Semi-Supervised Deep Learning for Estimating Fur Seal Numbers

Rujia Chen
Auckland University of Technology,
New Zealand
rujia.chen@aut.ac.nz

Roberto Chávez
Pontifical Catholic University of
Valparaíso Valparaíso
Chile
roberto@pucv.cl

Matías Pérez
Pontifical Catholic University of
Valparaíso Valparaíso
Chile
matias.perez@pucv.cl

Luciano Hiriart-Bertrand
Pontifical Catholic University of
Valparaíso Valparaíso
Chile

Victor Castillo-González
Pontifical Catholic University of
Valparaíso Valparaíso
Chile

Renato Borras-Chavez
Pontifical Catholic University of
Valparaíso Valparaíso
Chile

Ajit Narayanan
Auckland University of Technology,
New Zealand
ajit.narayanan@aut.ac.nz

Akbar Ghobakhlou
Auckland University of Technology,
New Zealand
akbar.ghobakhlou@aut.ac.nz

Abstract— **Having estimates of animal species is of growing importance for conservation and ecological reasons, given the increasing concern about the impact of climate change on fauna worldwide. However, it is difficult and sometimes dangerous to count animal numbers in the wild. Counting and detecting animals from drone images can be expected to become a crucial part of conservation policies based on obtaining up-to-date estimates of population numbers. This paper proposes a deep learning approach, the Faster- RCNN algorithm, to count fur seals on the Alejandro Selkirk Island using drone images. Using a semi-supervised approach, the experimental results show the overall precision to be 0.86. This preliminary research shows that machine learning for remote sensing via drone images is helpful for estimating fur seal numbers and could be extended to other areas where it is important to quickly estimate animal populations for the purpose of ecology and conservation.**

Keywords— Object detection, Deep learning, Remote monitoring, Faster R-CNN. *(key words)*

I. INTRODUCTION

Species abundance is one of the Essential Biodiversity Variables (EBV) according to the international effort GEO-BON (Group on Earth Observation - Biodiversity Observation Network) supporting the Convention on Biological Diversity. EBV is essential to understand animal population dynamics and generating appropriate actions for population conservation over time [1]. For endangered species, obtaining systematic information on population abundance over time is vital to assess trends and take necessary actions to avoid the collapse of the species.

Detecting individual animals is essential to observe the presence and distribution of endangered and invasive species [2,3]. Aerial photography and satellite imagery have been traditionally used by biologists to survey remote species across vast areas. Direct observation and traditional surveying are subject to observer bias, often laborious, expensive, logistically challenging, and possibly dangerous depending on the species. For example, in [2], a research, a team of three to five people camped overnight in stone huts to perform the ground count. In a comparison study, a voluntary project was launched to recruit over 5000 volunteers to help count the number of elephant seals, sea lions, and other species on Año Nuevo Island [3]. Periodically monitoring fur seal population changes can reveal the local ecosystem and environmental health to help us conserve the species.

Automated methods based on computer vision techniques for the detection and counting of individual animals could accelerate animal survey analyses. It can also reduce bias and increase the accuracy and precision of the wildlife survey census process. In recent years, deep Convolutional Neural Networks (CNN) have been shown to be successful in a variety of computer vision tasks such as image classification, semantic segmentation, and object detection [4]. CNN-based large-scale object detection methods have been used to estimate the number of buildings [5], cars [6], crowds [7] and many other objects [8].

The first problem for machine learning approaches in animal detection and counting from images is that there are no counter-examples to help the model converge to accurate prediction. That is, only the animals of interest are 'labelled' in training images through, typically, bounding boxes being placed around them manually by human experts. Rocks, other animals and features of no interest are not labelled at all. The machine learning task is to generalize from multiple labelled objects of interest (which may number hundreds or even thousands) in the training image to an unknown number of unlabelled objects of interest (which may also number hundreds or thousands) in a test image by locating bounding boxes around those objects of interest. Objects of no interest should have no bounding box in the test image, even though no counter-examples of objects of no interest are in the training set. After testing, the task is to count the number of bounding boxes in the test image and reach a conclusion on how many objects of interest there are in the test image. The generalization ideally should not have false positives (such as rocks or other objects of no interest). Hence, the approach is 'semi-supervised' in that objects of only one class are labelled.

The second problem is that, in most animal population studies, multiple images will be captured, for instance, through drones flying over the colony at a low height for maximum resolution. Ideally, an automatic population count or estimation system should learn from just one or two images manually labelled by experts to multiple images that are not or cannot be totally labelled because of the amount of time and labour required to for such labelling and counting. Even manually and accurately labelling one image containing several hundred objects of interest can take a human expert several hours. There is therefore a problem of how to ensure that, from a relatively small number of images labelled with all the objects of interest, there is enough information to allow for generalization to an unknown number of images with unknown numbers of objects of interest for testing. A suitable training and testing strategy involving images with known objects of interest must be identified to give confidence that the system accurately generalizes to an unknown number of unlabelled objects of interest. The approach adopted here is to identify test results only after the model makes a prediction of where an object of interest is located, and then assess that metrically against a 'ground truth' provided by experts. This is another form of 'semi-supervision'.

Traditional supervised machine learning methods like Support Vector Machine (SVM) can achieve up to 0.976 recall on monitoring cows distribution and abundance on high quality UAV videos [9], however, this result is based on small datasets. A real time Fast R-CNN model to detect large animals in UAV images can achieve 0.6 precision at 2.9 images per second [10], which is promising in real-time animal monitoring situations. Mask R-CNN is also a reliable algorithm for counting animals from quadcopter images with an accuracy of 0.94 in counting cattle on pastures and 0.92 in feedlots [11]. Detecting marine animals in UAV imagery by a fully automated and supervised CNN model distinguished between fur seals, sea turtles and gannets with recall of 0.94, 0.79 and 0.74, respectively [12]. However, a drop of precision to 0.27 for fur seals led the authors to recommend a semi-supervised method (that is, some manual detection and labelling of seals during training and testing) in future research to help improve precision. For livestock monitoring in Australia, deep learning algorithms are capable of remotely counting and locating animals to track cattle health [13]. These works indicate that a semi-supervised deep learning approach should be explored for fur seal number estimation at Alejandro Selkirk Island as well as for other animal surveys using UAV images.

The aim of this research is to test whether it is possible to estimate the number of seals in a large colony captured on multiple drone images by using a small number of labelled images for training and testing to allow for generalization to the remaining images which are not or cannot be labelled manually. The test images need to be assessed for accuracy using manual labelling depending on model predictions and labelling thresholds for calculating metrics.

II. Materials and Meethods

A. Data Collection

In this study, there are three high-resolution and aerial images of large areas (Figure 1-a) taken from the southwest corner of the island (shaded red in Figure 1-b). Seven smaller images were taken from these large area images for model construction. Figure 1-c shows Alejandro Selkirk Island, which is the largest and most westerly island in the Juan Fernández Archipelago of the Valparaíso Region of Chile. The Juan Fernández Archipelago in Chile is one of the 11 irreplaceable sites for global marine conservation and is one of the 10 priority sites for the conservation of the Valparaíso Region declared by the Ministry of the Environment, being the habitat of the only species of pinniped endemic to Chile, the Juan Fernández fur seal (Arctocephalus philippii) [14]. It was believed that the Juan Fernández fur seal was extinct until a small group of 200 was found on the Juan Fernández islands in the 1960s [15]. The height of the flight was around 50 meters from the ground, which gave a spatial resolution of 2 cm per pixel.

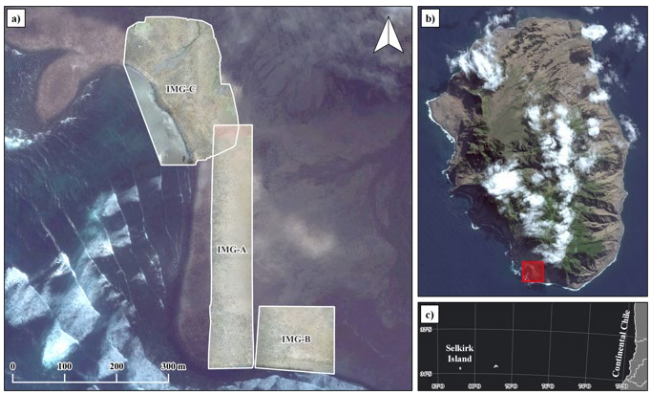

Fig. 1. Location of the Alejandro Selkirk Island.

B. Selection of ML Algorihtm

In this project, the Faster-region proposal networks (Faster-RCNN) [16] built-in the TensorFlow API [17] was adapted to design the semi-supervised model for detecting the number of fur seals on Alejandro Selkirk Island. The Region-based Convolutional Neural Networks (RCNN) [18] has three stages. The first stage is perform a selective search [19] to extract a set of object proposals (object candidate boxes). Then, according to a predefined ratio, a rescale is performed on each proposal to a fixed size image for input to a CNN model trained on a generic ImageNet (for example, ResNet[20]) to extract features. The final stage is to predict the presence of an object within each region using linear SVM classifiers. Faster-RCNN introduces the Regions of Interest (RoI) [16] to replace the selective search [18]. RoI speeds up the RCNN by enabling the algorithm simultaneously to train a detector and a bounding box regressor under the same network configurations.

III. Experimennt Design

A. Dataset

In this preliminary study, we trained and validated our model based on the seals on two smaller scale images before testing on five new smaller scale images, with 2073 fur seals in total. We manually labeled 800 fur seals for training and validation (see Figure 2) with the LabelImge tool [21]. Then, the 800 labeled data were partitioned into two sets for training and validation before testing. 700 labeled seals were used as the training set and 100 labeled seals as the validation set. All seals in the test set had to be manually labelled after they were labelled by the trained model in the test image to allow for metric calculations.

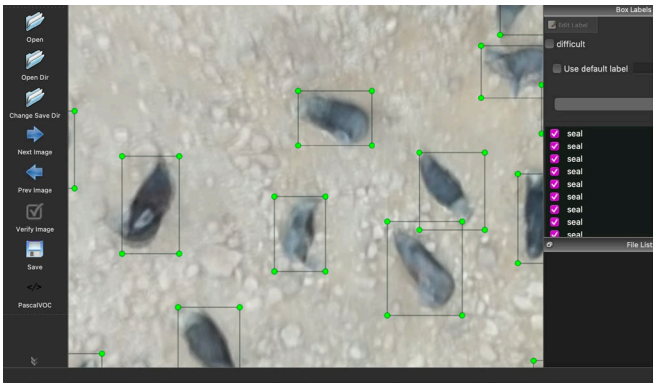

Fig. 2. Data labeling by the LabelImg Tool, where a bounding box is manually placed around seals. Such labelling occurs manually in the training and test images.

B. *Pretrainned Model*

The model used is the Faster R-CNN ResNet101 V1 1024x1024 model provided by the TensorFlow model zoo [17]. This model has been pre-trained on the Microsoft COCO-2017 dataset [22], which provides a relatively good trade-off between performance and speed.

C. *Evaluation Metrics*

To measure the object localization accuracy, the Intersection over Union (IoU) method is used to check whether the IoU between the predicted box and the ground truth box is greater than a predefined threshold, in this case, 0.5. If yes, the object will be identified as True Positive otherwise it will be identified as False Positive. The 0.5- IoU based Mean Average Precision (mAP) has become the de facto metric for object detection problems and uses the area of overlap between ground-truth (labelled) bounding boxes of objects and predicted ML bounding boxes for those same object with an overlap threshold to evaluate the accuracy of the classifier [8].

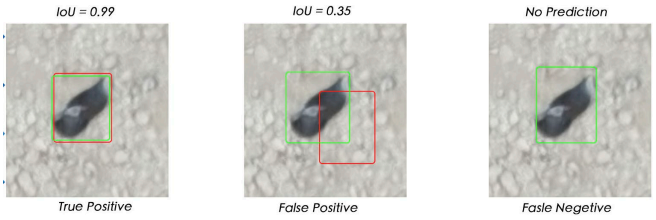

Fig. 3. Intersection over Union (IoU) is used to provide accuracy metrics between the predicted box and the ground truth box is greater than a predefined threshold. (a) Overlap greater than 0.5 is True Positive; (b) overlap less than 0.5 is False Positive; (c) no overlap is False Negative.

Equation (1) presents how the mean Average Precision is calculated based on IoU.

$$\Phi = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

D. *Training and Validation*

As noted above, in the training stage, 700 labeled seals were used as the training set and 100 labeled seals as the validation set before semi-supervised testing. Based on the IoU metrics, the training and validation mean Precision is 0.9367, indicating good training and identification of labelled seals.

Fig. 4. Loss value of the training model.

E. *Testing*

For testing, the 1,273 seals in the five test sub-images were labelled manually to compare predicted bounding boxes against ground-truth to calculate the IoU. These five sub-images from the original image (Figure 1-a, IMG-C, top left) were chosen because of their relative high clarity to ensure reliability of evaluation metrics (Figure 5).

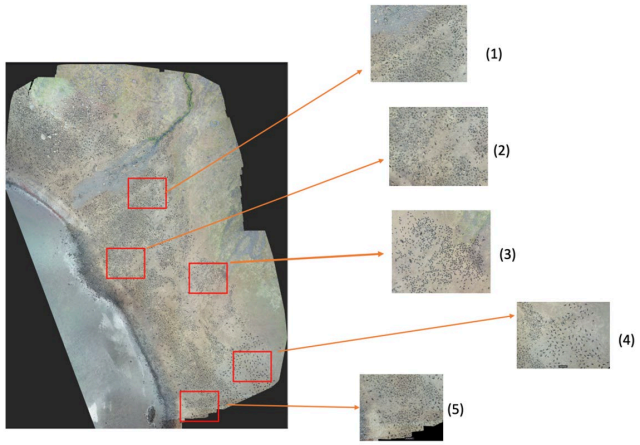

Fig. 5. Five sub-images are covering five different areas from the original image IMG-C are used for testing the model

F. *Backbone CNN*

ResNet-101 V1 1024x1024 is the backbone CNN of the TensorFlow pre-trained model. ResNet-101 V1 is a convolutional neural network that is 101 layers deep. The network has an image input size of 1024-by-1024.

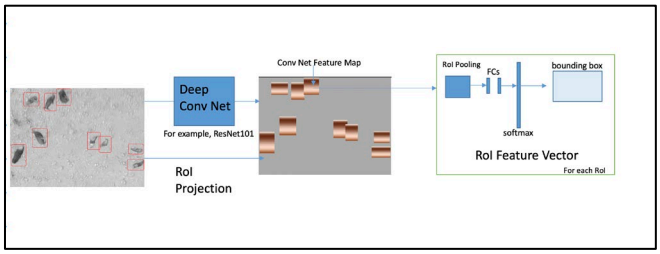

Fig. 6. The proposed Faster R-CNN model architecture.

As shown in Figure 6, the region of interest algorithm localizes the area of objects, and the backbone CNN has the task of determining whether an object is identified and counted.

IV. Result and Discussion

As noted above. the training, validation and test datasets were selected so that the images contained similarly clarity with fewer rocks and shadows. According to table 1, the True Positive, False Positive and False Negative are 1792, 281 and 518 using IoU, respectively, resulting in Recall of 0.7557 and the Overall Precision of training and testing of 0.8644.

TABLE I. Test Result

TP Number	*FP Number*	*FN Number*
1792	281	518

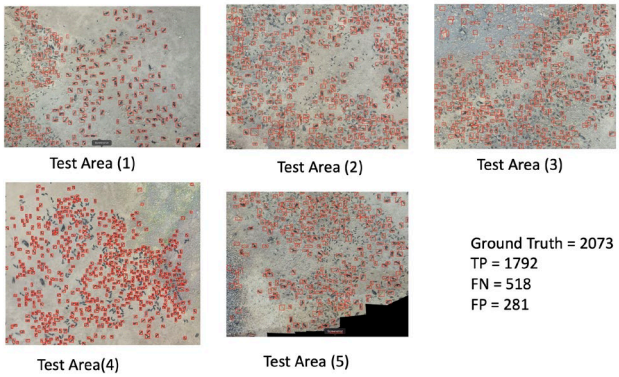

Fig. 7. Results of the test datasets and the overall precision is 0.86.

The 0.8644 Precision rate demonstrates the feasibility of our approach. However, 518 out of 2073 (24%) fur seals were not detected by the algorithm. This was due to overlapping fur seals, oddly shaped fur seals, and seals that are not distinct from the background and could not be detected properly by the algorithm (Figure 8). Precision will be dependent on drone image resolution. In this study, the photos were taken from 50 meters above ground, resulting in 2 cm per pixel resolution. It may also be because of the training and validation datasets chosen contained good fur seal samples. The model was not fine-tuned to deal with less clear samples.

Fig. 8. Example of a test result, the oddly shaped fur seals and the overlapped fur seals cannot be detected by the model (no bounding box produced by the classifier)

V. Conculusions and Futrue Work

As far as we are aware, this study implements the first semi-supervised machine learning model to count the number of fur seals from drone images. From 800 manually identified labels in the training set sub-images, testing was undertaken on another 1273 seals as they appeared and were labelled manually in the five test sub-images with 0.8644 precision but 24% false negatives. Given that there were no non-seals used for training, the high percentage of false negatives could be expected. If it is not possible to reduce this number of false negatives, counts of seals using the current architecture will need to take into account the 24% false negative rate to reach an evidenced estimate of the true number of seals in subsequent images.

These results show the feasibility of using Faster R-CNN for seal detection and counting from drone images. More work is required to determine whether such pretrained configurations can be applied successfully to other animal species or whether maximum precision can only be obtained by building specialized configurations from scratch for each species.

The next stage of our model is to fine-tune it to be able to detect fur seals with more disturbance, such as overlapping and oddly shaped instances. Experiments with different recognition thresholds will need to be tried systematically to evaluate the trade-off between precision and false negative rates. Furthermore, we may also want to extend the model to be able to identify different types of seals, for example, male, female, and puppy. While Faster-RCNN has been shown to be relatively successful in this study, other object detection models, such as YOLO[23] and RetinaNet [24], could be applied for comparison studies to find improvements in fur seal detection. However, the main problem with fur seal counting would appear to be the need to discriminate grey objects of interest against a similar shaded background of rocks. Future work may need to consider applying high contrast filters or image enhancement at the data pre-processing state.

Fully automatic recognition of animals from drone images without the need for manual labelling is the ultimate goal for animal surveys. There is little point in constructing automatic counting systems if human experts end up doing most of the manual labelling required to generate large numbers of training and test cases. The work described in this study provides an indication that manual and semi-supervised methods may need to be adopted in the near to mid future until deep learning methods coupled with high resolution images can be successfully combined in suitable architectures if high precision is to be obtained without significant human input. Another step could be to feed the outcomes of testing one by one back into the trained model to help it refine its architecture so that increased estimation accuracy is achieved as more and more test labels are manually introduced.

In summary, a semi-supervised approach involving only one object of interest has the problem of how to label all objects of interest in a large population to be used for training and testing. Our approach has shown that labelling 800 out of 2073 objects was sufficient for high precision on the remaining objects which were labelled by humans as they were identified in testing. Such an approach also resulted in many false negatives. However, by balancing out precision and false negatives through use of thresholds, it may be possible to apply the model to many other test images without further supervision and to attain an evidenced estimate if that is all that is required. Highly accurate census numbers will require a more thorough training and testing regime that will, in turn, require more human expert involvement and time.

ACKNOWLEDGMENT

This research was funded by FIPA, Luciano Hiriart-Bertrand, Victor Castillo-González, Renato Borras-Chavez, NGO-Costa Humboldt.

REFERENCES

[1] W. Jetz *et al.*, 'Essential biodiversity variables for mapping and monitoring species populations', *Nature ecology & evolution*, vol. 3, no. 4, pp. 539–551, 2019.

[2] R. R. McIntosh, R. Holmberg, and P. Dann, 'Looking without landing—using remote piloted aircraft to monitor fur seal populations without disturbance', *Frontiers in Marine Science*, vol. 5, p. 202, 2018.

[3] S. A. Wood, P. W. Robinson, D. P. Costa, and R. S. Beltran, 'Accuracy and precision of citizen scientist animal counts from drone imagery', *PloS one*, vol. 16, no. 2, p. e0244040, 2021.

[4] T. Hollings, M. Burgman, M. van Andel, M. Gilbert, T. Robinson, and A. Robinson, 'How do you find the green sheep? A critical review of the use of remotely sensed imagery to detect and count animals', *Methods in Ecology and Evolution*, vol. 9, no. 4, pp. 881–892, 2018.

[5] C. Meng *et al.*, 'IS-COUNT: Large-scale Object Counting from Satellite Images with Covariate-based Importance Sampling', in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 11, pp. 12034–12042.

[6] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, 'Drone-based object counting by spatially regularized regional proposal network', in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4145–4153.

[7] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, 'Crowdnet: A deep convolutional network for dense crowd counting', in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 640–644.

[8] Z. Zou, Z. Shi, Y. Guo, and J. Ye, 'Object detection in 20 years: A survey', *arXiv preprint arXiv:1905.05055*, 2019.

[9] J. C. van Gemert, C. R. Verschoor, P. Mettes, K. Epema, L. P. Koh, and S. Wich, 'Nature conservation drones for automatic localization and counting of animals', in *European Conference on Computer Vision*, 2014, pp. 255–270.

[10] B. Kellenberger, M. Volpi, and D. Tuia, 'Fast animal detection in UAV images using convolutional neural networks', in *2017 IEEE international geoscience and remote sensing symposium (IGARSS)*, 2017, pp. 866–869.

[11] B. Xu *et al.*, 'Automated cattle counting using Mask R-CNN in quadcopter vision system', *Computers and Electronics in Agriculture*, vol. 171, p. 105300, 2020.

[12] A. M. Dujon *et al.*, 'Machine learning to detect marine animals in UAV imagery: Effect of morphology, spacing, behaviour and habitat', *Remote Sensing in Ecology and Conservation*, vol. 7, no. 3, pp. 341–354, 2021.

[13] J. G. A. Barbedo, L. V. Koenigkan, T. T. Santos, and P. M. Santos, 'A study on the detection of cattle in UAV images using deep learning', *Sensors*, vol. 19, no. 24, p. 5436, 2019.

[14] J. Vanhulst, 'MENACES ET PERSPECTIVES POUR LA PRÉSERVATION DE LA BIODIVERSITÉ DE L'ARCHIPEL JUAN FERNÁNDEZ (CHILI)'.

[15] P. A. Folkens and R. R. Reeves, *Guide to marine mammals of the world*. AA Knopf, 2002.

[16] S. Ren, K. He, R. Girshick, and J. Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks', *Advances in neural information processing systems*, vol. 28, 2015.

[17] Martín Abadi *et al.*, 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems'. 2015. [Online]. Available: https://www.tensorflow.org/

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, 'Region-based convolutional networks for accurate object detection and segmentation', *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[19] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, 'Segmentation as selective search for object recognition', in *2011 international conference on computer vision*, 2011, pp. 1879–1886.

[20] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Tzutalin, 'LabelImg'. 2015. [Online]. Available: https://github.com/tzutalin/labelImg

[22] T.-Y. Lin *et al.*, 'Microsoft COCO: Common Objects in Context'. 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You only look once: Unified, real-time object detection', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, 'Focal loss for dense object detection', in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.