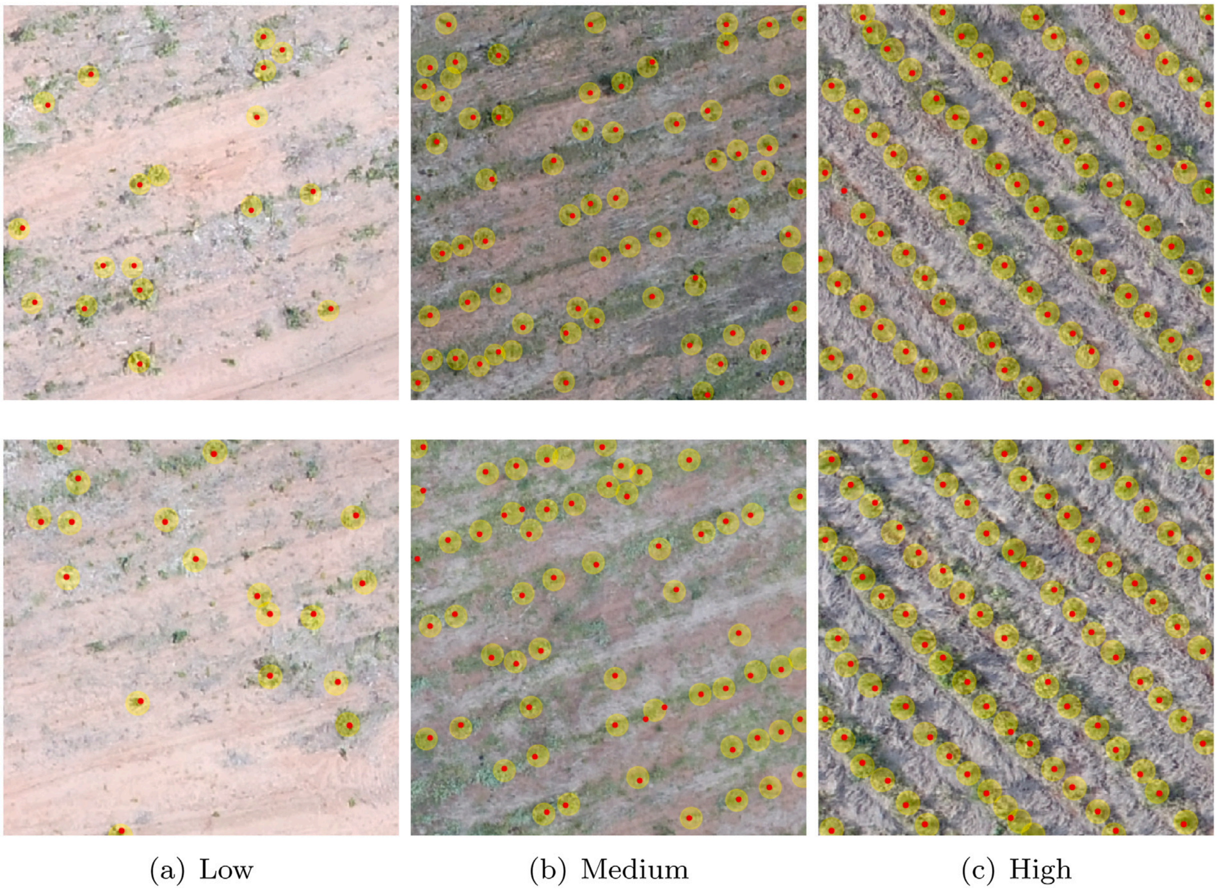


Fig. 9. Examples of the performance of the proposed approach at different levels of object densities. Column (a) shows the results for low densities, (b) for medium densities and (c) for high densities.

4.4.1. *Experiments on CARPK dataset*

The CARPK dataset (Hsieh et al., 2017) is composed of 989 training images (42,274 cars) and 459 test images (47,500 cars). The number of cars per image ranges from 1 to 87 in training images, and from 2 to 188 in test images.

Unlike the images of trees that we seek to cover its canopy, in the car images the confidence map seeks to cover the surface of the vehicle to correctly identify the objects. Table 7 presents the comparison with state-of-the-art methods. We can see that recent approaches such as Crowd-SDNet (Wang et al., 2021) and GAnet (YuanQiang et al., 2020) reached a MAE of 4.95 and 4.61, and an RMSE of 7.09 and 6.55, respectively. Traditional approaches such as Faster R-CNN, YOLO and RetinaNet achieved a MAE of 24.32, 45.36 and 16.62, and an RMSE of 37.62, 52.02 and 22.30. The proposed approach reached a MAE and an RMSE of 4.45 and 6.18, in addition it had a Precision, Recall and F-Measure of 0.767, 0.765 and 0.763, respectively.

Similar to this work, GSP (Aich & Stavness, 2019) also estimates an activation map indicating the positions of the objects. Although it obtains relevant results, the proposed method delivers a gain of 1.01 and 1.91 for MAE and RMSE, respectively. In Fig. 10 the visual comparison of the activations generated by the GSP and the proposed approach with its refinement in multiple stages is presented. We can observe that following the quantitative results the proposed approach delivers more refined predictions, achieving greater performance.

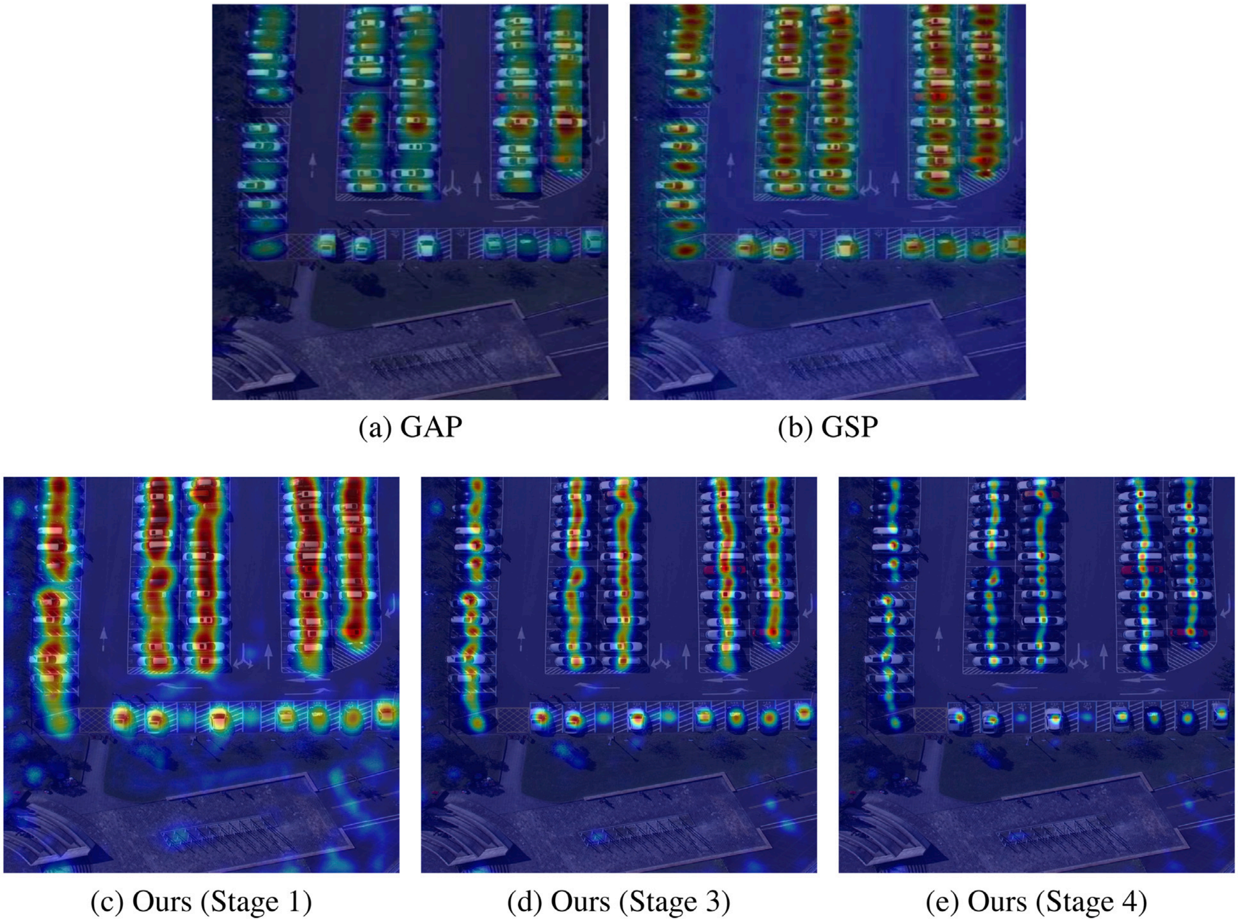
We observed that the proposed method achieved state-of-the-art performance in counting cars. As shown in Fig. 11, the proposed method improves the results by detecting more difficult true-positives. Some cars are partially covered by trees or shadows (Fig. 11(a)) while others are partially occluded (Fig. 11(b)) at the edge of the images. Our method was able to detect such cases. The PPM helped improve the object representation, while the multi-sigma refinement provided a better position in the center of the objects. These features, incorporated in our approach, provide to be important additions for the detection of objects in these challenging scenarios.

Table 7
CARPK comparative results.

Method	MAE	RMSE	R^2	Precision	Recall	F-Measure
One-Look Regression	59.46	66.84	–	–	–	–
IEP Counting	51.83	–	–	–	–	–
YOLO	48.89	57.55	–	–	–	–
YOLO9000	45.36	52.02	–	–	–	–
Faster R-CNN	24.32	37.62	–	–	–	–
RetinaNet	16.62	22.30	–	–	–	–
LPN	13.72	21.77	–	–	–	–
VGG-GAP	10.33	12.89	–	–	–	–
VGG-GAP-HR	7.88	9.30	–	–	–	–
Deep IoU CNN	6.77	8.52	–	–	–	–
GSP	5.46	8.09	–	–	–	–
Crowd-SDNet	4.95	7.09	–	–	–	–
GAnet	4.61	6.55	–	–	–	–
Proposed Method	4.45	6.18	0.975	0.767	0.765	0.763

4.4.2. *Experiments on PUCPR+ dataset*

PUCPR+ (Hsieh et al., 2017) is a subset of the PUCPR dataset 1u (de Almeida et al., 2015), and it is composed of 100 training images



(a) GAP

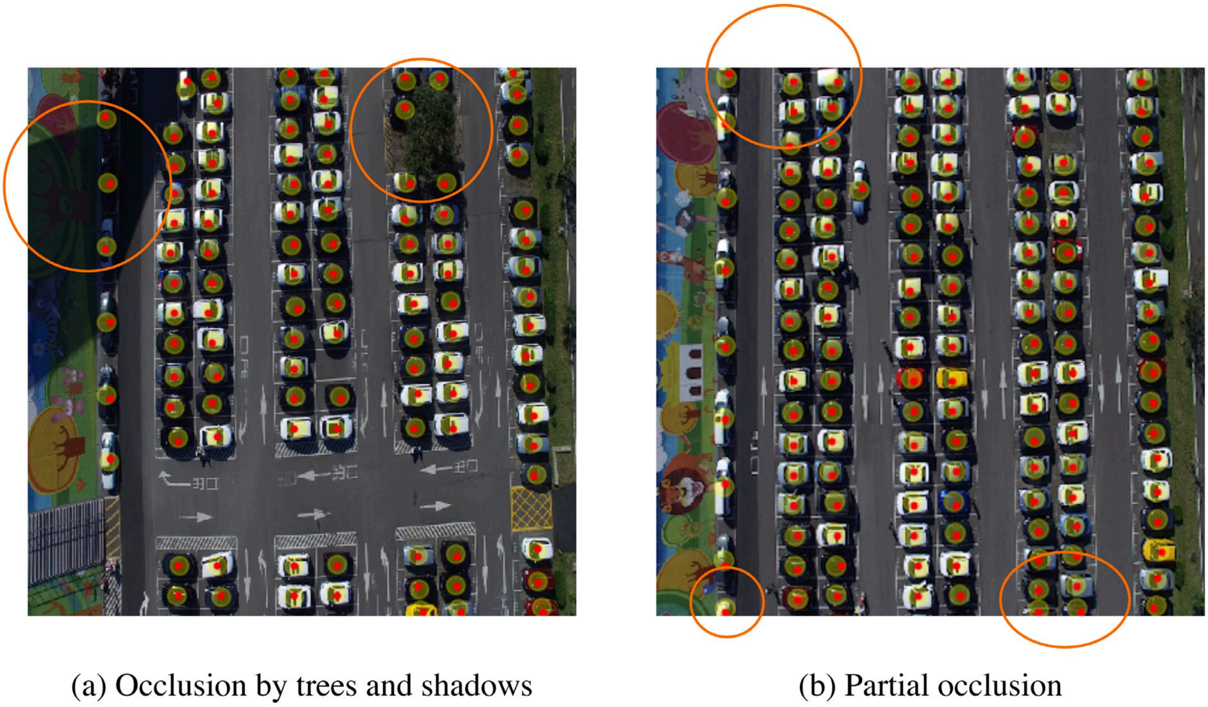
(b) GSP

(c) Ours (Stage 1)

(d) Ours (Stage 3)

(e) Ours (Stage 4)

Fig. 10. Comparison of the activations generated by GAP and GSP approaches (first row) adapted from Aich and Stavness (2019), and by the multiple stages of refinement of the proposed approach (second row).



(a) Occlusion by trees and shadows

(b) Partial occlusion

Fig. 11. Car detection by the proposed method on the CARPK dataset. Figure (a) shows the detections in scenarios of occlusions by trees and shadows, while figure (b) shows the cars partially hidden at the end of the image. Orange circles highlight challenging cases.

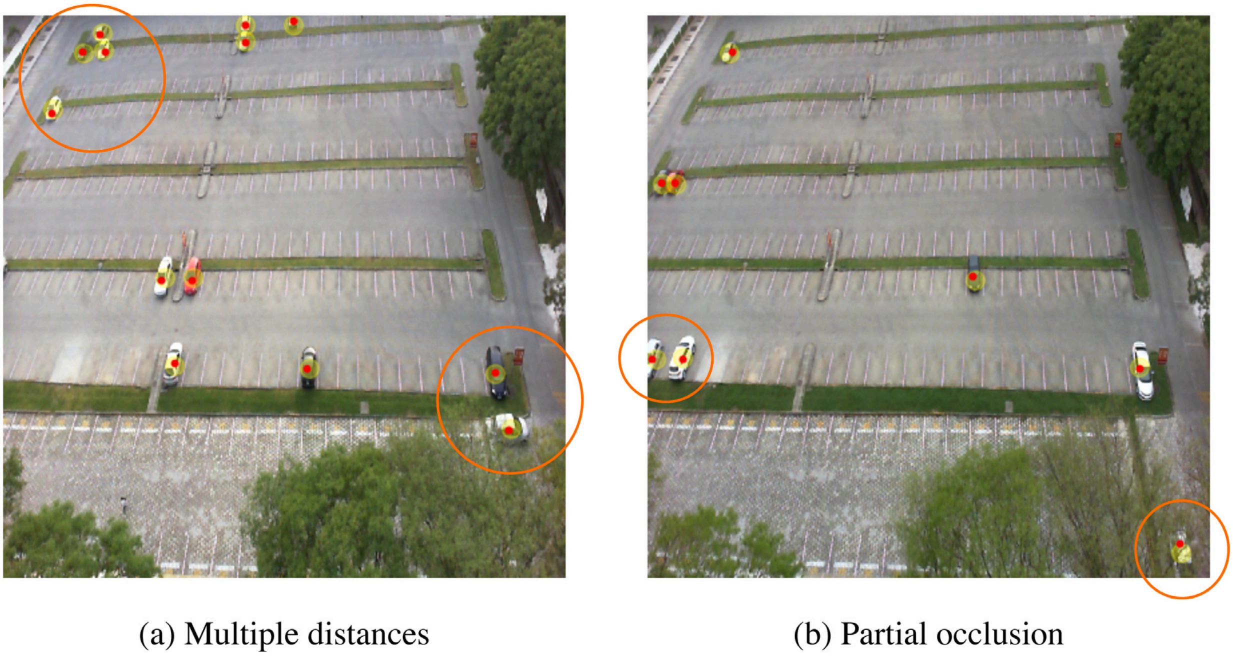


Fig. 12. Car detection by the proposed method on the PUCPR+ dataset. Figure (a) shows the detections in scenarios from multiple distances between overlapping objects and figure (b) shows the cars partially hidden by trees and at the end of the image. Orange circles highlight challenging cases.

Table 8
PUCPR+ comparative results.

Method	MAE	RMSE	R^2	Precision	Recall	F-Measure
YOLO	156.00	200.42	–	–	–	–
YOLO9000	130.40	172.46	–	–	–	–
Faster R-CNN	39.88	47.67	–	–	–	–
RetinaNet	24.58	33.12	–	–	–	–
One-Look Regression	21.88	36.73	–	–	–	–
IEP Counting	15.17	–	–	–	–	–
VGG-GAP	8.24	11.38	–	–	–	–
LPN	8.04	12.06	–	–	–	–
Deep IoU CNN	7.16	12.00	–	–	–	–
VGG-GAP-HR	5.24	6.67	–	–	–	–
GAnet	3.28	4.96	–	–	–	–
Crowd-SDNet	3.20	4.83	–	–	–	–
Proposed Method	3.16	4.39	0.999	0.832	0.829	0.830

and 25 test images. The training and test images contain respectively 12,995 and 3920 car instances.

Table 8 presents the comparison with 12 state-of-the-art methods for the PUCPR+ dataset. Again, we note that the approaches GAnet (YuanQiang et al., 2020) and Crowd-SDNet (Wang et al., 2021) reached a MAE of 3.28 and 3.20, and an RMSE of 4.96 and 4.83, respectively. In the same way as observed for the CARPK dataset, the traditional approaches Faster R-CNN, YOLO and RetinaNet achieved intermediate performances with MAE of 39.88, 130.40 and 24.58, and an RMSE of 47.67, 172.46 and 4.58. This shows that traditional methods of object detection are not suitable for dense scenes. The proposed approach reached a MAE and an RMSE of 3.16 and 4.39, and obtained a Precision, Recall and F-Measure of 0.832, 0.829 and 0.830, respectively.

Fig. 12 presents the detections obtained by the proposed approach on the PUCPR+ dataset. Due to the point of view of the camera, the cars appear closer and distant in the same image. Thus, the results help to assess the generalization of the approach to recognize objects at different scales and with overlap (Fig. 12(a)). Since PPM adds multi-scale information to objects and multi-sigma refines detections, especially in highly dense areas, we see that the proposed approach achieves good detections even in these challenging scenes. Following the results in the CARPK dataset, the proposed approach achieves good performance in occlusion situations (Fig. 12(b)).

5. Conclusion

In this study, we proposed a new method based on a CNN which returned state-of-the-art performance for counting and locating objects with a high-density in images. The proposed approach is based on a density estimation map with the confidence that an object occurs in each pixel. For this, our approach produces a feature map generated by a CNN, and then apply an enhancement with the PPM. To improve the predictions of each object, it uses a multi-sigma refinement process, and the object position is calculated from the peaks of the refined confidence maps.

Experiments were performed in three datasets with images containing eucalyptus trees and cars. Despite the challenges, the proposed method obtained better results than previous methods. Experimental results on the CARPK and PUCPR+ indicate that the proposed method improves MAE, e.g., from 6.77 to 4.45 on CARPK and 5.24 to 3.16 on database PUCPR+. The proposed method is suitable for dealing with high object-density in images, returning a state-of-the-art performance for counting and locating objects. Since this is the first object counting and locating CNN method based on a feature map enhancement and a multi-sigma refinement of a confidence map, other types of object detection approaches may benefit from the findings presented here.

Further research could be focused on investigating the impact on object counting for different choices of distribution (other than Gaussian) used to generate the groundtruth confidence map. Predictions other than the confidence map can also help in separating objects in high density, such as predicting the boundaries obtained from the Voronoi diagram.

CRediT authorship contribution statement

Mauro dos Santos de Arruda: Conceptualization, Methodology, Software, Writing – original draft. **Lucas Prado Osco:** Validation, Data curation, Writing – original draft. **Plabiany Rodrigo Acosta:** Methodology, Software. **Diogo Nunes Gonçalves:** Methodology, Software. **José Marcato Junior:** Conceptualization, Validation, Writing –

review & editing. **Ana Paula Marques Ramos:** Resources, Writing – review & editing. **Edson Takashi Matsubara:** Data curation, Writing – review & editing. **Zhipeng Luo:** Writing – review & editing. **Jonathan Li:** Conceptualization, Writing – review & editing. **Jonathan de Andrade Silva:** Methodology, Writing – review & editing. **Wesley Nunes Gonçalves:** Conceptualization, Methodology, Writing – original draft.

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 study was supported by the FUNDECT - State of Mato Grosso do Sul Foundation to Support Education, Science and Technology, Brazil, CAPES - Brazilian Federal Agency for Support and Evaluation of Graduate Education, and CNPq - National Council for Scientific and Technological Development, Brazil. The Titan V and XP used for this research was donated by the NVIDIA Corporation.

References

Aich, S., & Stavness, I. (2018). Improving object counting with heatmap regulation. arXiv:1803.05494.

Aich, S., & Stavness, I. (2019). Global sum pooling: A generalization trick for object counting with small datasets of large images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops*.

de Almeida, P. R., Oliveira, L. S., Britto, A. S., Silva, E. J., & Koerich, A. L. (2015). PKLot – a robust dataset for parking lot classification. *Expert Systems With Applications*, *42*(11), 4937–4949. http://dx.doi.org/10.1016/j.eswa.2015.02.009.

Castillo, O., Melin, P., Ramírez, E., & Soria, J. (2012). Hybrid intelligent system for cardiac arrhythmia classification with fuzzy K-nearest neighbors and neural networks combined with a fuzzy system. *Expert Systems With Applications*, *39*(3), 2947–2955. http://dx.doi.org/10.1016/j.eswa.2011.08.156.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), 1247–1250.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis, Vol. 326*. John Wiley & Sons.

Esen, H., Esen, M., & Ozsolak, O. (2017). Modelling and experimental performance analysis of solar-assisted ground source heat pump system. *Journal Of Experimental & Theoretical Artificial Intelligence*, *29*(1), 1–17. http://dx.doi.org/10.1080/0952813X.2015.1056242.

Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008a). Artificial neural networks and adaptive neuro-fuzzy assessments for ground-coupled heat pump system. *Energy And Buildings*, *40*(6), 1074–1083. http://dx.doi.org/10.1016/j.enbuild.2007.10.002.

Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008b). Forecasting of a ground-coupled heat pump performance using neural networks with statistical data weighting pre-processing. *International Journal Of Thermal Sciences*, *47*(4), 431–441. http://dx.doi.org/10.1016/j.ijthermalsci.2007.03.004.

Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008c). Performance prediction of a ground-coupled heat pump system using artificial neural networks. *Expert Systems With Applications*, *35*(4), 1940–1948. http://dx.doi.org/10.1016/j.eswa.2007.08.081.

Esen, H., Ozgen, F., Esen, M., & Sengur, A. (2009). Artificial neural network and wavelet neural network approaches for modelling of a solar air heater. *Expert Systems With Applications*, *36*(8), 11240–11248. http://dx.doi.org/10.1016/j.eswa.2009.02.073.

Ferrari, A., Lombardi, S., & Signoroni, A. (2017). Bacterial colony counting with convolutional neural networks in digital microbiology imaging. *Pattern Recognition*, *61*, 629–640. http://dx.doi.org/10.1016/j.patcog.2016.07.016, URL http://www.sciencedirect.com/science/article/pii/S0031320316301650.

Fiaschi, L., Koethe, U., Nair, R., & Hamprecht, F. A. (2012). Learning to count with regression forest and structured labels. In *Proceedings of the 21st international conference on pattern recognition* (pp. 2685–2688).

Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., & Hassner, T. (2019). Precise detection in densely packed scenes. In *IEEE conf. on computer vision and pattern recognition* (pp. 5227–5236). arXiv:1904.00853.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2020). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 386–397.

Hsieh, M., Lin, Y., & Hsu, W. H. (2017). Drone-based object counting by Spatially regularized regional proposal network. In *2017 IEEE international conference on computer vision* (pp. 4165–4173). http://dx.doi.org/10.1109/ICCV.2017.446.

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Máadeed, S., Rajpoot, N. M., & Shah, M. (2018). Composition loss for counting, density map estimation and localization in dense crowds. In *European conference on computer vision* (pp. 544–559). http://dx.doi.org/10.1007/978-3-030-01216-8_33.

Kokoulin, A. N., Tur, A. I., & Yuzhakov, A. A. (2018). Convolutional neural networks application in plastic waste recognition and sorting. In *Conference of russian young researchers in electrical and electronic engineering* (pp. 1094–1098). http://dx.doi.org/10.1109/EIConRus.2018.8317281.

Lin, T., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327. http://dx.doi.org/10.1109/TPAMI.2018.2858826.

Liu, Y., Shi, M., Zhao, Q., & Wang, X. (2019). Point in, box out: Beyond counting persons in crowds. In *The IEEE conference on computer vision and pattern recognition*. arXiv:1904.01333.

Ma, W., Wu, Y., Cen, F., & Wang, G. (2020). MDFN: Multi-scale deep feature learning network for object detection. *Pattern Recognition*, *100*, Article 107149. http://dx.doi.org/10.1016/j.patcog.2019.107149.

Mundhenk, T. N., Konjevod, G., Sakla, W. A., & Boakye, K. (2016). A large contextual dataset for classification, detection and counting of cars with deep learning. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision – ECCV 2016* (pp. 785–800). Cham: Springer International Publishing.

Ohn-Bar, E., & Trivedi, M. M. (2017). Multi-scale volumes for deep object detection and localization. *Pattern Recognition*, *61*, 557–572. http://dx.doi.org/10.1016/j.patcog.2016.06.002.

Osco, L. P., dos Santos de Arruda, M., Junior, J. M., da Silva, N. B., Ramos, A. P. M., Moryia, E. A. S., Imai, N. N., Pereira, D. R., Creste, J. E., Matsubara, E. T., Li, J., & Gonçalves, W. N. (2020). A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery. *ISPRS Journal Of Photogrammetry And Remote Sensing*, *160*, 97–106. http://dx.doi.org/10.1016/j.isprsjprs.2019.12.010.

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 6517–6525). http://dx.doi.org/10.1109/CVPR.2017.690.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. http://dx.doi.org/10.1109/TPAMI.2016.2577031.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal Of Computer Vision (IJCV)*, *115*(3), 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

d. S. de Arruda, M., Spadon, G., Rodrigues, J. F., Gonçalves, W. N., & Machado, B. B. (2018). Recognition of endangered pantanal animal species using deep learning methods. In *International joint conference on neural networks* (pp. 1–8). http://dx.doi.org/10.1109/IJCNN.2018.8489369.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.

Sindagi, V. A., & Patel, V. M. (2018). A survey of recent advances in CNN-based single image crowd counting and density estimation. *Pattern Recognition Letters*, *107*, 3–16. http://dx.doi.org/10.1016/j.patrec.2017.07.007.

Stahl, T., Pintea, S. L., & van Gemert, J. C. (2019). Divide and count: Generic object counting by image divisions. *IEEE Transactions On Image Processing*, *28*(2), 1035–1044. http://dx.doi.org/10.1109/TIP.2018.2875353.

Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2014). *Mathematical statistics with applications*. Cengage Learning.

Wang, Y., Hou, J., Hou, X., & Chau, L. P. (2021). A self-training approach for point-supervised object detection and counting in crowds. *IEEE Transactions On Image Processing*, *30*, 2876–2887. http://dx.doi.org/10.1109/TIP.2021.3055632.

Xu, J., Wang, W., Wang, H., & Guo, J. (2020). Multi-model ensemble with rich spatial information for object detection. *Pattern Recognition*, *99*, Article 107098. http://dx.doi.org/10.1016/j.patcog.2019.107098.

Yuan, J., Xiong, H.-C., Xiao, Y., Guan, W., Wang, M., Hong, R., & Li, Z.-Y. (2019). Gated CNN: Integrating multi-scale feature layers for object detection. *Pattern Recognition*, Article 107131. http://dx.doi.org/10.1016/j.patcog.2019.107131.

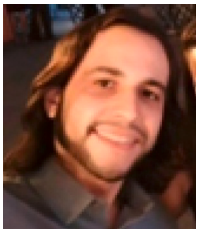
YuanQiang, C., Du, D., Zhang, L., Wen, L., Wang, W., Wu, Y., & Lyu, S. (2020). Guided attention network for object detection and counting on drones. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 709–717). Association for Computing Machinery, http://dx.doi.org/10.1145/3394171.3413816.

Zhang, Y., Bai, Y., Ding, M., Li, Y., & Ghanem, B. (2018). Weakly-supervised object detection via mining pseudo ground truth bounding-boxes. *Pattern Recognition*, *84*, 68–81. http://dx.doi.org/10.1016/j.patcog.2018.07.005.

Zhang, S., Li, H., Kong, W., Wang, L., & Niu, X. (2019). An object counting network based on hierarchical context and feature fusion. *Journal Of Visual Communication And Image Representation*, *62*, 166–173. http://dx.doi.org/10.1016/j.jvcir.2019.05.003.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *The IEEE conference on computer vision and pattern recognition*. arXiv:1612.01105.

Mauro dos Santos de Arruda received his B.Sc. degree in information systems and M.Sc. degrees in computer science from Federal University of Mato Grosso do Sul in 2016 and 2018, respectively. He is completing his Ph.D. thesis at the Faculty of Computer Science of Federal University of Mato Grosso do Sul, Brazil. His current research interests are computer vision, pattern recognition, machine learning, deep neural networks for object detection, classification and segmentation.



Lucas Prado Osco received the Ph.D. degree in agronomy science from the University of Western São Paulo, Brazil. He is currently a Postdoctoral researcher with the Natural Resources Program at the Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil. His current research interests include remote sensing, proximal sensing, vegetation analysis, precision agriculture, machine learning and deep neural networks for object detection, classification and segmentation.

Plabiany Rodrigo Acosta received his B.Sc. degree in computer science from Federal University of Mato Grosso do Sul, Brazil. He is a master student at the Faculty of Computer Science, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil. His current research interests include computer vision and deep neural networks for object detection, classification and segmentation.

Diogo Nunes Gonçalves received his B.Sc. degree and M.Sc. degrees in computer science from Federal University of Mato Grosso do Sul in 2016 and 2018, respectively. He is completing his Ph.D. thesis at the Faculty of Computer Science of Federal University of Mato Grosso do Sul, Brazil. His current research interests are computer vision, machine learning, and deep neural networks for classification and segmentation.

José Marcato Junior received the Ph.D. degree in cartographic science from the São Paulo State University, Brazil. He is currently a Professor with the Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil. His current research interests include UAV photogrammetry and deep neural networks for object detection, classification and segmentation.

Ana Paula Marques Ramos received the Ph.D. degree in cartographic science from the SǒPaulo State University, Brazil. She is currently a Professor with the Environment and Regional Development Program, University of Western São Paulo, Presidente Prudente, SP, Brazil. Her current research interests include geosciences, cartograph, environmental science, geoprocessing and remote sensing in environmental applications.

Edson Takashi received the Ph.D. degree in computer science from the University of São Paulo, Brazil. He is currently a Professor with the Faculty of Computer Science, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil. His current research interests include computer vision, machine learning, natural language processing, and deep neural networks for text and images.

Zhipeng Luo received his B.Sc. degree in mathematics from Zhangzhou Normal University in 2013 and M.Sc. degrees in computer science and technology from Fuzhou University, China in 2016, respectively. He is completing his Ph.D. thesis at the School of Informatics of Xiamen University, China. His current research interests are 3D computer vision, 3D pattern recognition, and machine learning for point cloud processing.

Jonathan Li received the Ph.D. degree in geomatics engineering from the University of Cape Town, South Africa. He is currently professor of geomatics and systems design engineering with the University of Waterloo, Canada. His current research interests include information extraction from LiDAR point clouds and earth observation images. He has co-authored more than 400 publications, over 200 of which were published in refereed journals including PR, IEEE-TGRS, IEEE-TITS, IEEE-GRSL, IEEE-JSTARS, ISPRS-JPRS, and RSE. He is Chair of the ISPRS Working Group I/6 on LiDAR for Airborne and Spaceborne Sensing (2016-2020), Chair of the ICA Commission on Sensor-driven Mapping (2019-2023), and Associate Editor of IEEE-TITS, CJRS, and IEEE-JSTARS.

Jonathan de Andrade Silva received the Ph.D. degree in computer science from the University of São Paulo, Brazil. He is currently a Professor with the Faculty of Computer Science, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil. His current research interests include computer vision, machine learning, data stream, and deep neural networks for audio and images.

Wesley Nunes Gonçalves received the Ph.D. degree in computational physics from the University of São Paulo, Brazil. He is currently a Professor with the Faculty of Computer Science and Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil. His current research interests include computer vision, machine learning, texture analysis, and deep neural networks for object detection, classification and segmentation.