Remote Sensing in Ecology and Conservation
Open Access
ZSL

RESEARCH ARTICLE

Gray whale detection in satellite imagery using deep learning

Katherine M. Green[1,2], Mala K. Virdee[2], Hannah C. Cubaynes[1], Angelica I. Aviles-Rivero[3], Peter T. Fretwell[1], Patrick C. Gray[4], David W. Johnston[4], Carola-Bibiane Schönlieb[3], Leigh G. Torres[5] & Jennifer A. Jackson[1]

[1]British Antarctic Survey, Cambridge, UK
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[3]Department of Applied Mathematics Theoretical Physics, University of Cambridge, Cambridge, UK
[4]Marine Laboratory, Duke University, Durham, North Carolina, USA
[5]Marine Mammal Institute, Oregon State University, Corvallis, OR, USA

Keywords
CNN, *Eschrichtius robustus*, gray whale, machine learning, remote sensing, VHR satellite imagery

Correspondence
Katherine M. Green, British Antarctic Survey, Cambridge, UK. Tel.: 07738227476. E-mail: kmmg2@cam.ac.uk

Funding Information
The study was funded by British Antarctic Survey and UK Research and Innovation (EP/S022961/1).

Editor: Nathalie Pettorelli
Associate Editor: Nicola Quick

Received: 9 December 2022; Revised: 2 May 2023; Accepted: 25 May 2023

doi: 10.1002/rse2.352

Remote Sensing in Ecology and Conservation 2023;**9** (6):829–840

Abstract

The combination of very high resolution (VHR) satellite remote sensing imagery and deep learning via convolutional neural networks provides opportunities to improve global whale population surveys through increasing efficiency and spatial coverage. Many whale species are recovering from commercial whaling and face multiple anthropogenic threats. Regular, accurate population surveys are therefore of high importance for conservation efforts. In this study, a state-of-the-art object detection model (YOLOv5) was trained to detect gray whales (*Eschrichtius robustus*) in VHR satellite images, using training data derived from satellite images spanning different sea states in a key breeding habitat, as well as aerial imagery collected by unoccupied aircraft systems. Varying combinations of aerial and satellite imagery were incorporated into the training set. Mean average precision, whale precision, and recall ranged from 0.823 to 0.922, 0.800 to 0.939, and 0.843 to 0.889, respectively, across eight experiments. The results imply that including aerial imagery in the training data did not substantially impact model performance, and therefore, expansion of representative satellite datasets should be prioritized. The accuracy of the results on real-world data, along with short training times, indicates the potential of using this method to automate whale detection for population surveys.

Introduction

Whales are ecosystem ‘engineers’, generating and transporting large quantities of nutrients through the water column, and sequestering carbon in the ocean (Roman et al., 2014; Savoca et al., 2021). However, whale populations were hunted to near-extinction during the 19th and 20th centuries (Reeves & Smith, 2010; Rocha Jr. et al., 2015), and contemporary populations are still recovering, with many species still classified as endangered (Clapham, 2016). In the case of gray whales (*Eschrichtius robustus*), the focus of this study, global populations are believed to have been historically three to five times larger than current populations, indicating the potential for further recovery (Alter et al., 2007). As well as recovering from the impacts of commercial whaling, whales are faced with anthropogenic threats such as ship strikes, entanglement, pollution, and climate change (de Vos et al., 2016; Nicol et al., 2020; Silber et al., 2020). This is true for gray whales since their range is strongly coincident with human development (Scordino et al., 2018; Silber et al., 2020).

Monitoring whale abundance and distribution is key to understanding how these threats can impact species' recovery. Whale population surveys are traditionally conducted via visual observations from aircraft, ships, or land (Hammond et al., 2021; Noad et al., 2019), either through line-transect sampling or through photographic capturerecapture techniques. These surveys provide important insights into whale density and abundance patterns, but their spatial and temporal coverage is patchy due to the costs and logistical limitations of ship-based surveys. Consequently, in many areas it is difficult to maintain the monitoring effort required to accurately measure important conservation parameters such as abundance and habitat overlap with areas of concentrated human activities (Kaschner et al., 2012).

Remote sensing, in particular very high resolution (VHR) satellite imagery, provides vast amounts of data that can increase the coverage of whale surveys and fill data gaps (Charry et al., 2021; Corrêa et al., 2022; Cubaynes et al., 2019; Fretwell et al., 2014; Hodul et al., 2023), particularly in remote locations where traditional coverage is sparse (Bamford et al., 2020; Charry et al., 2021). In comparison with land-based species, the detection of whales in remote sensing data is more challenging since even if a whale is present in the image, it must be sufficiently close to the surface and in relatively calm conditions to be detected. Therefore, collecting accurate ground-truth data is difficult, and in order to expand these methods to calculate abundance estimates, accurate methods of accounting for whales that may have been missed will need to be devised. Detecting whales via VHR satellite imagery is in its infancy, but is already being applied to multiple studies of species distribution (e.g., Bamford et al., 2020; Charry et al., 2021; Corrêa et al., 2022). For more advanced goals such as measuring whale abundance, more work is required to investigate how whale detection probabilities vary by species, whale depth, body position (Cubaynes et al., 2020), and sea state (Bamford et al., 2020). Automatic tools for detecting whales in images more efficiently will help to advance all of these studies.

Object detection is a task within the field of computer vision: a bounding box is specified around each object of interest and labelled by the class it belongs to (Liu et al., 2020; Zou et al., 2023). Machine learning is typically leveraged for this task, and most state-of-the-art approaches use convolutional neural networks (CNNs). CNNs ingest multidimensional arrays, such as imagery or video, and transform that input data into higher-level features that, in theory, represent more abstracted and semantically meaningful information that can be used to classify the data. Specifically, CNNs extract features from imagery data with a series of layers that convolve imagery elements with moving kernels that have learnable weights (Dumoulin & Visin, 2018). Through multiple convolutional layers, where the output of one is passed on to the other, higher-level features are derived that represent aspects of the imagery important to the task at hand (LeCun et al., 2015). The filter weights are learned by providing the network examples of the task being done successfully, in this case bounding boxes and classes of the desired objects. Using these examples, the network minimizes a loss function and iteratively improves the network weights and thus improves at the specified task (Khan et al., 2020). The first layers of a CNN typically create maps of features such as edges and curves. The feature maps from deeper layers are more abstract and combine the previous layer's feature maps; in our case, this may indicate flippers, flukes, or body shapes.

Developments in high-performance computing systems and network architectures have increased the capabilities of CNNs for object detection. For example, the YOLO (You Only Look Once) object detection framework (Redmon et al., 2016) produced state-of-the-art results in terms of accuracy and speed. There have been many iterations of the YOLO framework since its initiation, with continual improvements (Bochkovskiy et al., 2020). Increases in speed and accuracy of object detection using deep learning now make it a good candidate for remote sensing tasks (Li et al., 2020). These developments are beneficial for conservation, as the combination of artificial intelligence and remote sensing has the potential to increase the scale and speed of whale detection in satellite imagery and reduce manual input through automation (Lamba et al., 2019). In addition to the potential for whale population monitoring, these techniques, combined with the major increase in available VHR satellite imagery (Maxar, 2022; Planet Labs, 2022; UP42 GmbH, 2019), open up new possibilities for whale conservation, such as rapid response to ship collision risks.

Object detection with deep learning has been successful in identifying whales in VHR satellite imagery (Houegnigan et al., 2022; Kapoor et al., 2023; Khan et al., 2023). Training of CNNs has been previously performed using aerial imagery or a mixture of VHR satellite and aerial imagery (Borowicz et al., 2019; Guirado et al., 2019), due to the lack of labelled VHR satellite imagery, which takes time to acquire and must be manually requested and annotated by users (Cubaynes & Fretwell, 2022; Höschle et al., 2021). Both studies used satellite images collected in sheltered conditions, providing optimum data for evaluating model performance under similar conditions. However, whales occupy offshore locations where conditions render whales less detectable visually (Bortolotto et al., 2016; Marsh & Sinclair, 1989; Panigada et al., 2011). Therefore, it is necessary to investigate how different combinations of satellite and aerial images influence model performance in real-world conditions.

Although deep learning for image recognition has been previously applied for whale detection (e.g., Borowicz et al., 2019; Guirado et al., 2019), the real-world application and contributing factors when using different types of imagery remain largely unexplored. Here, we systematically investigate performance of the YOLOv5 deep learning model (Jocher et al., 2021; Redmon et al., 2016), to automatically detect gray whales on VHR satellite images in sheltered and more exposed areas. The gray whale, the target species of this study, is a baleen whale, which is restricted to the North Pacific, following centuries of hunting in the North Atlantic and Pacific Oceans (Alter et al., 2007). In summer, the species feeds in shallow subpolar and polar waters (the north-west Bering, southern Chukchi, and Beaufort Seas), as well as coastal waters of North America between California and Alaska (Caretta et al., 2017). In winter, gray whales mainly calve and congregate around lagoons in Baja California, Mexico (Würsig et al., 2008). We chose the gray whale as it represents an ideal candidate for which remote monitoring with VHR satellites may be helpful: During the calving season, it is often located in calm lagoons where detection is feasible and so there is a need to monitor shifts in distribution and identify overlaps with risk factors in future, including for example, ship strike risks while on migration. Specifically, we address the following questions: How do different types of imagery affect model performance? What is the most important training imagery type for automatic whale detection?

Materials and Methods

Data description and preprocessing

Imagery description

Four VHR satellite images from Baja California Peninsula, Mexico, were used for this analysis (Fig. 1), from Laguna Ojo de Liebre (areas 1 and 2), Laguna San Ignacio (area 3), and the mouth of Laguna San Ignacio (area 4). The first three images were selected due to the absence of clouds, calm sea state, and abundance of gray whales. The fourth image shows a slightly rougher sea state, so was included to investigate the model's ability to detect whales in a less sheltered area. Satellite images (Table 1) were provided as two image types: a multispectral color image (resolution 1.24 m) and a panchromatic grayscale (resolution 0.31 m).

Guirado et al. (2019) previously incorporated aerial imagery of varying resolutions showing differing sea states and positions of whales into the training dataset, to

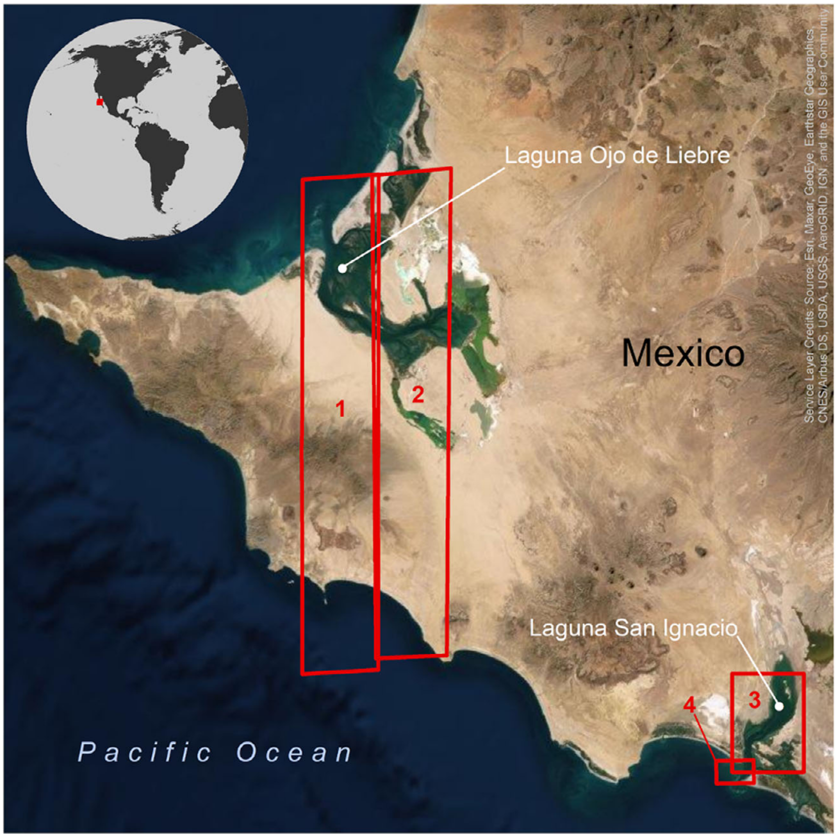
Laguna Ojo de Liebre
Mexico
1
2
Laguna San Ignacio
3
4
Pacific Ocean
Service Layer Credits: Source: Esri, Maxar, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Figure 1. Map of the west coast of Mexico illustrating the locations from which the four satellite images used in the study were taken.

Table 1. Specifications for each satellite image used in this study.

Area ID	Location	Satellite	Product type	Date	Catalog ID
1	Laguna Ojo de Liebre – West	WorldView-2	Standard 2A	03.01.2013	103001001E12C300
2	Laguna Ojo de Liebre – East	WorldView-2	Standard 2A	25.01.2015	103001003C8B2300
3	Laguna San Ignacio	WorldView-3	Standard 2A	20.02.2017	104001002959ED00
4	Mouth of Laguna San Ignacio	GeoEye-1	Standard 2A	28.02.2009	1050410001FEA500

Area ID refers to area number provided in Figure 1.

improve the CNN's ability to detect whales in satellite images. We also included aerial images in our training dataset. Downsampling has previously been used on aerial images to convert them to the same resolution as the satellite images for model training (Borowicz et al., 2019). Properly transforming an aerial image to match a satellite image requires downsampling to the spatial resolution of the satellite image, as well as accounting for natural atmospheric distortion and the algorithm for atmospheric correction of the image. This transformation is difficult to perform in a standardized manner and thus was not done, and we investigate whether full-resolution aerial images improve model performance.

We used two 0.005 m resolution aerial image datasets from Oregon waters to train the CNN. Firstly, we incorporated stills from video footage of gray whales, recorded by a DJI Phantom 3 Pro or 4 Advanced UAS during a previous study (Torres et al., 2018). Using VLC Media Player, one scene was extracted per second from these videos. The images contained whales in a variety of positions and scenarios, such as diving or with visible spray from blowing, and were taken across a variety of sea colors and states. Secondly, we used a collection of 89 aerial images of gray whales shown in 'ideal' states, fully visible from directly above (Burnett et al., 2019), using cameras attached to the same UAS (Fig. 2).

Image preprocessing

All satellite images were pan-sharpened [combining the multispectral image with the panchromatic image for each area, using the Gram-Schmidt algorithm in ArcGIS 10.8 (ESRI)], producing multispectral images of 0.31 m resolution. These images were systematically scanned at a scale of 1:2000 by two observers and gray whales and boats labelled. Whales in each of these images were assigned a certainty of 'definite', 'probable', or 'possible' (Cubaynes et al., 2019). To ensure only confident whale identifications were present in the training data, only samples labelled as 'definite' or 'probable' by either observer were included in the whales class in the data. Labels were then merged to form a single dataset. Duplicate samples, where the same whale was labelled by both observers, were merged to single points using a buffer of 3 m to define a duplicate. This buffer was required as the two observers often did not label each whale in the same location, for

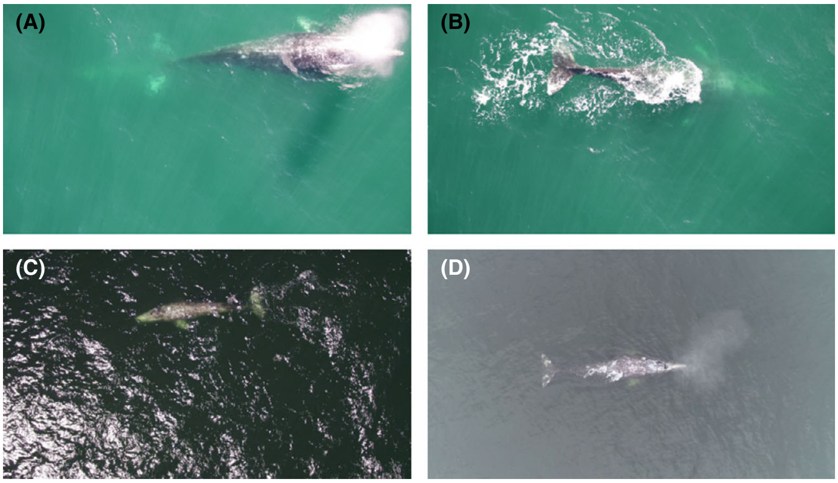

Figure 2. Aerial imagery samples. (A) A whale partly obscured by spray from blowing, (B) a diving whale, (C) a different sea state to demonstrate the variability in the aerial dataset, and (D) a still image sample demonstrating an 'ideal' whale image where all the features are visible from directly overhead. Images collected under NOAA/NMFS permit #16111.

example, one label may be closer to the tail and the other the center. The resulting images were manually scanned to ensure there were no duplicate points remaining in the set. Bounding boxes were added around each point labelling a whale or a boat using a 10 m buffer to ensure the inclusion of the whole object.

The pan-sharpened images were converted from four-band images (red, green, black, and near-infrared), with 16-bit pixel depth to three-band RGB images with 8-bit pixel depth, as required by YOLOv5. To help with stability of training, the 8-bit values were normalized between 0 and 255 and scaled using the 98th percentile of the 16-bit pixel values. This ensured that the image pixel values were distributed over the full range of possible values to aid in whale detection, which tend to have pixel values toward the lower end of the spectrum.

The normalized images were cropped into 512 × 512-pixel tiles centered around the whale and boat bounding boxes to ensure sufficient context was included (Fig. 3).

For the aerial imagery, LabelMe software was used (Wada, 2021) to add bounding boxes around whales. This resulted in 1019 images of whales identified by video (1124 whale instances and 53 unique animals), and 78 images containing 89 individual whales identified on still images. In computer vision tasks, particularly those with small datasets or high-class imbalance, augmentations are commonly used on input imagery to increase the amount of training data (Shorten & Khoshgoftaar, 2019). Two augmentation regimes were applied to the satellite dataset, each randomly generating one augmentation per whale tile and eight per boat tile to address the class imbalance in the dataset. The first augmentation regime included geometric transformations only, applying a flip in the horizontal axis, a flip in the vertical axis, or a rotation between −180 and 180°. The second regime also included color space transformations; blurring, altering brightness or contrast, and adding Gaussian noise (Fig. 4).

Deep learning model

YOLOv5 architecture

There are multiple versions of the YOLOv5 architecture (Jocher et al., 2021), which vary in size and depth. In this study, YOLOv5s was selected due to its speed and successful use in similar tasks (Chen et al., 2021). This speed may be beneficial in future for the automation of the workflow as the detection of whales in large satellite images is time and computationally intensive.

Transfer learning

Transfer learning is commonly employed in deep learning tasks to prevent models from overfitting on small training datasets and improve their ability to generalize (Yosinski et al., 2014). In this study, the model was pretrained on the MS-COCO dataset (Lin et al., 2014), an image dataset containing over 300 000 images of everyday objects with bounding boxes and class labels. While whales are not in the COCO dataset, the features learnt by the pretrained network are helpful to prevent the model overfitting when trained on the whale dataset. Many of the most fundamental features such as edges, curves, and color gradients are expected to be the same no matter the final class. Previously, Gray et al. (2019) successfully used this approach when classifying aerial images of whales with a small training dataset. Here, we used a multi-stage transfer learning approach, whereby a model pretrained on MS-COCO was subsequently trained on the entire aerial imagery dataset before training on the satellite imagery.

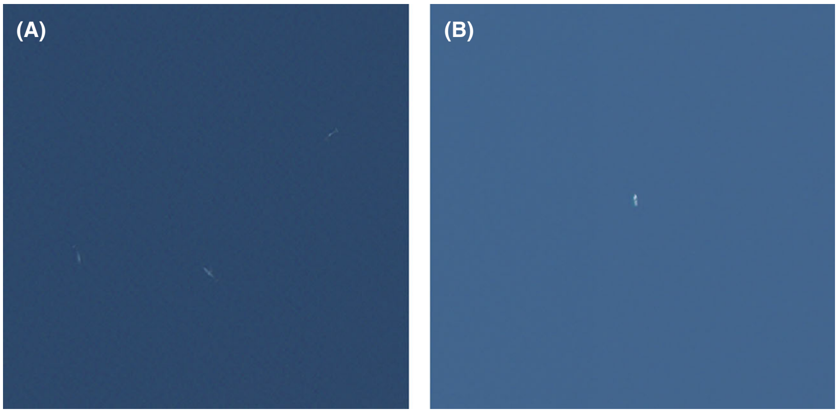

Figure 3. Two satellite tiles from the training set. (A) A tile containing three whales and (B) A tile containing a small boat. Satellite image © 2022 Maxar Technologies.

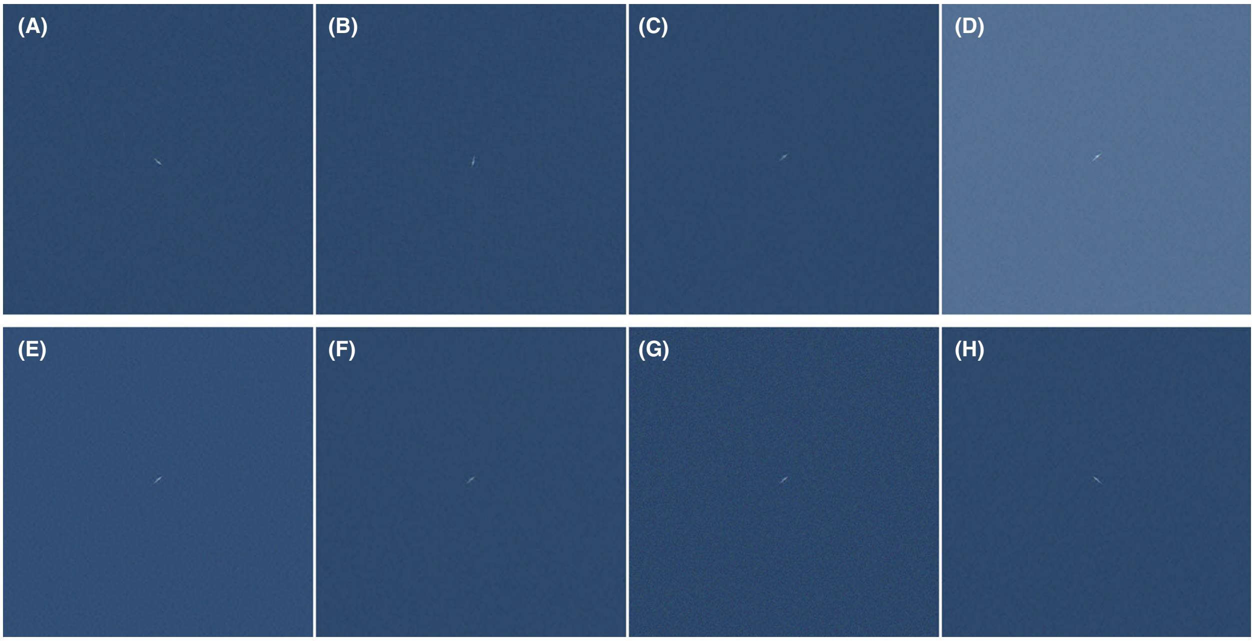

Figure 4. Example satellite tile demonstrating the effects of various augmentations. (A) Vertical flip, (B) rotation, (C) blur, (D) brightness, (E) contrast, (F) Gaussian blur, (G) Gaussian noise, and (H) horizontal flip. Satellite image © 2022 Maxar Technologies.

Training scheme

Yolov5s CNNs were trained on 10 training sets containing different combinations of satellite and aerial imagery (Table 2). For the first eight implementations, all satellite images were processed and cropped to produce 313 tiles, with 503 whale instances and 103 boat instances. These tiles were split using a 70:15:15 ratio into training, validation, and test sets, with each image contributing roughly equally to each set. This resulted in a total of 344 whale instances and 65 boat instances (training set), 69 and 21 (validation set), and 90 and 17 (test set). The baseline (implementation 1) was trained on the satellite data only with no additional augmentations. Augmentation regimes increased the number of satellite whale tiles and boat tiles by two and eight times respectively for implementations 2, 3, and 7.

To examine how results compared when only ‘ideal’ data were used to train the CNN, we ran a sensitivity test using the best image of the four, with no rough water areas (area 2 in Fig. 1), to generate the training and validation sets (198 whales and three boats in the training

Table 2. Results of implementations on the various test sets, where the best results for each metric from implementations 1–8 are shown in bold.

Training data description	Test data description	mAP@0.5	Whale precision	Whale recall	Count deviation
1. Satellite imagery only	Full satellite test set	0.823	0.889	**0.889**	0.222
2. Satellite imagery + geometric augmentations		**0.922**	0.899	**0.889**	0.211
3. Satellite imagery + full augmentations		0.904	0.909	**0.889**	**0.189**
4. Satellite + drone video imagery		0.873	0.814	0.878	0.311
5. Satellite + drone still imagery		0.839	0.898	0.878	0.211
6. Satellite + all drone imagery		0.840	0.800	**0.889**	0.322
7. Satellite + full augmentations + all drone imagery		0.915	0.832	0.878	0.300
8. Transfer learning with aerial pretraining		0.884	**0.939**	0.843	**0.189**
9. Reduced ideal satellite imagery only (image 2)	Reduced ideal satellite test set (image 2)	0.965	0.977	0.843	0.176
10. Reduced ideal satellite imagery only (image 2)	Reduced challenging satellite test set (images 1, 3, and 4)	0.330	0.600	0.625	0.795

Precision and recall for the boat class are shown in the Supplementary Information.

set, 45 and seven in the validation set), with a test set derived: (i) from the same image (51 whales, three boats); (ii) from the other three satellite images (39 whales, 14 boats).

Each implementation was trained on a single Tesla V100 32GB GPU for 300 epochs (batch size = 32). To prevent overfitting, the validation set was used to select the best weights for evaluation on the test set in each implementation. For the transfer learning approach (implementation 8), the model was trained on aerial imagery (150 epochs) and parameters were updated by training on the satellite imagery for another 150 epochs. Recommended hyperparameters (Jocher et al., 2021) were used throughout.

Evaluation protocol

Following standard deep learning protocols model performance was evaluated using three metrics, plus a fourth task-specific metric, calculated on the test set. Model precision (1) measures proportion of correct detections where TP (true positive) is the number of correct predictions, and FP (false positive) is the number of predictions made incorrectly.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Model recall (2) measures the model's ability to detect all possible objects where FN (false negative) is the number of labelled objects that the model fails to detect.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

The confidence threshold used to calculate precision and recall was 0.25. Therefore, any model predictions with a confidence above this threshold are included in the results. This threshold was chosen to balance the trade-off between minimizing FNs and FPs.

The third metric used to assess model performance was the mean average precision (mAP). The average precision was calculated for each class from the area under the precision-recall curve at different intersection over union (IoU) thresholds. IoU is defined as the intersection between the labelled bounded box and the model's output bounding box divided by the union of those two bounding boxes. The IoU threshold is the minimum fraction by which the predicted bounding box and ground-truth bounding box need to overlap for the prediction to be considered a true positive. The mAP is the mean of the average precision scores across all classes. Here, the metric reported was mAP@0.5, that is, the IoU threshold was set to 0.5. This threshold is widely used in other studies of object detection and instance segmentation (He et al., 2018).

The count deviation (3) (Rodofili et al., 2022), on the whale class, was calculated to provide a measure of the cumulative mistakes made by the model as a fraction of the total number of samples.

$$\text{Count deviation} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN}} \tag{3}$$

Results

Summary metrics are shown in Table 2. The results from the sensitivity test show high performance in ideal sea states (implementation 9, precision 0.977) but an inability to generalize to more challenging images (implementation 10, precision 0.600). Across implementations 1–8 the highest mAP@0.5 on the test set was 0.922 by implementation 2, but the performance across the other metrics was not an improvement over baseline implementation 1. The highest precision on the whale class was 0.939 (implementation 8), and the highest recall of 0.889 was achieved by implementations 1, 2, 3, and 6. The count deviation values agree with the other calculated metrics with the best values achieved in implementations 3 and 8. Implementation 3 (satellite images plus full augmentations) showed the best performance across all metrics (Table 2; Fig. 5).

On average, across implementations 1–8, 13% of predicted whales were found to be false positives. However, the majority of these errors arose from areas of sea being misidentified as whales (e.g., Fig. 6) with few boats misidentified as whales (0.1%). False-negative predictions, where labelled whales were missed, resulted in an average of 11% of whale identifications missed altogether and 0.6% mislabelled as boats across all implementations.

Discussion

Here, we use CNNs to identify whales from space using the largest dataset to date (503 identified gray whales), showing the importance of environmental context when automating whale identification in satellite images. Using imagery from a sheltered location with a calm sea state yielded high model performance and precision when applied to animals in the same area (> 0.95). Model training using satellite images from a mixture of sheltered and exposed locations generated lower model performance and whale precision (> 0.90) but performed substantially better than when a model trained in a sheltered location was tested on imagery from more exposed areas (whale precision = 0.60). These results highlight the need for CNNs to incorporate satellite images from the full range of environmental contexts in the training and validation stage.

In our study, image augmentations were the most important feature helping to improve model performance.

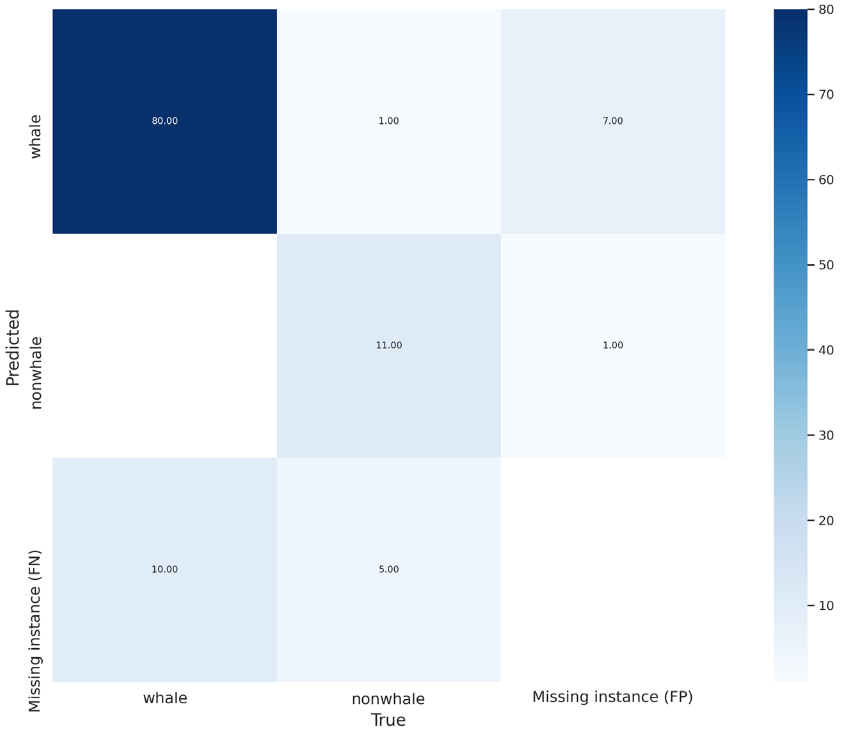

Figure 5. Confusion matrix from test predictions of implementation 3, where the test set contains 90 whales. The 'non-whale' class contains boat classifications. The predicted labels (*y*-axis) are plotted against the ground-truth labels (*x*-axis) for the whale and boat classes. The major diagonal is the correct detection and any falling off the major diagonal is incorrect. For example, if the model predicts a wave as a whale this would be shown at the intersection of the FP column and whale row.

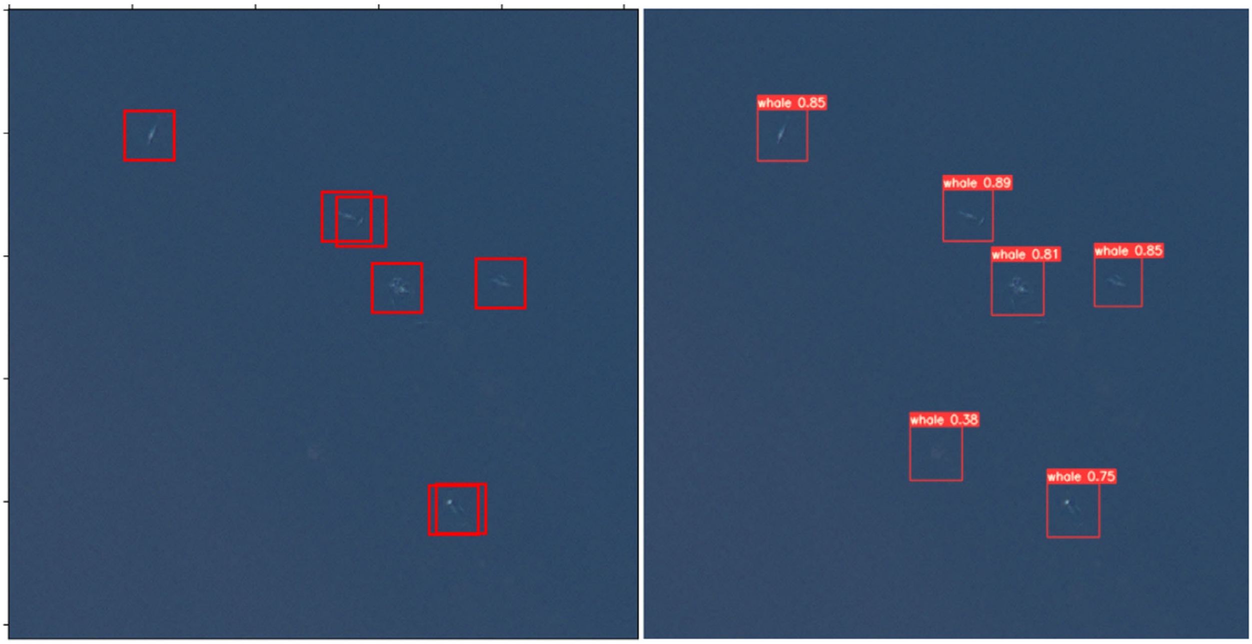

Figure 6. Correct and incorrect predictions from implementation 3 with corresponding confidence scores. Left panel: true labels. Right panel: Model predictions. Two FNs where labelled calves are not detected and one FP. Satellite image © 2022 Maxar Technologies.

Overhead images collected by UAS did not substantially improve whale precision and had mixed outcomes relative to the reference case. For example, adding aerial imagery to the training dataset was most effective when used as part of a transfer learning approach as it increased precision, but this implementation also had lower recall than all others, with more FNs (missed detections). These results suggest that overhead imagery might be best incorporated with a transfer learning approach when using CNNs to automatically detect whales, and combined with augmentations to improve performance and recall. In comparison with previous studies (Borowicz et al., 2019; Guirado et al., 2019) this study explores the application of a different CNN architecture to a satellite image dataset containing a larger number of labelled whales (503 compared with 32 and 62 respectively). Guirado et al. (2019) adopted a two-step approach where satellite images are initially scanned for tiles containing whales and the individual whales within these tiles and then counted, achieving a whale precision of 0.878 and recall of 0.763. Although a two-step approach allows larger areas to be scanned for whales with fewer resources, the first step may potentially discard images where whales are present but difficult to detect due to challenging sea state [i.e., poorer than average (Bamford et al., 2020)]. This can be mitigated by including training imagery (with whale identifications) from the areas and sea states where the model could be applied.

Our results highlight the importance of using training data from multiple sea states when developing CNNs for use in whale identification over broad areas. Further investigation into model performance in different sea states could be performed to quantify a threshold sea state where counting whales in satellite imagery is too unreliable, as is the case with traditional visual surveys (Bortolotto et al., 2016; Marsh & Sinclair, 1989; Panigada et al., 2011; see Bamford et al., 2020) for a classification of sea state in satellite imagery. As it is more difficult to confidently identify whales in rough seas, high-quality training data are more difficult to produce, limiting capacity to train models on these sea states. With increasing amounts of imagery, the identification threshold could be investigated, as well as whether identification issues could be overcome using semi-supervised or unsupervised learning approaches.

There are many potential goals for automation of whale identifications on satellite imagery. A simple goal may be detection of whales in an area of interest (presence information, e.g., in relation to human activities such as ship-strike risks, or understudied areas; Cubaynes et al., 2019). When being conservative about detection (e.g., allowing false positives), an automation approach that has high recall but low precision can work in this space (e.g., Borowicz et al., 2019). For example, if a semi-automated whale counting approach was being developed, whale recall may be the desirable metric to maximize, since false positives could be discarded manually by an expert. More challenging goals may be to identify relative abundance changes across time or space (informing areas of elevated habitat importance, or seasonal patterns). Most ambitious would be to generate results which are informative about absolute whale density (e.g., Bamford et al., 2020), for example, to measure local abundance. In the latter cases, it is particularly important to minimize bias generated by poor precision or recall metrics (Rodofili et al., 2022). The mAP@0.5 is the metric used to judge overall model performance and is maximized by the model in this study, but consideration should be given to the importance of each metric, particularly as the confidence at which the precision and recall are calculated is different from that of the mAP.

To minimize false-positive whale detections, including confounding classes is important in object detection (Borowicz et al., 2019). Here, we considered two classes (whales and boats), as boats can be of similar size and shape to whales. However, boats are a very variable category, and poor performance of the model in relation to this class impacts overall mAP. Our implementations showed promising performance, with mAP@0.5 scores ranging from 0.823 to 0.922. Most whale samples (80 of 90) were correctly identified but one boat was identified as a whale. This could be addressed by including more boat samples in the training data. However, most mistakes made by the model were either FNs or FPs (Fig. 6). This issue may be improved by including further ‘non-whale’ classes in the training data, such as ‘white caps’. It could also be beneficial to introduce a separate class for ‘calves’ or ‘whales with calves’ as many FN predictions were of calves labelled in the image but missed by the model, and separation of very close objects is a common problem for object detection (Diwan et al., 2022).

Mosaic augmentations are automatically applied by YOLOv5 to all input data to improve model performance. Using additional augmentation regimes improved mAP@0.5 and whale precision in comparison with the baseline. Augmentations increased variability in the training data, improving the ability of the network to generalize in testing. The use of image augmentation could be further investigated through implementation of stronger augmentations, such as random erasing (Shorten & Khoshgoftaar, 2019), on the training data. This could be approached via techniques such as RandAugment (Cubuk et al., 2019), which automatically searches the space of possible augmentations and selects the optimal ones.

Inclusion of aerial imagery in the training set has previously shown potential to improve results (e.g., Guirado et al., 2019). All implementations including aerial imagery marginally improved mAP@0.5 scores in relation to the baseline but had variable success across the other metrics and stronger improvement was seen with the addition of augmentations. These results suggest that including a wide range of aerial imagery in the training set may confuse the model, rather than improving its ability to generalize. However, implementation 8, which used transfer learning on the aerial data, achieved the highest mAP@0.5 (0.884) and a high whale precision (0.939) showing promise for this method of imagery combination. Inclusion of aerial imagery could be further investigated through optimizing the transfer learning process or investigating data fusion techniques (e.g., Duarte et al., 2018). In our study, many of the images taken from videos were very similar (multiple stills of the same individual over a short time), possibly providing little new information for the model to learn. The aerial images were collected from a different location to the satellite imagery which may have limited their ability to improve model performance in this case. A wider range of aerial imagery, in particular images taken from desired testing locations, may better improve model performance. However, this study suggests that the best method for improving model performance, particularly on inference in differing sea states, is to expand the training dataset to include a larger variety of satellite imagery both through expanding existing datasets and applying augmentations.

Conclusion

This work demonstrates the capability of CNNs, specifically YOLOv5, to detect whales in satellite imagery with good levels of precision and recall. This architecture provides accurate results with short training and detection times, making it ideal to scale up to larger volumes of satellite imagery both in training and testing. The inclusion of aerial imagery in the training dataset showed slight improvements on overall model performance, but improvements were not as strong as those achieved by basic augmentations. When formulating a workflow for automated whale detection, training data from all areas and sea states where detection is to be performed should be included for optimal performance, or performance can degrade drastically. While further details could be investigated and more training samples must be incorporated, the possibility for fully automated satellite-based detection of whales is a potentially transformative conservation tool that this work demonstrates is within our reach.

Code Availability

All of the codes used throughout this study are provided in a GitHub repository. The code used for model training, provided here (https://github.com/KMacfarlaneGreen/yolov5), was forked from the original repository by Jocher et al. (2021), which is continually updated.

Acknowledgements

This study formed part of the Ecosystems component of the British Antarctic Survey Polar Science for Planet Earth Programme, funded by The Natural Environment Research Council. We are thankful to the DigitalGlobe Foundation for providing free satellite imagery. UAS images were collected by Leigh Torres with permit number: NOAA/NMFS permit #16111. This work was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (reference EP/S022961/1).

References

Alter, S.E., Rynes, E. & Palumbi, S.R. (2007) DNA evidence for historic population size and past ecosystem impacts of gray whales. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 15162–15167.

Bamford, C.C.G., Kelly, N., Dalla Rosa, L., Cade, D.E., Fretwell, P.T., Trathan, P.N. et al. (2020) A comparison of baleen whale density estimates derived from overlapping satellite imagery and a shipborne survey. *Scientific Reports*, **10**, 12985.

Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y.M. (2020) YOLOv4: optimal speed and accuracy of object detection. *arXiv*. arXiv:2004.10934 [cs, eess].

Borowicz, A., Le, H., Humphries, G., Nehls, G., Höschle, C., Kosarev, V. et al. (2019) Aerial-trained deep learning networks for surveying cetaceans from satellite imagery. *PLoS One*, **14**, e0212532.

Bortolotto, G.A., Danilewicz, D., Andriolo, A., Secchi, E.R. & Zerbini, A.N. (2016) Whale, whale, everywhere: increasing abundance of Western South Atlantic humpback whales (*Megaptera novaeangliae*) in their wintering grounds. *PLoS One*, **11**, e0164596.

Burnett, J.D., Lemos, L., Barlow, D., Wing, M.G., Chandler, T. & Torres, L.G. (2019) Estimating morphometric attributes of baleen whales with photogrammetry from small UASs: A case study with blue and gray whales. *Marine Mammal Science*, **35**, 108–139. Available from: https://doi.org/10.1111/mms.12527

Caretta, J.V., Muto, M.M., Greenman, J., Wilkinson, K., Lawson, D., Viezbicke, J. et al. (2017) Sources of human-related injury and mortality for U.S. pacific west coast marine mammal stock assessments, 2011-2015. Available from: https://doi.org/10.7289/V5/TM-SWFSC-579

Charry, B., Tissier, E., Iacozza, J., Marcoux, M. & Watt, C.A. (2021) Mapping Arctic cetaceans from space: a case study for beluga and narwhal. *PLoS One*, **16**, e0254380.

Chen, Y., Zhang, C., Qiao, T., Xiong, J. & Liu, B. (2021) Ship detection in optical sensing images based on YOLOv5. *Twelfth international conference on graphics and image processing (ICGIP 2020)*. 117200E.

Clapham, P. (2016) Managing leviathan: conservation challenges for the great whales in a post-whaling world. *Oceanography*, **29**, 214–225.

Corrêa, A.A., Quoos, J.H., Barreto, A.S., Groch, K.R. & Eichler, P.P.B. (2022) Use of satellite imagery to identify southern right whales (*Eubalaena australis*) on a Southwest Atlantic Ocean breeding ground. *Marine Mammal Science*, **38**, 87–101. Available from: https://doi.org/10.1111/mms.12847

Cubaynes, H.C. & Fretwell, P.T. (2022) Whales from space dataset, an annotated satellite image dataset of whales for training machine learning models. Available from: https://doi.org/10.17863/CAM.84957

Cubaynes, H.C., Fretwell, P.T., Bamford, C., Gerrish, L. & Jackson, J.A. (2019) Whales from space: Four mysticete species described using new VHR satellite imagery. *Marine Mammal Science*, **35**, 466–491. Available from: https://doi.org/10.1111/mms.12544

Cubaynes, H.C., Rees, W.G., Jackson, J.A., Moore, M., Sformo, T.L., McLellan, W.A. et al. (2020) Spectral reflectance of whale skin above the sea surface: a proposed measurement protocol. *Remote Sensing in Ecology and Conservation*, **6**, 411–423. Available from: https://doi.org/10.1002/rse2.155

Cubuk, E.D., Zoph, B., Shlens, J. & Le, Q.V. (2019) RandAugment: Practical automated data augmentation with a reduced search space. *arXiv*. arXiv:1909.13719 [cs]. Available at: http://arxiv.org/abs/1909.13719

de Vos, A., Brownell, R.L., Tershy, B. & Croll, D. (2016) Anthropogenic threats and conservation needs of blue whales, *Balaenoptera musculus indica*, around Sri Lanka. *Journal of Marine Biology*, **2016**, e8420846.

Diwan, T., Anirudh, G. & Tembhurne, J.V. (2022) 'Object detection using YOLO: Challenges, architectural successors, datasets and applications', multimedia tools and applications. Available from: https://doi.org/10.1007/s11042-022-13644-y

Duarte, D., Nex, F., Kerle, N. & Vosselman, G. (2018) Multi-resolution feature fusion for image classification of building damages with convolutional neural networks. *Remote Sensing*, **10**, 1636.

Dumoulin, V. & Visin, F. (2018) A guide to convolution arithmetic for deep learning. *arXiv*. arXiv:1603.07285 [cs, stat]. Available at: http://arxiv.org/abs/1603.07285

Fretwell, P.T., Staniland, I.J. & Forcada, J. (2014) Counting southern right whales by satellite. *PLoS One*, **9**, e88655.

Gray, P.C., Bierlich, K.C., Mantell, S.A., Friedlaender, A.S., Goldbogen, J.A. & Johnston, D.W. (2019) Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry. *Methods in Ecology and Evolution*, **10**, 1490–1500. Available from: https://doi.org/10.1111/2041-210X.13246

Guirado, E., Tabik, S., Rivas, M.L., Alcaraz-Segura, D. & Herrera, F. (2019) Whale counting in satellite and aerial images with deep learning. *Scientific Reports*, **9**, 14259.

Hammond, P.S., Francis, T.B., Heinemann, D., Long, K.J., Moore, J.E., Punt, A.E. et al. (2021) Estimating the abundance of marine mammal populations. *Frontiers in Marine Science*, **8**, 735770.

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2018) Mask R-CNN. *arXiv*. arXiv:1703.06870 [cs]. Available at: http://arxiv.org/abs/1703.06870

Hodul, M., Knudby, A., McKenna, B., James, A., Mayo, C., Brown, M. et al. (2023) Individual North Atlantic right whales identified from space. *Marine Mammal Science*, **39**, 220–231. Available from: https://doi.org/10.1111/mms.12971

Höschle, C., Cubaynes, H.C., Clarke, P.J., Humphries, G. & Borowicz, A. (2021) The potential of satellite imagery for surveying whales. *Sensors*, **21**, 963.

Houegnigan, L., Merino, E.R., Vermeulen, E., Block, J., Safari, P., Moreno-Noguer, F. et al. (2022) Wildlife and marine mammal spatial observatory: Observation and automated detection of southern right whales in multispectral satellite imagery. *bioRxiv*. Available from: https://doi.org/10.1101/2022.01.20.477141v1

Jocher, G., Stoken, A., Borovec, J., Chaurasia, A., Xie, T., Changyu, L. et al. (2021) ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. Available at: https://zenodo.org/record/4679653#.YMDXlzZKh8Z [Accessed 6th September 2021].

Kapoor, S., Kumar, M. & Kaushal, M. (2023) Deep learning based whale detection from satellite imagery. *Sustainable Computing: Informatics and Systems*, **38**, 100858.

Kaschner, K., Quick, N.J., Jewell, R., Williams, R. & Harris, C.M. (2012) Public library of science, global coverage of cetacean line-transect surveys: status quo, data gaps and future challenges. *PLoS One*, **7**, e44075.

Khan, A., Sohail, A., Zahoora, U. & Qureshi, A.S. (2020) A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, **53**, 5455–5516.

Khan, C.B., Goetz, K.T., Cubaynes, H.C., Robinson, C., Murnane, E., Aldrich, T. et al. (2023) A biologist's guide to the galaxy: leveraging artificial intelligence and very high-resolution satellite imagery to monitor marine mammals from space. *Journal of Marine Science and Engineering*, **11**, 595.

Lamba, A., Cassey, P., Segaran, R.R. & Koh, L.P. (2019) Deep learning for environmental conservation. *Current Biology*, **29**, R977–R982.

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

Li, K., Wan, G., Cheng, G., Meng, L. & Han, J. (2020) Object detection in optical remote sensing images: a survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, **159**, 296–307.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D. et al. (2014) Microsoft COCO: common objects in context. *Computer vision – ECCV 2014*. Cham, pp. 740–755.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X. et al. (2020) Deep learning for generic object detection: a survey. *International Journal of Computer Vision*, **128**, 261–318.

Marsh, H. & Sinclair, D.F. (1989) Correcting for visibility bias in strip transect aerial surveys of aquatic fauna. *The Journal of Wildlife Management*, **53**, 1017–1024.

Maxar. (2022) Worldview legion - it takes a legion. Available at: https://www.maxar.com/splash/it-takes-a-legion [Accessed 10th October 2022].

Nicol, C., Bejder, L., Green, L., Johnson, C., Keeling, L., Noren, D. et al. (2020) Anthropogenic threats to wild cetacean welfare and a tool to inform policy in this area. *Frontiers in Veterinary Science*, **7**, 57.

Noad, M.J., Kniest, E. & Dunlop, R.A. (2019) Boom to bust? Implications for the continued rapid growth of the eastern Australian humpback whale population despite recovery. *Population Ecology*, **61**, 198–209. Available from: https://doi.org/10.1002/1438-390X.1014

Panigada, S., Lauriano, G., Burt, L., Pierantonio, N. & Donovan, G. (2011) Monitoring winter and summer abundance of cetaceans in the Pelagos sanctuary (northwestern Mediterranean Sea) through aerial surveys. *PLoS One*, **6**, e22878.

Planet Labs. (2022) Our next-generation satellite constellation Pelican is expected to deliver very-high-resolution and rapid-revisit capabilities. Available at: https://www.planet.com/pulse/our-next-generation-satellite-constellation-pelican-is-expected-to-deliver-very-high-resolution-and-rapid-revist-capabilities/ [Accessed 10th October 2022].

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016) You only look once: unified, real-time object detection. *arXiv*. arXiv:1506.02640 [cs].

Reeves, R.R. & Smith, T.D. (2010) Commercial whaling, especially for gray whales, *Eschrichtius robustus*, and humpback whales, *Megaptera novaeangliae*, at California and Baja California Shore Stations in the 19th century (1854–1899). *Marine Fisheries Review*, **25**, 1–25.

Rocha, R.C., Jr., Clapham, P.J. & Ivashchenko, Y. (2015) Emptying the oceans: a summary of industrial whaling catches in the 20th century. *Marine Fisheries Review*, **76**, 37–48.

Rodofili, E.N., Lecours, V. & LaRue, M. (2022) Remote sensing techniques for automated marine mammals detection: a review of methods and current challenges. *PeerJ*, **10**, e13540.

Roman, J., Estes, J.A., Morissette, L., Smith, C., Costa, D., McCarthy, J. et al. (2014) Whales as marine ecosystem engineers. *Frontiers in Ecology and the Environment*, **12**, 377–385. Available from: https://doi.org/10.1890/130220

Savoca, M.S., Czapanskiy, M.F., Kahane-Rapport, S.R., Gough, W.T., Fahlbusch, J.A., Bierlich, K.C. et al. (2021) Baleen whale prey consumption based on high-resolution foraging measurements. *Nature*, **599**, 85–90.

Scordino, J., Carretta, J. & Cottrell, P. (2018) *Bycatch and ship strikes of gray whales in U.S. and Canadian waters, 2008–2012*. Paper. SC/65b/BRG21 presented to the IWC Scientific Committee, May 2014.

Shorten, C. & Khoshgoftaar, T.M. (2019) A survey on image data augmentation for deep learning. *Journal of Big Data*, **6**, 60.

Silber, G., Weller, D., Reeves, R., Adams, J. & Moore, T. (2020) Co-occurrence of gray whales and vessel traffic in the North Pacific Ocean. *Endangered Species Research*, **44**, 201.

Torres, L.G., Nieukirk, S.L., Lemos, L. & Chandler, T.E. (2018) Drone up! Quantifying whale behavior from a new perspective improves observational capacity. *Frontiers in Marine Science*, **5**, 319.

UP42 GmbH. (2019) Reveal new insights with Pléiades Neo Data. Available at: https://up42.com/goingup/pleiades-neo [Accessed 10th October 2022].

Wada, K. (2021) wkentaro/labelme. Available at: https://github.com/wkentaro/labelme

Würsig, B., Perrin, W., Würsig, B. & Thewissen, J.G.M. (Eds.). (2008) *Encyclopedia of marine mammals*, 2nd edition. Cambridge, MA: Academic Press. Available from: https://www.elsevier.com/books/encyclopedia-of-marine-mammals/wursig/978-0-12-373553-9 [Accessed 10th October 2022].

Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014) How transferable are features in deep neural networks? *arXiv*. arXiv:1411.1792 [cs].

Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. (2023) Object detection in 20 years: A survey. *arXiv*. arXiv:1905.05055 [cs]. Available at: http://arxiv.org/abs/1905.05055

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Precision and recall results on the boat class.
Figures S1–S9. Confusion matrices for all implementations.