Beluga whale detection from sliced aerial remote sensing images using object detection pipelines

Muhammed Patel University of Waterloo
Xinwei Chen University of Waterloo
Neil C. Brubacher University of Waterloo
Linlin Xu University of Waterloo
David A. Clausi University of Waterloo
Email: {m32patel, xinwei.chen, ncbrubacher, l44xu, dclausi}@uwaterloo.ca

Abstract

The emergence of high-resolution remote sensing imagery greatly facilitates the activities related to conservation biology, including whale counting. As manual annotating is laborious and subjected to human-induced bias, it is necessary to introduce automatic approaches for whale detection from the large remote sensing dataset based on machine learning-based techniques. In this paper, we implement two deep neural network-based object detection models (i.e., RetinaNet and faster RCNN) to detect the presence of whale in aerial remote sensing images obtained from a survey conducted on Cumberland Sound Bay, Nunavut in 2014. To tackle the difficulties in effective detection caused by the sparse occurrence of whales in the large image, an image-slicing approach is adopted to increase the ratio between the size of whale sample bounding boxes and the input image of the model. Testing results on both viusal evaluation and numerical analysis show that compared to downsample on the original image directly, the proposed image-slicing approach boosts the detection accuracy significantly.

1 Introduction

Monitoring whale population in remote areas is important for the preservation and sustainable hunt of this population [1, 2]. Remote sensing (RS) imagery obtained from various platforms, such as satellites, aircraft, and drones, can efficiently monitor vast areas with rich spatial-spectral information for target discrimination. It has become an essential tool for the detection and localization of whale presence over vast ocean areas in various whale protection and law enforcement applications.

Over the past few years, multiple deep learning-based approaches have been proposed for automatic whale detection using RS images [3, 4]. Although they are much more efficient than manual labeling, challenges still exist in the accuracy and generality of these methods. Due to the innate rareness of whale occurrence, as well as the tiny scale of whale samples relative to large RS images, conventional DL-based image classification and object detection models face key challenges in performance and efficiency in detecting such rare, small targets. A method that can perform well in tiny object detection tasks (i.e. whale detection in aerial RS images) should be investigated.

In this paper, a framework designed for performing large scale object detection called slicing aided hyper inference (SAHI) [5] is combined with object detection baselines to further improve the detection accuracy of whales in aerial RS imagery. Section 2 gives an introduction of the dataset used in this study. The overall procedures of the proposed pipeline is illustrated in Section 3. Finally, experimental results are presented and discussed in Section 4.

2 Data Overview

The RS images used in this study were obtained from aerial surveys of the Cumberland Sound Bay, Nunavut on August, 2014, as indicated in Fig. 1. The belugas population in this area, known as the Cumberland Sound belugas, are fairly isolated and genetically distinct from other beluga populations. The aircraft that conducted the survey was equipped with a 36.15MP Nikon D810 camera to take aerial images geo-coded with an onboard GPS unit at a target altitude of 610m, with an interval of approximately 4 seconds. The size of each image is 4912×7360 pixels, and the pixel spacing is around 10 cm. An example image taken in the survey is presented in Fig. 2, with enhancement showing the presence of belugas. Further details about the survey and the dataset can be found in [4]. Among all the images collected in the 2014 survey, a total of 467 images that contain whale samples are included for model training and evaluation. 80% of those images are randomly selected for model training (300 images) and validation (75 images), while the remaining 20% (92 images) are used for testing.

Fig. 1: The location of aerial survey in Cumberland Sound Bay, Nunavut (outlined in yellow).

3 Methodology

To ensure the accuracy of the model, whales that are present in the images should be labeled precisely. The locations of the whales in the images were annotated by expert annotators from Department of Fisheries and Oceans, Canada. We take the pixel position of whales and assign a bounding box to each whale as indicated in Fig. 3. These annotations are then stored in the popular MSCOCO dataset format [6]. An open-source framework for object detection named MMdetection [7] is used to build our models based on deep neural networks. In this study, two deep neural networks, i.e., RetinaNet [8] and FasterRCNN [9], are adopted to build the detection models due to their better performance compared to other types of networks. To reduce the computational cost, an initial approach to model training involved resizing the images into 1333×800 pixels. However, the size of bounding boxes were found to be prohibitively small compared to the image size after downscaling, which may negatively affect the model performance. Hence, to improve the detection accuracy, an image slicing approach designed for object detection proposed in [5] is applied to slice each original image into patches with 768×768 pixels with an overlap ratio of 0.2 between those patches. Those patches are then input into the networks for training. In this way, the original resolution is retained and the ratio between bounding box size and input image size becomes larger. Those two approaches are both implemented for comparison. For both approaches, we use the SGD optimizer with a linear learning rate decay policy. The models are trained until convergence which takes approximately 3 hours on a single NVIDIA V100 Tensor Core with 32GB memory.

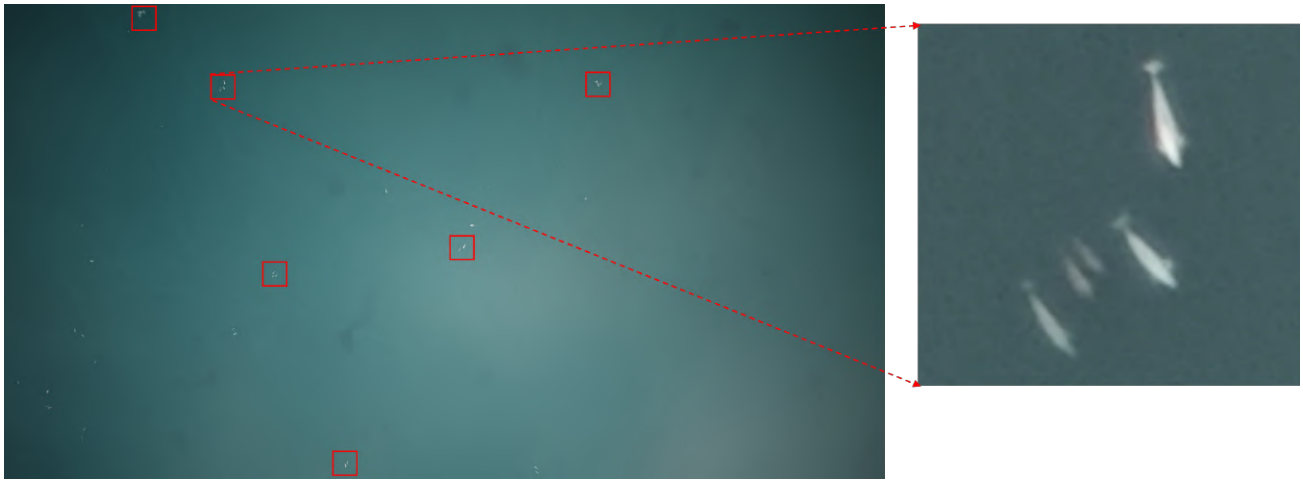

Fig. 2: Left: an example of an aerial image obtained from the survey. Regions with whales are outlined in red. Right: an enlarged example of a region in the image on the left with whales.

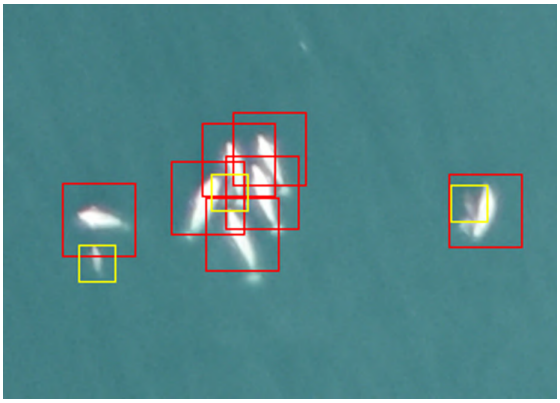

Fig. 3: An example of annotations of whales, with 40×40 bounding boxes for the adults (outlined in red) and 20×20 bounding boxes for the cubs (outlined in yellow).

4 Discussion

Fig. 4 shows the prediction results on a testing image using two different networks with the two image preprocessing schemes mentioned above. In can be observed that the RetinaNet-based model with the image downsampling scheme is unable to detect any whales (top right), whereas the one trained with sliced images is able to detect the presence of whales (bottom left). Nevertheless, it also produces a false positive prediction on the rocks. In comparison, the Faster RCNN-based model trained on the sliced images could predict all the whales correctly without producing false positives.

To evaluate the testing results numerically, the MSCOCO evaluation protocol [6] has been adopted for the evaluation and more details can be found in [1]. In terms of mAP, with an IOU threshold of 0.75, the RetinaNet-based model with downsampling scheme has an overall mAP of 0.06, whereas the one trained on sliced patches achieves an overall mAP of 0.20. Thus, the mAP improves by more than 3 times when using the slicing approach. Besides, faster RCNN trained with sliced images further improves the performance by around 40% with an maP of 0.28, which is 0.08 higher in comparison to the RetinaNet-based model. Error analysis of all the detectors can be found in Fig. 5. In conclusion, slicing the large image into patches for model training effectively solves the tiny object issue and improves detection rates significantly. As for model selection, it is found that faster RCNN out performs RetinaNet with less false positives. Future works will focus on further improving the accuracy and robustness of the detection model by incorporating more data for training and implementing on more state of the art network models.

[1] https://cocodataset.org/detection-eval

Acknowledgments

The authors would like to thank the Canadian National Engineering Science and Research Council (NSERC) and Whale Seeker for supporting this research.

References

[1] M. Marcoux and M. O. Hammill, *Model estimates of Cumberland Sound beluga (Delphinapterus leucas) population size and total allowable removals*. Canadian Science Advisory Secretariat, 2016.

[2] H. C. Cubaynes and P. T. Fretwell, "Whales from space dataset, an annotated satellite image dataset of whales for training machine learning models," *Sci. Data*, vol. 9, no. 1, pp. 1–8, 2022.

[3] E. Guirado, S. Tabik, M. L. Rivas, D. Alcaraz-Segura, and F. Herrera, "Whale counting in satellite and aerial images with deep learning," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, 2019.

[4] P. Q. Lee, K. Radhakrishnan, D. A. Clausi, K. A. Scott, L. Xu, and M. Marcoux, "Beluga whale detection in the cumberland sound bay using convolutional neural networks," *Can. J. Remote Sens.*, vol. 47, no. 2, pp. 276–294, 2021.

[5] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," *2022 IEEE Intl. Conf. Image Proc. (ICIP)*, pp. 966–970, 2022.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Comput Vis ECCV*. Springer, 2014, pp. 740–755.

[7] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv Neural Inf Process Syst.*, vol. 28, 2015.

Fig. 4: Top left: the ground truth labels of whales outlined in green in a sample image. Top right: predictions of the sample image using RetinaNet without slicing. The RetinaNet is unable to detect any whales. Bottom left: predictions of RetinaNet trained with sliced images outlined in red with confidence scores. The model could detect the GT correctly but it also produces a false positive near the rock. Bottom right: predictions of FasterRCNN trained with sliced images. The model is able to detect the GT correctly as well as eliminate the false positive that RetinaNet produced.

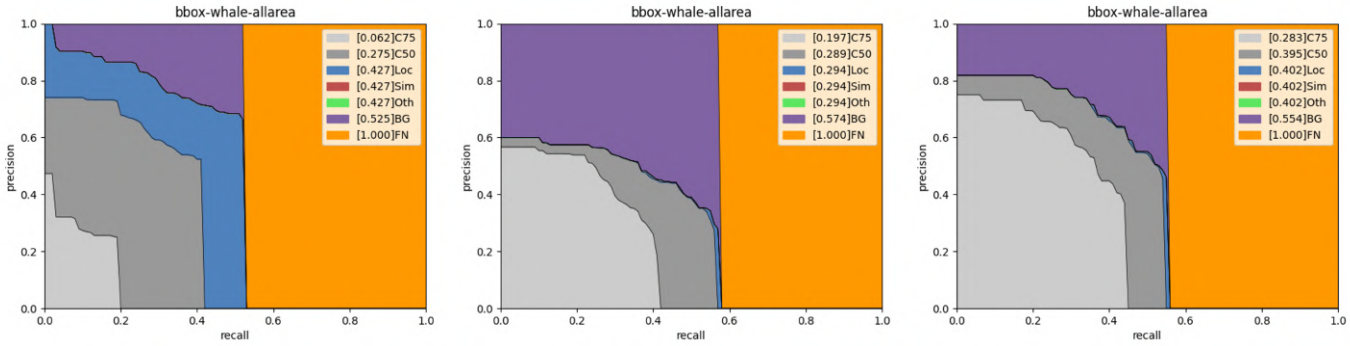

Fig. 5: Left: Error analysis curve for RetinaNet without slicing. Middle: error analysis curve for RetinaNet trained with sliced images. Right: error analysis curve for Faster RCNN trained on sliced images.