Canadian Science Publishing

ARTICLE

Detection and tracking of belugas, kayaks and motorized boats in drone video using deep learning[1]

Madison L. Harasyn, Wayne S. Chan, Emma L. Ausen, and David G. Barber

Abstract: Aerial imagery surveys are commonly used in marine mammal research to determine population size, distribution and habitat use. Analysis of aerial photos involves hours of manually identifying individuals present in each image and converting raw counts into useable biological statistics. Our research proposes the use of deep learning algorithms to increase the efficiency of the marine mammal research workflow. To test the feasibility of this proposal, the existing YOLOv4 convolutional neural network model was trained to detect belugas, kayaks and motorized boats in oblique drone imagery, collected from a stationary tethered system. Automated computer-based object detection achieved the following precision and recall, respectively, for each class: beluga = 74%/72%; boat = 97%/99%; and kayak = 96%/96%. We then tested the performance of computer vision tracking of belugas and occupied watercraft in drone videos using the DeepSORT tracking algorithm, which achieved a multiple-object tracking accuracy (MOTA) ranging from 37% to 88% and multiple object tracking precision (MOTP) between 63% and 86%. Results from this research indicate that deep learning technology can detect and track features more consistently than human annotators, allowing for larger datasets to be processed within a fraction of the time while avoiding discrepancies introduced by labeling fatigue or multiple human annotators.

Key words: computer vision, deep learning, unmanned aerial vehicle (UAV), beluga, object detection, object tracking.

Résumé : Les relevés par imagerie aérienne sont couramment utilisés dans la recherche sur les mammifères marins pour déterminer la taille de la population, sa répartition et l'utilisation de l'habitat. L'analyse des photos aériennes implique des heures d'identification manuelle des individus présents dans chaque image et la conversion des chiffres bruts en statistiques biologiques utilisables. Notre recherche propose l'utilisation d'algorithmes d'apprentissage en profondeur pour augmenter l'efficacité du flux de recherche sur les mammifères marins. Pour mettre à l'essai la faisabilité de cette proposition, le modèle de réseau de neurones à convolution YOLOv4 existant a été entraîné pour détecter les bélugas, les kayaks et les embarcations motorisées dans des images de drones obliques, recueillies à partir d'un système fixe relié. La détection automatisée d'objets par ordinateur a atteint la précision et le rappel suivants, respectivement, pour chaque classe : béluga : 74 %/72 %; bateau : 97 %/99 %; kayak : 96 %/96 %. Les auteurs ont ensuite testé la performance de poursuite au moyen de la vision par ordinateur des bélugas et des motomarines dans des

Received 20 August 2021. Accepted 14 December 2021.

M.L. Harasyn,[*,†] W.S. Chan,[†] E.L. Ausen, and D.G. Barber. Centre for Earth Observation Science, Department of Environment and Geography, University of Manitoba, Winnipeg, Canada.
Corresponding author: Madison L. Harasyn (e-mail: madison.harasyn@usask.ca).

[*] Present address: Coldwater Laboratory, 1151 Sidney Street, Canmore, AB T1W 3G1, Canada.
[†] Both authors contributed equally.
[1] This article is part of a collection on Unoccupied Vehicle Systems in Arctic Research and Monitoring jointly published by Arctic Science and Drone Systems and Applications.

Drone Syst. Appl. **10:** 77–96 (2022) dx.doi.org/10.1139/juvs-2021-0024 Published at www.cdnsciencepub.com/dsa on 4 January 2022.

vidéos de drones à l'aide de l'algorithme de poursuite DeepSORT, qui a obtenu une exactitude de poursuite des objets multiples (« MOTA ») allant de 37 à 88 % et une précision de poursuite des objets multiples (« MOTP ») allant de 63 à 86 %. Les résultats de cette recherche indiquent que la technologie d'apprentissage profond peut détecter et suivre les caractéristiques plus régulièrement que les annotateurs humains, permettant de traiter des ensembles de données plus volumineux en une fraction de temps tout en évitant les écarts introduits par la fatigue d'étiquetage ou de multiples annotateurs humains. [Traduit par la Rédaction]

Mots-clés : vision par ordinateur, apprentissage en profondeur, véhicule aérien sans pilote (UAV), béluga, détection d'objets, poursuite d'objets.

1. Introduction

Marine wildlife population statistics provide information on the success of these species under human influence and are used to develop conservation strategies for individual species and populations. Monitoring these populations is particularly important for Arctic species, which are facing a dramatic change in their habitat due to climate change and the resulting change in sea ice cover and prey distribution (Kelley et al. 2010). An increasing threat for Arctic marine wildlife is noise pollution generated by shipping traffic and resource exploration, which can mask communication between individuals and increase mammal stress levels, leading to a lowered immune response and reproductive success (Rolland et al. 2012; Erbe et al. 2016). Another rapidly increasing source of anthropogenic influence on marine wildlife is ecotourism (Giampiccoli et al. 2020), however the impacts of these activities are dependent on the species and population in question. For example, dwarf minke whales (*Balaenoptera acutorostrata* Lacépède, 1804) were observed more than expected within 60 m of whale tourism boats in the Great Barrier Reef, showing attraction (Mangott et al. 2011), whereas populations of killer whales (*Orcinus orca* (Linnaeus, 1758)) in British Colombia, and beluga (*Delphinapterus leucas* (Pallas, 1776)) in Onega Bay both displayed behaviors including fleeing and diving to avoid nearby tourist vessels (Williams et al. 2002; Krasnova et al. 2020). The range in response to anthropogenic activities based on species and population highlights the importance of researching the influence of human activities on wildlife at a group or population scale.

Population statistics are commonly obtained from aerial surveys, collected via visual observation from aircraft or from aerial imagery (e.g., Lowry et al. 2017; Stafford et al. 2018). Recently, drones have become a popular method for cetacean monitoring due to their low cost, minimal disruption to wildlife, and the reduction of risk to humans in comparison to occupied aircraft (e.g., Hodgson et al. 2013; Koski et al. 2015; Ferguson et al. 2018; Torres et al. 2018). An array of individual and population variables can be gathered from aerial observations: population size, sex and age distribution, individual health, and migratory patterns of groups (Ferguson et al. 2018). These variables are derived from aerial imagery by a trained observer, who manually counts individuals and makes note of the size, colour, and identifying features of individuals.

Manual photo analysis expends countless work hours for scientists and can be prone to errors as a result of observer fatigue (Wang et al. 2019). In response to this, researchers have begun to develop methods of automated detection of marine wildlife in aerial imagery. Borowicz et al. (2019) and Guirado et al. (2019) have both applied convolutional neural networks (CNN) to high-resolution satellite imagery for automated detection of whales in mid-latitudes, achieving detection rates over 90%. CNNs have also been applied to individual whale identification, with Bogucki et al. (2019) correctly identifying individual whales in 87% of aerial photos. The high detection metrics achieved by these studies is encouraging for the transition towards adopting machine learning algorithms as an assistive technology for cetacean detection in aerial imagery.

The studies mentioned above involved the *detection* of wildlife in imagery, rather than the *tracking* of wildlife over time. In a temporally ordered sequence of images (i.e., a video), detection algorithms can only identify objects in each image, but they are not able to match individual objects between successive images unless there is only one object being tracked. For example, a detection algorithm may determine that there are four people in a video frame and the same number in the next frame, but it is not able to keep track of individuals from frame to frame. Detection only provides a snapshot of a population at a particular time, limiting the types of statistics that can be derived from the data. In contrast, tracking algorithms follow the movements of individual objects through a sequence of images and can provide information on interactions between individuals within a group, or interactions between individuals and other objects within the frame.

Wildlife tracking can be achieved using GPS microchips (e.g., Chambault et al. 2020), radio-frequency identifier (RFID) telemetry (e.g., Bridge et al. 2019), acoustic telemetry (e.g., Hauser et al. 2014; Hastie et al. 2019), or through video tracking systems (Manabe 2017). GPS, RFID and acoustic tracking all involve the attachment of a transmitter tag to the animal, ranging in size from millimetres to several centimetres (e.g., Robbins et al. 2013; van Harten et al. 2019). Transmitter tags allow for remote tracking of tagged individuals over a large geographic area, including vertical movement of marine animals in the water column (Citta et al. 2013; Hauser et al. 2014; Stafford et al. 2018). Despite these benefits, the number of individuals which can be tracked using these methods is limited by cost and time, and tag deployment can cause physiological damage to tagged individuals (Walker et al. 2012). Video tracking of wildlife provides a non-invasive alternative to tagging. Although video-based tracking is limited both spatially by the field-of-view of the camera, and temporally by the time interval of data collection, large populations of animals can be tracked without interference to the species or habitat using video methods. As well, video capture can provide information on environmental variables and potential interferences to the population (e.g., Hodgson et al. 2013) which may not be captured by tag-based methods of tracking.

Object *detection* algorithms process each frame individually, denoting the location and classification probability of each object on a frame-by-frame basis (Redmon et al. 2016). Object *tracking*, on the other hand, assigns a unique identifier to each object in the first video frame, and then identifies the location of each object in consecutive frames using location and displacement information from the proceeding frames (Betke et al. 2000). This allows for the researcher to extract information on the movement of individuals within a scene, and to account for potential outliers or unique movement patterns of individuals. The simultaneous tracking of multiple objects generates challenges such as occlusion, which occurs when objects overlap or are hidden from view (Papadourakis and Argyros 2010; Motro and Ghosh 2018).

Initial development of tracking algorithms for behaviour classification of animals in video was conducted in controlled laboratory settings on small animals such as pigeons (Pear 1985), with extensive development leading to 90% of detections being correct in lab-based tracking (Burgos-Artizzu et al. 2012; Giancardo et al. 2013; Hong et al. 2015; Robie et al. 2017). Application of these tracking algorithms becomes complicated in a natural setting, due to the variability of background colour/texture, solar illumination, and the increased potential for the feature of interest to be occluded between video frames (Wang et al. 2019). Despite these complications, previous studies have applied tracking algorithms to wildlife videos, achieving accuracies ranging from 86.6% for honeybees (Ratnayake et al. 2021) to 91.3% for elephants and humans (Bondi et al. 2020).

Our research aims to expand on the previous literature by addressing the following questions:

1. How difficult is it for human observers to accurately identify and track whales, kayaks and motorized boats in drone video?
2. Can an existing deep learning algorithm be used to detect whales, kayaks and motorized boats in drone video with performance that is comparable or superior to human observers?
3. If detection is successful, can an existing multi-object tracking algorithm be used to track individual whales, kayaks and motorized boats in drone video over time?

This research is a part of a collaborative research project led by the Centre for Earth Observation Science (CEOS) at the University of Manitoba, and Fisheries and Oceans Canada (DFO). The research project aims to quantify the effect of watercraft/ship presence on beluga behaviour in the Churchill River estuary, in efforts to inform policy development for the protection of the Western Hudson Bay beluga population. The goal of the present study is to provide automated methods of obtaining quantitative statistics on the presence and movement of beluga and watercraft in the estuary, which can be interpreted by biologists in future research to investigate behavioural responses of beluga to changes in the environment.

2. Materials and methods

2.1. Study area

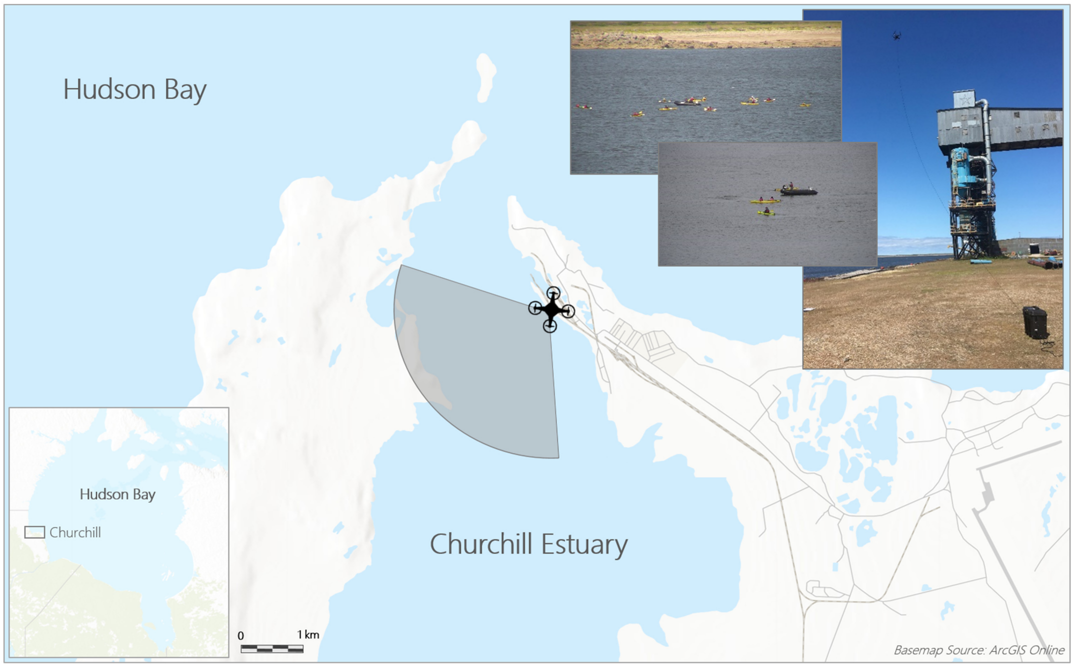
Drone video data were collected between 28 July and 9 August 2019 in the Churchill River estuary, located in northern Manitoba, Canada, off the southwest coast of Hudson Bay (Fig. 1). The Churchill River estuary is home to a subset of the Western Hudson Bay beluga population each year from June to September. This population has an estimated size of 55 000–60 000 belugas and during the last photographic aerial survey completed in 2015, an estimated 3 000 belugas were counted in the Churchill River estuary (Matthews et al. 2016). The Churchill River estuary is a unique environment for beluga due to frequent interactions with humans and small watercraft resulting from tourist activities. Unlike offshore waters near other Arctic communities, beluga are not often hunted within the Churchill River estuary, which may contribute to beluga being less avoidant of boats (Caron and Smith 1990; Tyrrell et al. 2007; Malcolm and Penner 2011; Doniol-Valcroze et al. 2013).

Tourist activities in the estuary include Zodiac inflatable boat tours and kayaking, which allow for close interaction between humans and belugas (Manitoba's Western Hudson Bay Ad Hoc Beluga Habitat Sustainability Plan Committee 2016). The Port of Churchill is a shipping hub for grain exports out of Hudson Bay, generating large-sized vessel traffic in the port during the summer months (COSEWIC 2014). The effects of these anthropogenic activities on beluga in the Churchill River estuary have been researched by Malcolm and Penner (2011), who concluded, based on frequent interactive behaviour demonstrated by belugas within 25 m of vessels, that beluga potentially show attraction to tourist boats in the Churchill River estuary. These conclusions resulted in the recommendation for the development of a travel corridor in the estuary, and certain restrictions of motorized vessels around beluga, such as a minimum approach distance and maximum travel speed (Malcolm and Penner 2011).

2.2. Drone system/Data collection

Aerial video data were captured using a M210 RTK quadcopter (DJI 2018) equipped with a DJI Zenmuse Z30 2 MP camera with a 30× optical zoom, recording video at 30 fps (Fig. 1). The Zenmuse Z30 gimbal has a full 360° horizontal rotation and 180° vertical rotation, allowing for video capture of the full estuary area. The drone was attached to the NTP PowerLine

Fig. 1. Location of drone video capture, (inset) setup of tethered drone and example imagery from video capture.



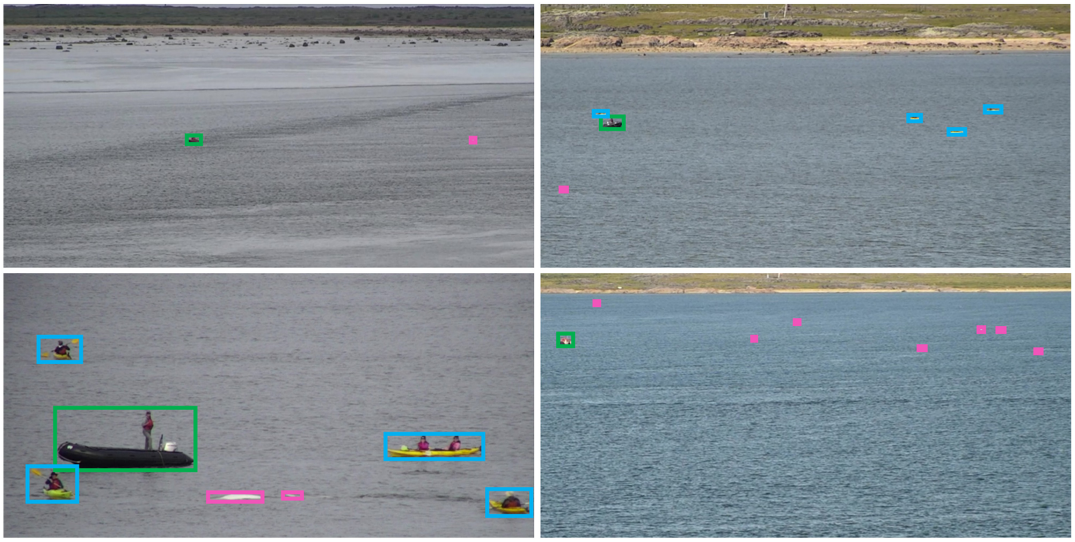
(Menet Aero 2017), a tethered power supply which provides constant power to the drone. This allowed for the drone to remain airborne continuously throughout data collection.

Drone video data were captured at a constant altitude of 18 m, and at a specified location (58.7733°N 94.1917°W) on the eastern estuary shore located within the Churchill Port grounds, at a minimum horizontal distance of 500 m away from the closest observed beluga (Fig. 1). The flight permits obtained for this study only allowed a stationary platform at a given location and height, limiting video collection to an oblique view. Data collection was conducted when weather conditions permitted (winds < 10 m/s, no precipitation), resulting in approximately 6 hours total of video over 6 collection days. In all videos, water conditions were calm with no white caps; however, sun illumination varied between days of data collection resulting in different water illumination between datasets (Fig. 2). For every day of data collection, video capture was collected at different magnitudes of zoom and camera orientations to provide different scenes and levels of detail of features within the videos, offering a variety of conditions for computer algorithm training and testing.

2.3. Object detection

In recent years, significant strides have been made in the application of CNNs to computer vision tasks such as object detection. CNNs are a class of machine learning models that gain their inspiration from biological neural networks. State-of-the-art CNNs contain many layers, which led to the use of the term “deep learning” to denote the training of a neural network with multiple layers. Starting with the first layer, which receives the original image as input, each subsequent layer receives an image from the previous layer as input and performs a transformation on it, which is then output to the next layer. The term “convolutional” comes from the use of the mathematical operation called *convolution* (the combination of two functions to produce a third function) in CNNs (Goodfellow et al. 2016).

Fig. 2. Examples of validation label sets from video frames captured on different dates, used for YOLOv4 algorithm training. Video frame on the bottom right is an example of 0× zoom, where belugas appear quite small in the frame making manual identification difficult. Classes are as follows: belugas (pink), boats (green), kayaks (blue).



2.3.1. *Object detection algorithm*

There were several desired specifications in selecting a deep learning algorithm for object detection: it should be open-source, able to be trained on custom dataset/classes, and able to classify features relatively fast to allow for rapid processing of frames within a video. You Only Look Once (YOLO) v4 (Bochkovskiy et al. 2020), based on the Darknet framework, fit all these criteria. There also exists comprehensive tutorials and documentation for YOLOv4 online, making it more accessible to a wide variety of users. YOLOv4 is based on a single CNN framework, which simultaneously predicts bounding boxes and class probabilities for each image/video frame based on pixel colour/texture. This allows the YOLOv4 algorithm to process images at a speed of around 65 fps (with a Tesla V100 GPU) while still delivering reliable results, making it a suitable choice for processing hours of marine mammal video.

2.3.2. *Training parameters*

The YOLOv4 model was trained using images with width and height of 768 pixels × 768 pixels, compressed from their original dimensions (1920 pixels × 1088 pixels). Images had to be reduced in size due to GPU memory constraints, with the selected dimensions (768 × 768) being the largest that could be reliably handled by the computer system used for training. Aspect ratio was not preserved in the resizing. A parameter called *letter_box* that maintained aspect ratio during image resizing was tested but did not give improved results.

The number of batches was set to 64 and the subdivisions to 16. These values were arrived at through experimentation. The initial learning rate was 0.001, with a decay rate of 0.0005 and momentum of 0.949, which were the default values used by YOLOv4. The starting set of weights was pre-trained on the Microsoft COCO (Common Objects in Context) dataset, which contains 80 classes representing a variety of common objects (Lin et al. 2015).

YOLOv4 training was performed on a GPU node (4 × NVIDIA Tesla V100 – 32 GB, dual Intel Xeon Gold 5218 Cascade Lake CPUs, 192 GB RAM) of the University of Manitoba's Grex computing cluster (Grex Technical Specifications 2018), which had CUDA 10.2 and cuDNN 7.6.5 installed. Training was performed for 18 000 iterations, which took 39 h to complete. Additional training runs up to 36 000 iterations were performed but did not yield better results.

2.3.3. Object detection ground truth

YOLOv4 requires a correctly labelled set of images to serve as ground truth for training. For training, a dataset of images was generated from every 90th frame (or every 3 s) of selected drone videos. Videos of varying zoom, scene, environmental conditions, and number of objects were selected for training image generation to provide the algorithm with a wide range of object appearances (Fig. 2).

A bounding box must be drawn around each identified object in an image and labelled with its object class (e.g., "beluga"). The ground truth dataset was created by using a graphical image annotation tool called LabelImg (Tzutalin 2015) to draw the bounding boxes and label them with their object class. Three classes of objects were identified in the images: belugas, kayaks, and motorized boats. The kayak class included one- and two-person kayaks and the boat class included Zodiac inflatable boats and hard-hulled motorized vessels. Belugas were only annotated when they appeared above water. In total, 721 images were extracted from drone videos, with 75% (541 images) being used for algorithm training and 25% (180 images) being used for validation (used for calculations in Section 3). Of the 721 images, 535 images contained belugas, 544 images contained boats, 481 images contained kayaks, and 40 images had none of these classes.

All manual labeling for the ground truth dataset was performed by one of the authors (M. Harasyn). It is important to note that in this research, we use the term 'ground truth' to represent the human annotated dataset used to train the deep learning algorithm. The term does not reflect the correctness of the dataset, as there is no way to verify the annotations. We consider it to be an "expert" labeling done by an experienced annotator, but it may still contain subjective labeling decisions and human error.

2.4. Object tracking

Once object detection was performed by YOLOv4, the processed frames were passed to a multiple object tracking algorithm to conduct the tracking phase of the computer vision workflow, in which individual objects were tracked through consecutive frames in the video.

2.4.1. Object tracking algorithm

After training YOLOv4, the weights from the model were imported into an implementation of DeepSORT (Simple Online and Real-time Tracking with a Deep Association Metric) (Wojke et al. 2018), a tracking-by-detection algorithm, to track individual belugas, boats, and kayaks. DeepSORT, which is an extension of the original SORT algorithm (Bewley et al. 2016), combines Kalman filtering within image space and frame-by-frame data association to measure bounding box overlap to identify corresponding detections in consecutive video frames (Wojke et al. 2018). DeepSORT is an online algorithm that only considers information about the current video frame and past frames, rather than processing the entire video before making decisions. An existing implementation of DeepSORT called YOLOv4-DeepSORT (The AI Guy 2007) was selected as it is compatible with YOLOv4 weights and is simple to implement on the same framework as YOLOv4, allowing for a simple transition from detection to tracking.

As input, a weights file from a previously trained YOLOv4 model and a RPAS video file were provided to YOLOv4-DeepSORT. The program performed object detection using the previously trained weights and then passed the processed frames to the DeepSORT algorithm, which performed multiple object tracking. A new video file was output, consisting of the original video annotated with bounding boxes around the tracked objects. Each box was labelled with an object class and an identification number assigned to the object.

2.4.2. Tracking parameters

The *max_age* parameter (maximum length of time before a specific object track is deleted because it is considered to have left the field of view) in DeepSORT was set to 30 frames, or 1 s, and the *n_init* parameter (number of consecutive detections before a track is confirmed) was set to 5. These parameter values were chosen as they deliver the optimal tracking metrics for our dataset (MOTA, MOTP; see Section 2.6.2), determined through systematic variation of the parameters.

2.4.3. Object tracking ground truth

A ground truth dataset for object tracking was generated using the web-based Computer Vision Annotation Tool (CVAT; OpenVINO Toolkit 2021). The ground truth set was comprised of five 30-minute segments of videos independent from the detection training set, manually labeled frame-by-frame. CVAT assists in manual tracking by automating the movement of a label in a linear direction based on movement detected in previous frames. This feature was used in manual labeling, such that track movement by CVAT was used when it appropriately described object movement. The five manually labeled videos, numbered 5, 9, 10, 15, and 16, were selected to represent a range of object counts and zoom magnitudes (see Section 3.2.1). All manual labeling for the object tracking ground truth videos was performed by one of the authors (M. Harasyn).

2.5. Human-annotated datasets

2.5.1. Image subsets

To gauge human performance on the object detection task, a subset of 95 images from the full 721 image dataset was annotated independently by two summer students and by one of the authors (E. Ausen). The students were trained by our team for object labeling using the LabelImg tool. E. Ausen was experienced in image labeling and did not participate in the training. The three manually labelled image subsets were used to calculate the interobserver agreement (Johnston et al. 2020) between the three human annotators and the ground truth on the task of identifying and labeling the class of each object in the images. The inter-observer agreement was calculated between all datasets using an open-sourced mean average precision (mAP) program (Cartucho et al. 2018), which determines the agreement between a ground truth dataset and a test dataset by calculating the intersection over union (IoU) for pairwise bounding boxes (Cartucho et al. 2018). The IoU is defined as the intersection (overlap) of the areas of two bounding boxes, divided by the union (sum) of the areas of the bounding boxes minus the intersection of the areas. Bounding boxes were considered a match if the IoU was greater than 50% and the class label (e.g., “beluga”, “boat”, or “kayak”) of the boxes was the same. The class-based average precision (AP; see Section 2.6.1 for definition) scores calculated between the ground truth dataset and the three human-annotated datasets will be used as a baseline for detection accuracy, to compare with the AP of each class produced by the YOLOv4 object detection algorithm. A comparison to other human annotators will act as a check of observer variability in classification and provide an indication of the degree of difficulty of the labeling task.

2.5.2. Video subsets

To evaluate human performance on object tracking, human-annotated subsets were also produced for three of the videos used for DeepSORT tracking (video 5, 10 and 16). The full duration of each video was labeled by two of the authors (W. Chan and E. Ausen) using the CVAT tool, following the same methods used by M. Harasyn. MOTA and MOTP were calculated between the ground truth and the three subsets to determine the human accuracy of tracking beluga and watercraft.

2.6. Analyses

2.6.1. Object detection performance

To assess the effectiveness of the trained object detection model, the precision, recall, and F1 score were calculated as follows:

$$(1)\quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$(2)\quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$(3)\quad \text{F1Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP is the number of true positives (correctly identified object), FP is the number of false positives (object is misidentified, or identified but no object present), and FN is the number of false negatives (object exists but is not identified by the model) (see Padilla et al. 2021).

Precision indicates the amount of misclassification occurring in the model, whereas recall indicates how many detections were missed, with values closer to 1 representing a more accurate model. The F1 score calculates the balance between precision and recall, meaning a higher F1 score represents a model that is both robust and precise. Precision and recall are used to calculate the AP, or the overall performance of the model under different confidence thresholds (the overlap percentage needed to be classified as a true positive):

$$(4)\quad \text{AP} = \sum_{n} (r_{n+1} - r_n) \max_{\tilde{r} \geq r_{n+1}} p(\tilde{r})$$

where $p(\tilde{r})$ is the measured precision at recall $\tilde{r}$ (Padilla et al. 2021). AP is often reported for each class and is then averaged across all classes and reported as the mAP for overall object detection. Object detection AP was calculated using the previously referenced mAP program (Cartucho et al. 2018).

2.6.2. Object tracking performance

The performance of multiple object tracking with DeepSORT was evaluated using the py-motmetrics library (Heindl et al. 2020), which calculates the multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP). MOTA balances the error associated with false negatives (FN), false positives (FP) and identity switches (changing of the ID value assigned to an object) of a tracked object (IDSW), calculated as follows:

$$(5)\quad \text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t}$$

where t is the frame index, and GT is the number of ground truth objects in the respective frame (Milan et al. 2016). MOTP indicates the dissimilarity between all true positive bounding boxes, and their respective ground truth bounding boxes, calculated as:

$$\text{(6)} \quad \text{MOTP} = \frac{\sum_{t,i} d_{t,i}}{\sum_{t} c_t}$$

where c_t is the number of true positives in frame t, and $d_{t,i}$ is the bounding box overlap of detected object i with its respective ground truth bounding box (Milan et al. 2016). MOTA indicates the tracking model's ability to maintain accurate object trajectories, whereas MOTP indicates the ability for the model to estimate the precise location of objects.

2.6.3. Human performance

For each of the three object classes, the AP (see Section 2.6.1) was calculated between the three manually labelled image subsets and the corresponding images in the ground truth dataset, otherwise known as the inter-observer agreement. This was done to determine how well humans can perform the task of identifying belugas, boats, and kayaks in the drone imagery. In addition, multiple object tracking metrics (MOTA and MOTP) were calculated between three manually labelled video subsets and the corresponding videos in the ground truth dataset, to determine how well humans can perform on the task of tracking individual belugas, boats, and kayaks through consecutive drone video frames. The image and video subsets both serve as baseline conditions for comparison with the performance of machine learning on the tasks.

3. Results

3.1. Object detection

3.1.1. Model performance

The highest AP achieved during training for YOLOv4 detection of each class was: 61.17% for belugas, 98.58% for boats, and 95.97% for kayaks (at 0.5 confidence threshold). This resulted in an overall mAP of 85.2% for the trained YOLOv4 model (precision = 89%, recall = 92%, F1 score = 0.9). A lower precision in comparison to recall indicates that detection for all classes has a higher rate of false positives than false negatives (Table 1), caused by model over-detection. Considering each class individually, it is apparent that the model is the least accurate at detecting belugas, which accounts for the majority of the false positives and false negatives (Table 1). Beluga misidentification is often a result of clustered belugas being identified as a single beluga or belugas in the background being missed, resulting in false negatives (Fig. 3). False positive occurrence is related to belugas generating water disturbances while they are near the water surface (Fig. 3). Boats and kayaks are identified very well by the trained YOLOv4 model and are surprisingly well identified even when they are partially off screen or when appearing quite small within the image. When watercraft are overlapping, the detection model can correctly detect and classify watercraft, however the precise locations of watercraft do not always agree with the ground truth leading to misidentifications (Fig. 3).

3.1.2. Human performance

As shown in Table 2, the high inter-observer agreement between the ground truth labels and the subset labels for boats and kayaks indicates that the labeling of these object classes can be independently replicated by human annotators with high accuracy. However, beluga labels varied significantly between annotators (Fig. 4). Ground truth labels for boats and kayaks are in higher agreement with the summer students' labels than with E. Ausen's labels, despite Ausen's greater experience with data annotation (Table 2).

Table 1. Number of true positives (TP), false positives (FP) and false negatives (FN) for each class with the trained YOLOv4 model run on the full ground truth dataset.

Class	Total number in ground truth set	Number of detected objects	TP	FP	FN	Precision (%)	Recall (%)	F1 score
Beluga	414	403	300	103	114	74.44	72.46	0.73
Boat	150	152	148	4	2	97.37	98.67	0.98
Kayak	648	649	625	24	23	96.30	96.45	0.96
Total	1212	1204	1073	131	139	89.12	88.53	0.89

Fig. 3. Examples of where YOLOv4 detection fails. Dark blue boxes represent true positives, red boxes represent false positives and light blue/white boxes represent ground truth boxes. The top row displays difficulty of detection in crowded scenes, classification of a water disturbance as a beluga, and difficulty of identifying belugas in the background, from left to right. The bottom row displays instances where YOLOv4 prediction boxes and ground truth boxes for belugas did overlap, but IoU was too low to be classified as a match.

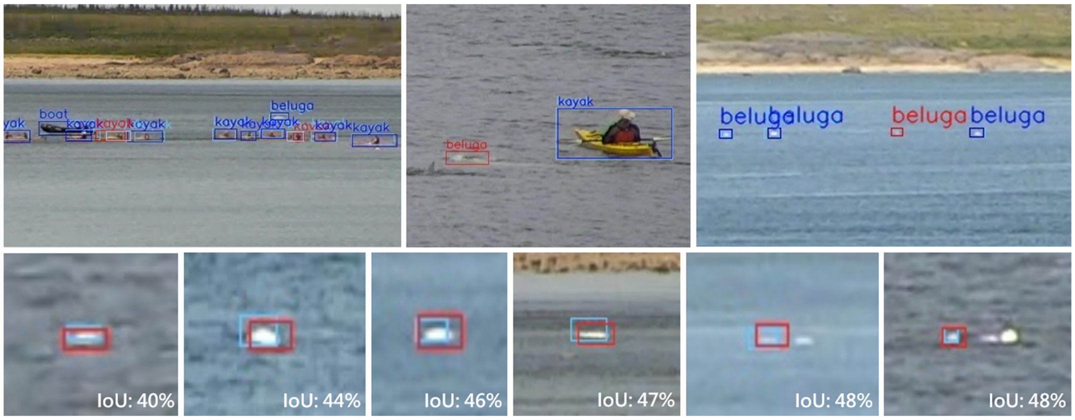
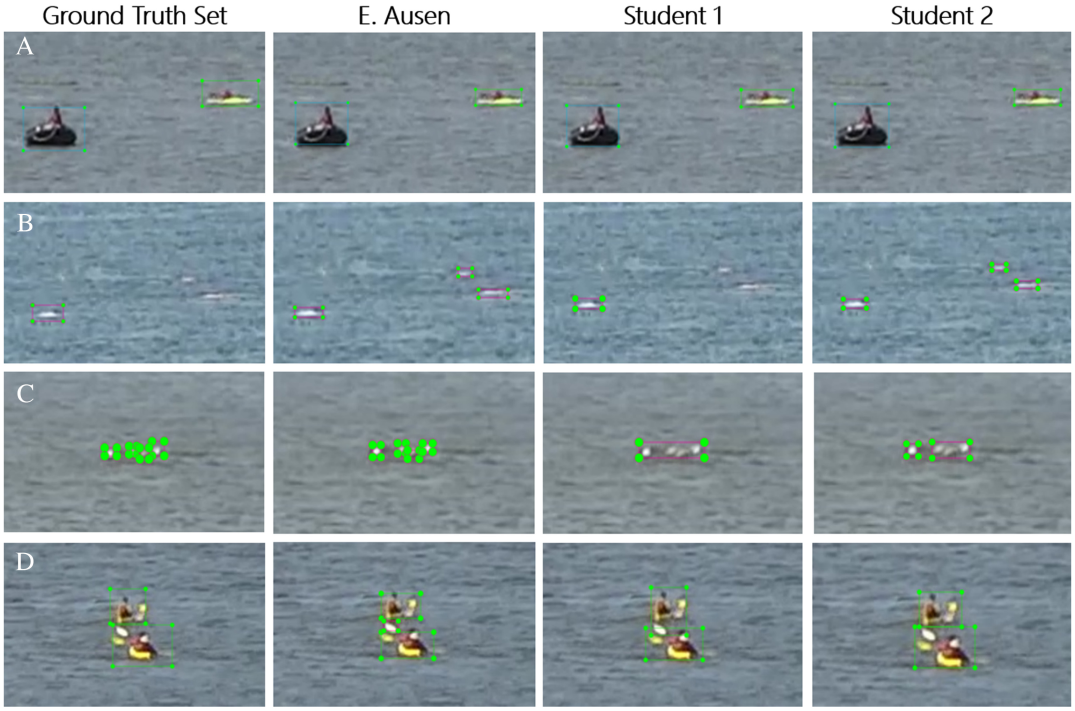


Table 2. Average precision (AP) for the three human annotators, compared to the ground truth dataset.

	E. Ausen (%)	Student 1 (%)	Student 2 (%)
Beluga AP	24.7	18.2	3.0
Boat AP	85.4	100.0	87.8
Kayak AP	85.0	87.8	88.7
Mean AP (mAP)	65.0	68.7	59.8

This may be due to the students being specifically trained on how to use the LabelImg tool and on how tightly the bounding boxes should be drawn around objects. It was observed that each human annotator drew bounding boxes with varying degrees of tightness around each object, which is a factor in calculating IoU and therefore whether a detection bounding box is considered to match a ground truth bounding box (Fig. 4). Low beluga AP values for each label set comparison (Table 2) highlight the inherent difficulty in manually identifying belugas, particularly from oblique imagery, as water disturbances or sun glint can often be misclassified as a beluga, multiple belugas can be mistaken as a single beluga if they are travelling in pods, and kayak paddles can sometimes be misidentified as belugas (Fig. 4).

Fig. 4. Examples of label bounding boxes generated by M. Harasyn (ground truth), E. Ausen, and the two students. The following labeling discrepancies are shown in each image set: (A) differences in how constrained bounding boxes were drawn; (B) discrepancies in labeling belugas near the surface; (C) beluga traveling in a pod labeled as a single beluga and; (D) kayak paddles being labeled as a beluga and discrepancies in including kayak paddles within bounding boxes (E. Ausen).



Comparing YOLOv4 AP values to the inter-observer agreement calculated between the ground truth and the three human annotators, YOLOv4 predictions agree more closely with our ground truth with a mAP of 85.2%, in comparison to the highest inter-observer agreement mAP of 68.7% among the three human annotators. It is notable that YOLOv4 achieves an AP for belugas of 61.17%, whereas the highest inter-observer agreement beluga AP is 24.67%. These results suggest that YOLOv4 can reduce errors introduced by bias or fatigue in beluga labeling by human annotators.

3.2. Object tracking

3.2.1. Model performance

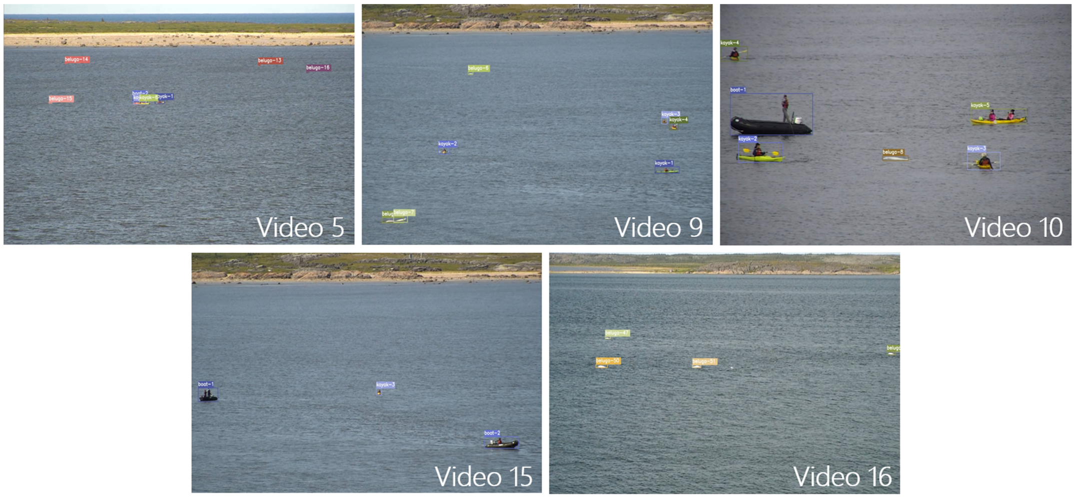
Due to the oblique perspective of the videos, we found that individual belugas could not be reliably tracked once they are fully submerged. Therefore, we treat each event where a beluga appears above water as an individual beluga. For videos containing both belugas and watercraft (video 5, 9 and 10), the DeepSORT tracking model achieves a MOTA between 37.4% and 87.9% and a MOTP between 63.3% and 86% (Table 3). Considering videos containing only boats and kayaks (video 15; Table 3), DeepSORT achieves a MOTA of 98.7% and MOTP of 82.6%, indicating that the trajectories of watercraft are well-identified, and the model can determine the precise location of these objects. DeepSORT achieves a low MOTA (12.2%) when applied to a video containing only belugas (video 16), which is influenced by the short duration belugas are present within the frame. Most videos have a high

Table 3. Multiple object tracking accuracy (MOTA) and precision (MOTP) for DeepSORT outputs of five different drone videos (Figure 5). Number of identified belugas, boats and kayaks provided for the ground truth label set (manually identified).

Video ID	Video length (s)	No. belugas	No. boats	No. kayaks	FP	FN	MOTA (%)	MOTP (%)
5	30	65	1	3	1442	2230	37.4	74.0
9	30	12	0	5	692	412	73.7	72.2
10	19	5	1	4	100	277	87.9	86.0
15	10	0	2	1	12	0	98.7	82.6
16	10	28	0	0	306	724	12.2	63.3

Note: false positives (FP) and false negatives (FN) are counted for each frame of the video and summed.

Fig. 5. Screenshots of the DeepSORT output for videos used in metrics calculations. Videos were chosen for metrics calculations which exhibit a variety of camera zoom configurations, water conditions, and number of objects present.



false negative to false positive ratio which indicates that under-identification is occurring in DeepSORT tracking.

Accuracy of the DeepSORT model in simultaneously tracking belugas and watercraft is dependent on the number of objects in the video, and the size of the features within the video (relating to zoom). Videos 5, 9, and 10 sequentially increase in zoom and decrease in total number of objects present within the video (Fig. 5). MOTA increases as the complexity of the video decreases (increase in size of objects, decrease in number of objects). MOTA and MOTP have a negative correlation with the number of belugas present in the video, related to the magnitude of belugas present in each video (MOTA: R^2 = 0.51, MOTP: R^2 = 0.22).

DeepSORT tracking results fall within the range of MOTA and MOTP scores from inter-observer agreement analysis, indicating that DeepSORT can track beluga and watercraft with a similar accuracy to human annotators. Interestingly, DeepSORT provides higher tracking metrics for the least complex video (video 10) in comparison to inter-observer agreement metrics (Table 4), showing that deep learning algorithms can replicate ground truth labels for simple videos with an accuracy higher than human annotators.

Table 4. Multiple-object tracking accuracy (MOTA) and precision (MOTP) for two human annotators, compared to the ground truth dataset.

				MOTA (%)		MOTP (%)	
Video	No. belugas	No. boats	No. kayaks	W. Chan	E. Ausen	W. Chan	E. Ausen
5	65	1	3	42.5	26.4	70.1	67.0
10	5	1	4	83.3	82.2	80.4	83.6
16	28	0	0	33.0	−51.8	71.1	58.4

3.2.2. Human performance

Inter-observer agreement was calculated for three videos of varying complexity (zoom and number of objects present; Fig. 5, Table 4). Video 5 contained 65 belugas, 1 boat and 3 kayaks, and was at a low zoom. Video 10 contained 5 belugas, 1 boat and 4 kayaks, and was at a high zoom. Video 16 contained 28 belugas, no watercraft, and was at a medium zoom.

Inter-observer agreement for tracking is high between the ground truth and subset labels for the three test videos (Table 4). The MOTA and MOTP for both label subsets agree between videos, with more complex videos (i.e., more objects, lots of belugas) having lower tracking metrics. Despite video 5 containing the highest number of belugas, the presence of watercraft in the frame increases the tracking metrics, as watercraft are present in all frames (900 frames) whereas each beluga is present in approximately 30 frames. Video 16 has the lowest MOTA for both label subsets, indicating that there is a large error associated with manually tracking belugas. The negative MOTA statistic for E. Ausen's label set for video 16 is a result of a high number of false positives, resulting from E. Ausen's label set containing more beluga identifications than the ground truth. This is due to the inherent difficulty in identifying and labeling belugas, demonstrated by results from the inter-observer agreement for detection (Table 2), along with the subjectivity associated with identifying the first frame where a beluga is above or below water. MOTP scores above 50% indicate that the placement of bounding boxes agrees between datasets; however, varying MOTA scores indicate that the presence/absence or duration of object tracks are in varying agreement between datasets.

4. Discussion and conclusions

4.1. Potential for deep learning use in marine mammal research

Accuracies achieved in this research support the use of deep learning techniques as an assistive technology in marine mammal research. The YOLOv4 detection model predicts ground truth labels to an overall mAP of 85.2% (precision = 89%, recall = 92%), which is significantly higher than the agreement between human-annotated labels (maximum mAP of 68.67%). These results indicate that with a reliable training dataset, YOLOv4 can identify beluga and watercraft in a fraction of the time, and reduce errors and discrepancies introduced by multiple human annotators or fatigue. One factor that may impact model detection is the size of the training dataset. It is recommended to train YOLOv4 with at least 2000 images per object class (Bochkovskiy 2021). Due to the laborious process of annotating images for ground truth, we were only able to train the model on a total of 721 images. Future work should consider training a similar computer vision framework on a larger training dataset.

The AP values only represent detections which have 50% or greater IoU between the ground truth box and the YOLOv4 prediction box (Padilla et al. 2021). An IoU of 50% was chosen as it is the standard matching overlap percentage used in previous research (Everingham et al. 2010). However, this value can be adjusted based on the nature of the task. Most of the mismatches in our dataset have an IoU of 40% or greater, which appears to the human eye as a correct detection of an individual beluga (Fig. 3). If the IoU were to be decreased to 40%, beluga AP would increase to 76.3%. This indicates that YOLOv4 is significantly more applicable to marine mammal detection than our AP results may suggest.

When comparing our results to previous research, it is important to consider the type of data being used. Marine mammal detection is often completed using overhead nadir (overhead) imagery, allowing for greater differentiation between animals traveling in pods. Guirado et al. (2019) and Borowicz et al. (2019) trained CNNs on satellite imagery of whales (15 cm – 6 m resolution). Guirado et al. (2019) achieved an F1 value of 0.94 for detecting and counting individual whales, whereas Borowicz et al. (2019) achieved a precision of 100% and recall of 93.7% for identifying images where a whale is present. Considering oblique imagery, Chalmers et al. (2021) reported a mAP of 83% for detecting both rhinos and cars within oblique 20 MP drone imagery, indicating that our results are on par with other models trained on a similar dataset type.

In regard to tracking belugas and watercraft using the DeepSORT model, our results indicate that its performance is competitive with manual feature tracking. DeepSORT tracking metrics fell within the range of inter-observer agreement tracking metrics, demonstrating that deep learning tracking algorithms can replicate a ground truth label set with a similar degree of error as that introduced by multiple humans labeling a dataset. DeepSORT was successful at tracking kayaks and boats (multiple-object tracking accuracy; MOTA = 98.7%) and precisely locating them (multiple-object tracking precision; MOTP = 82.6%). The main source of error in watercraft tracking was the loss of a track while vessels were perfectly overlapping, however DeepSORT effectively tracked vessels even while partially overlapping (Fig. 5). Application of DeepSORT to videos containing only belugas resulted in a lower MOTA score (12.2%; Table 3), influenced by the short duration of belugas being on screen (~1 s). Belugas appearing and disappearing frequently on screen allows for error between the ground truth and DeepSORT to be propagated, as mismatches between the frame where each beluga appears and disappears can easily be defined differently between each dataset. For example, if DeepSORT detected all belugas 10 frames (0.33 s) after they are labeled in the ground truth for video 5, 650 false negatives would be applied to the MOTA score. This is demonstrated by the low inter-observer agreement accuracies for video 16 (Table 4; MOTA = 33% and −51.8% respectively, for the two human annotators), which shows that belugas are inherently difficult to track from an oblique view.

Similar computer vision tracking algorithms have been applied to oblique wildlife video in previous research. Bondi et al. (2020) applied existing tracking algorithms to aerial thermal infrared video of humans and elephants, achieving a single-object tracking precision of 48%, and a MOTA of 61.6% and MOTP of 100% (for small-sized objects). Ratnayake et al. (2021) developed a deep-learning program to track the movement of honeybees within a tripod-mounted camera video to a rate of 86.6%. Computer vision tracking has also been applied to underwater video of fish in Blowers et al. (2020), who were able to track fish across a controlled background to a precision of 84% and recall of 73%. The highest MOTA (87.9%) and corresponding MOTP (86.0%) achieved in our study for videos containing both belugas and watercraft is comparable to values reported in previous research. Limitations in beluga tracking for our research is linked to the short duration of time belugas are visible above water, providing a limited reference of the trajectory of belugas for DeepSORT to base predictions on. Beluga tracking accuracy would likely increase if tracking algorithms were applied to a dataset where belugas remain in frame for a long duration, such as nadir video where belugas can be seen while surfaced or at a shallow depth.

The challenging nature of the tracking problem in our study cannot be overstated, and was due to three main factors: (*i*) the very high occlusion rate in which belugas were only visible for a small fraction of the total time, (*ii*) the homogeneous appearance of the belugas in the relatively low resolution drone videos that made it difficult to differentiate individual belugas, and (*iii*) the tendency for belugas to travel in pods, which further declined the ability to track them reliably. If a beluga surfaced briefly, disappeared from view, and then a second beluga appeared nearby, it was difficult to determine if the belugas were in fact a single whale that appeared twice, or whether they were two distinct whales. Compared to other tracking challenges, such as the MOTChallenge benchmarks (https://motchallenge.net/) which are designed to evaluate state-of-the-art tracking algorithms on difficult multiple object tracking datasets, the problem of tracking belugas in oblique video seems to be at least on par — if not more difficult — than these challenge benchmarks, which involve tracking vehicles and pedestrians, for which the highest reported MOTA scores are currently below 80%.

4.2. Future research

The most substantial factor limiting the results of this study is the resolution and perspective of the drone video, resulting from the opportunistic nature of data collection in the field. Both the low resolution of the camera (2 MP) and the oblique perspective of the video made the video frames initially difficult to manually label for the ground truth, resulting in subpar data quality used to train the deep learning algorithm. The low agreement between the four manually labeled datasets (ground truth and the three human-annotated datasets) demonstrates the difficulty in manually labeling the drone dataset. A higher-resolution dataset would increase the accuracy of manual-labeling, hence rendering a more accurate ground truth dataset for training a more accurate detection model.

The oblique perspective of the video led to full occlusion of underwater belugas, limiting the extent to which they could be tracked. With nadir videos, belugas can be tracked while submerged at a shallow depth, allowing for the tracking of belugas over a longer timeframe. This would increase the accuracy of beluga tracking, as tracking algorithms would have a longer reference period of beluga movement on which to base trajectory predictions. High-resolution nadir video would allow for a wider variety of classes to be detected, such as a range in motorized boat sizes (small, medium, large) and age classes of beluga based on colour and length. Nadir videos would also allow for the calculation of the relative distances between belugas and watercraft, using the height and focal length of the camera. In doing so, reliable quantitative population statistics can be automatically derived from drone video, which can be used to study beluga behaviour in the presence and absence of watercraft to better inform shipping traffic policies in the Churchill River estuary, and other biologically significant waterways impacted by vessel traffic.

Further research into the ways in which the human annotators differ in their labeling is recommended, which could help to better understand the types of errors, biases, and subjectivity that may come into play while labeling marine mammals in datasets. It is also recommended that other deep learning object detection algorithms, such as Faster R-CNN (Ren et al. 2017) or RetinaNet (Lin et al. 2020), as well as alternative tracking algorithms be tested on wildlife imagery and videos to determine the optimal framework for wildlife data processing.

The presented research, along with previous work (e.g., Borowicz et al. 2019; Guirado et al. 2019), demonstrates the potential for deep learning methods to be used in marine wildlife research. Automated methods of detecting and tracking marine wildlife and ships will allow for processing large datasets in a fraction of the time, resulting in larger data availability for marine wildlife research. The use of drones to collect aerial imagery/video reduces the risk and cost of data collection, and a consistent GPS location can be maintained while hovering, allowing for a constant video reference frame for calculations of absolute movement of marine wildlife and watercraft. With increase in high-resolution data availability from drones, and the ease of data processing introduced by deep learning algorithms, biological data can become more widely available for this research, leading to an advancement of our knowledge on the behavioural patterns of marine wildlife and how anthropogenic activities are impacting these species.

Acknowledgements

The authors would like to thank the members of the Churchill Port Authority, particularly Randy Spence, who provided us access to the Churchill Port grounds, and provided logistical and technical support throughout our data collection. Thank you to the staff at the Churchill Northern Studies Centre for their hospitality and logistical support during the field season. We would like to acknowledge the time that Casey Clair and Ryan Kum dedicated to labeling our image set for inter-observer comparison, and to Cooper Harasyn for support during code development. We thank Brittany Cook and Joseph Pear for discussions about inter-observer agreement, and Grigory Shamov for assistance with the Grex HPC system. The authors would also like to thank the Editor-in-Chief Dr. Dominique Chabot, and three anonymous reviewers whose comments greatly improved the paper. This project is part of the UM Canada Excellence Research Chair (CERC) program and the Arctic Science Partnership.

Competing interests

The authors declare there are no competing interests.

Author contributions

MH: Conceptualization, methodology, software, writing – original draft, visualization. **WC:** Methodology, software, validation, resources, writing – review & editing. **EA:** Validation, writing – review & editing. **DB:** Supervision, funding, writing – review & editing.

Funding

This research was supported by the Canada Excellence Research Chair (CERC) and the Canada Research Chair (CRC programs).

Data availability statement

All data used for this research have been archived on the Canadian Watershed Information Network (CanWIN; http://lwbin-datahub.ad.umanitoba.ca). Dataset is available from the following direct link: http://lwbin-datahub.ad.umanitoba.ca/dataset/drone.

References

Betke, M., Haritaoglu, E., and Davis, L.S. 2000. Real-time multiple vehicle detection and tracking from a moving vehicle. Mach. Vis. Appl. **12**(2): 69–83. doi: 10.1007/s001380050126.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. 2016. Simple online and realtime tracking. *In* Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, Arizona. pp. 3464–3468. doi: 10.1109/ICIP.2016.7533003.

Blowers, S., Evans, J., and McNally, K. 2020. Automated identification of fish and other aquatic life in underwater video. Scottish Mar. Freshw. Sci. **11**(18): 1–62. doi: 10.7489/12333-1.

Bochkovskiy, A. 2021. GitHub. Available from https://github.com/AlexeyAB/darknet [accessed 8 September 2019].

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y.M. 2020. YOLOv4: Optimal speed and accuracy of object detection. Available from http://arxiv.org/abs/2004.10934.

Bogucki, R., Cygan, M., Khan, C.B., Klimek, M., Milczek, J.K., and Mucha, M. 2019. Applying deep learning to right whale photo identification. Conserv. Biol. **33**(3): 676–684. doi: 10.1111/cobi.13226. PMID: 30259577.

Bondi, E., Jain, R., Aggrawal, P., Anand, S., Hannaford, R., Kapoor, A., Piavis, J., et al. 2020. BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos. *In* Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020. 1–5 March 2020, Aspen, CO. pp. 1736–1745. doi: 10.1109/WACV45572.2020.9093284.

Borowicz, A., Le, H., Humphries, G., Nehls, G., Höschle, C., Kosarev, V., and Lynch, H.J. 2019. Aerial-trained deep learning networks for surveying cetaceans from satellite imagery. PLoS One, **14**(10): e0212532. doi: 10.1371/journal.pone.0212532.

Bridge, E.S., Wilhelm, J., Pandit, M.M., Moreno, A., Curry, C.M., Pearson, T.D., et al. 2019. An Arduino-based RFID platform for animal research. Frontiers in Ecology and Evolution, **10**: 1–10. doi: 10.3389/fevo.2019.00257.

Burgos-Artizzu, X.P., Dollar, P., Lin, D., Anderson, D.J., and Perona, P. 2012. Social behavior recognition in continuous video. *In* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 16–21 June 2012, Providence, RI. pp. 1322–1329. doi: 10.1109/CVPR.2012.6247817.

Caron, L., and Smith, T. 1990. Philopatry and site tenacity of belugas, *Delphinapterus leucas*, hunted by the Inuit at the Nastapoka estuary, eastern Hudson Bay. *In* Advances in research on the Beluga Whale, *Delphinapterus leucas*, 224th edition. *Edited by* T. Smith, D. St. Aubin, and J. Geraci. Canadian Bulletin of Fisheries and Aquatic Sciences, Ottawa. pp. 69–79.

Cartucho, J., Ventura, R., and Veloso, M. 2018. Robust object recognition through symbiotic deep learning in mobile robots. *In* Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1–5 October 2018, Madrid, Spain. pp. 2336–2341.

Chambault, P., Dalleau, M., Nicet, J.B., Mouquet, P., Ballorain, K., Jean, C., et al. 2020. Contrasted habitats and individual plasticity drive the fine scale movements of juvenile green turtles in coastal ecosystems. Mov Ecol, **8**(1): 1–15. doi: 10.1186/s40462-019-0184-2.

Citta, J.J., Suydam, R.S., Quakenbush, L.T., Frost, K.J., and O'Corry-Crowe, G.M. 2013. Dive behavior of eastern Chukchi beluga whales (*Delphinapterus leucas*), 1998–2008. Arctic, **66**(4): 389–406. doi: 10.14430/arctic4326.

COSEWIC. 2004. COSEWIC assessment and update status report on the Beluga Whale *Delphinapterus Leucas* in Canada. Committee on the Status of Endangered Wildlife in Canada (COSEWIC), Ottawa, ON.

DJI. 2018. M210 RTK Quadcopter. Shenzhen, China.

Doniol-Valcroze, T., Gosselin, J.-F., and Hammill, M.O. 2013. Population modeling and harvest advice under the precautionary approach for eastern Hudson Bay beluga (*Delphinapterus leucas*). DFO Can. Sci. Advis. Sec. Res. Doc. 2012, 168: 1–31.

Erbe, C., Reichmuth, C., Cunningham, K., Lucke, K., and Dooling, R. 2016. Communication masking in marine mammals: A review and research strategy. Mar. Pollut. Bull., **103**(1–2): 15–38. doi: 10.1016/j.marpolbul.2015.12.007. PMID: 26707982.

Everingham, M., van Gool, L., Williams, C.K.I., Winn, J., and Zisserman, A. 2010. The Pascal visual object classes (VOC) challenge. Int. J. Comput. Vis., **88**(2): 303–338. doi: 10.1007/s11263-009-0275-4.

Ferguson, M.C., Angliss, R.P., Kennedy, A., Lynch, B., Willoughby, A., Helker, V., et al. 2018. Performance of manned and unmanned aerial surveys to collect visual data and imagery for estimating arctic cetacean density and associated uncertainty. J. Unman. Veh. Sys. **6**(3): 128–154. doi: 10.1139/juvs-2018-0002.

Giampiccoli, A., Mtapuria, D.O., and Jugmohan, S. 2020. Community-based tourism and animals: Theorising the relationship. Cogent Soc. Sci. **6**(1). doi: 10.1080/23311886.2020.1778965.

Giancardo, L., Sona, D., Huang, H., Sannino, S., Managò, F., Scheggia, D., et al. 2013. Automatic visual tracking and social behaviour analysis with multiple mice. PLoS ONE, **8**(9): e74557. doi: 10.1371/journal.pone.0074557.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. Deep learning. MIT Press, Cambridge, USA.

Grex Technical Specifications. 2018. Available from https://www.westgrid.ca/support/systems/Grex [accessed 8 September 2020].

Guirado, E., Tabik, S., Rivas, M.L., Alcaraz-Segura, D., and Herrera, F. 2019. Whale counting in satellite and aerial images with deep learning. Sci Rep, **9**(1): 1–12. doi: 10.1038/s41598-019-50795-9.

Hastie, G.D., Wu, G.M., Moss, S., Jepp, P., MacAulay, J., Lee, A., et al. 2019. Automated detection and tracking of marine mammals: A novel sonar tool for monitoring effects of marine industry. Aquat. Conserv. Mar. Freshwater Ecosyst., **29**(S1): 119–130. doi: 10.1002/aqc.3103.

Hauser, D.D.W., Laidre, K.L., Suydam, R.S., and Richard, P.R. 2014. Population-specific home ranges and migration timing of Pacific Arctic beluga whales (*Delphinapterus leucas*). Polar Biol., **37**(8): 1171–1183. doi: 10.1007/s00300-014-1510-1.

Heindl, C., Toka, and Valmadre, J. 2020. py-motmetrics. Available from www.github.com/cheind/py-motmetrics [accessed 8 September 2019].

Hodgson, A., Kelly, N., and Peel, D. 2013. Unmanned aerial vehicles (UAVs) for surveying Marine Fauna: A dugong case study. PLoS ONE, **8**(11): e79556–16. doi: 10.1371/journal.pone.0079556. PMID: 24223967.

Hong, W., Kennedy, A., Burgos-Artizzu, X.P., Zelikowsky, M., Navonne, S.G., Perona, P., and Anderson, D.J. 2015. Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. Proc. Natl. Acad. Sci. United States Am. **112**(38): E5351–E5360. doi: 10.1073/pnas.1515982112.

Johnston, J.M., Pennypacker, H.S., and Green, G. 2020. Strategies and Tactics of behavioral research and practice. Routledge, New York.

Kelley, T.C., Loseto, L.L., Stewart, R.E.A., Yurkowski, M., and Ferguson, S.H. 2010. Importance of Eating Capelin: Unique Dietary Habits of Hudson Bay Beluga. *In* A little less Arctic: Top predators in the world's largest Northern Inland Sea, Hudson Bay. *Edited by* S.H. Ferguson, L.L. Loseto, and M.L. Mallory. Dordrecht: Springer Netherlands, New York. doi: 10.1007/978-90-481-9121-5.

Koski, W.R., Gamage, G., Davis, A.R., Mathews, T., Leblanc, B., and Ferguson, S.H. 2015. Evaluation of UAS for photographic re-identification of bowhead whales, *Balaena mysticetus*. J. Unman. Veh. Sys. **3**(1): 22–29. doi: 10.1139/juvs-2014-0014.

Krasnova, V., Prasolova, E.A., Belikov, R.A., Chernetsky, A.D., and Panova, E.M. 2020. Influence of boat tourism on the behaviour of Solovetskiy beluga whales (*Delphinapterus leucas*) in Onega Bay, the White Sea. Aquat. Conserv. Mar. Freshwater Ecosyst. **30**(10): 1922–1933. doi: 10.1002/aqc.3369.

Lin, T.Y., Goyal, P., Girshick, R., He, K., and Dollar, P. 2020. Focal loss for dense object detection. IEEE PAMI. **42**(2): 318–327. doi: 10.1109/TPAMI.2018.2858826.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. 2015. Microsoft COCO: common objects in context. *In* Proceedings of the European Conference on Computer Vision. Zurich, Switzerland. pp. 1–15. doi: 10.1007/978-3-319-10602-1_48.

Lowry, L.F., Kingsley, M.C.S., Hauser, D.D.W., Clarke, J., and Suydam, R. 2017. Aerial survey estimates of abundance of the eastern Chukchi Sea stock of beluga whales (*Delphinapterus leucas*) in 2012. Arctic, **70**(3): 273–286. doi: 10.14430/arctic4667.

Malcolm, C.D., and Penner, H.C. 2011. Behavior of Belugas in the presence of whale-watching vessels in Churchill, Manitoba and recommendations for local Beluga-watching activities. *In* Polar Tourism: Human, Environmental and Governance Dimensions. *Edited by* P. Maher, E. Stewart, and M. Lück. Cognizant Communications, Putnam Valley, NY. pp. 54–79.

Manabe, K. 2017. The Skinner box evolving to detect movement and vocalization. Revista Mexicana de Analisis de la Conducta, **43**(2): 192–211. doi: 10.5514/rmac.v43.i2.62313.

Mangott, A., Birtles, R., and March, H. 2011. Attraction of dwarf minke whales Balaenoptera acutorostrata to vessels and swimmers in the Great Barrier Reef World Heritage Area – The management challenges of an inquisitive whale. J. Ecotour. **10**: 64–76. doi: 10.1080/14724041003690468.

Manitoba's Western Hudson Bay Ad Hoc Beluga Habitat Sustainability Plan Committee. 2016. Manitoba's Beluga habitat sustainability plan. Manitoba Conservation and Water Stewardship, Winnipeg, Manitoba. pp. 1–30.

Matthews, C.J.D., Watt, C.A., Asselin, N.C., Dunn, J.B., Young, B.G., Montsion, L.M., et al. 2016. Estimated abundance of the Western Hudson Bay beluga stock from the 2015 visual and photographic aerial survey. Fisheries and Oceans Canada, Canadian Science Advisory Secretariat Research Document, Ottawa, ON. pp. 1–25.

Menet Aero. 2017. NTP powerline tether. Ringwood, USA.

Milan, A., Leal-Taixe, L., Reid, I., Roth, S., and Schindler, K. 2016. MOT16: a benchmark for multi-object tracking. Available from http://arxiv.org/abs/1603.00831.

Motro, M., and Ghosh, J. 2018. Measurement-wise occlusion in multi-object tracking. *In* Proceedings of the 2018 21st International Conference on Information Fusion, FUSION 2018. pp. 2384–2391. doi: 10.23919/ICIF.2018.8455339.

Open VINO Toolkit. 2021. CVAT. Available from www.github.com/openvinotoolkit/cvat [accessed 8 January 2021].

Padilla, R., Passos, W.L., Dias, T.L.B., Netto, S.L., and da Silva, E.A.B. 2021. A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics (Switzerland), **10**(3): 1–28. doi: 10.3390/electronics10030279.

Papadourakis, V., and Argyros, A. 2010. Multiple objects tracking in the presence of long-term occlusions. Comput. Vis. Image Underst., **114**(7): 835–846. Elsevier Inc. doi: 10.1016/j.cviu.2010.02.003.

Pear, J.J. 1985. Spatiotemporal patterns of behavior produced by variable-interval schedules of reinforcement. J Exp Anal Behav, **44**(2): 217–231. doi: 10.1901/jeab.1985.44-217. PMID: 16812432.

Ratnayake, M.N., Dyer, A.G., and Dorin, A. 2021. Tracking individual honeybees among wildflower clusters with computer vision-facilitated pollinator monitoring. PLoS ONE, **16**(2): e0239504–20. doi: 10.1371/journal.pone.0239504. PMID: 33571210.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. 2016. You only look once: Unified, real-time object detection Joseph. IEEE. doi: 10.1021/je00029a022.

Ren, S., He, K., Girshick, R., and Sun, J. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell., **39**(6): 1137–1149. doi: 10.1109/TPAMI.2016.2577031. PMID: 27295650.

Robbins, J., Zerbini, A.N., Gales, N., Gulliand, F.M.D., Double, M., Clapham, P.J., et al. 2013. Satellite tag effectiveness and impacts on large whales: preliminary results of a case study with Gulf of Maine humpback whales. *In* Scientific Committee of the International Whaling Commission, 3–15 June 2013, Seogwipo-si, Jeju-do, Korea. pp. 1–10.

Robie, A.A., Seagraves, K.M., Egnor, S.E.R., and Branson, K. 2017. Machine vision methods for analyzing social interactions. J. Exp. Biol., **220**(1): 25–34. doi: 10.1242/jeb.142281.

Rolland, R.M., Parks, S.E., Hunt, K.E., Castellote, M., Corkeron, P.J., Nowacek, D.P., et al. 2012. Evidence that ship noise increases stress in right whales. Proc. R. Soc. B: Biol. Sci. **279**(1737): 2363–2368. doi: 10.1098/rspb.2011.2429.

Stafford, K., Ferguson, M., Hauser, D., Okkonen, S., Berchok, C., Citta, J., et al. 2018. Beluga whales in the western Beaufort Sea: Current state of knowledge on timing, distribution, habitat use and environmental drivers. Deep Sea Res. II, **152**: 182–194. doi: 10.1016/j.dsr2.2016.11.017.

The AI Guy. 2017. YOLOv4-DeepSORT. Available from www.github.com/theAIGuysCode/yolov4-deepsort [accessed 8 September 2019].

Torres, L.G., Nieukirk, S.L., Lemos, L., and Chandler, T.E. 2018. Drone up! Quantifying whale behavior from a new perspective improves observational capacity. Front. Mar. Sci., **5**: 1–14. doi: 10.3389/fmars.2018.00319.

Tyrrell, M., Idle, P.D., Doniol-Valcroze, T., Gosselin, J.-F., and Hammill, M.O. 2007. Population modeling and harvest advice under the precautionary approach for eastern Hudson Bay beluga (*Delphinapterus leucas*). Human Ecol. **35**(5): 575–586. doi: 10.1007/s10745-006-9105-2.

Tzutalin. 2015. LabelImg. Available from www.github.com/tzutalin/labelImg [accessed 8 September 2019].

van Harten, E., Reardon, T., Lumsden, L.F., Meyers, N., Prowse, T.A.A., Weyland, J., and Lawrence, R. 2019. High detectability with low impact: Optimizing large PIT tracking systems for cave-dwelling bats. Ecol Evol, **9**(19): 10916–10928. doi: 10.1002/ece3.5482. PMID: 31641445.

Walker, K.A., Trites, A.W., Haulena, M., and Weary, D.M. 2012. A review of the effects of different marking and tagging techniques on marine mammals. Wildlife Res. **39**(1): 15–30. doi: 10.1071/WR10177.

Wang, D., Shao, Q., and Yue, H. 2019. Surveying wild animals from satellites, manned aircraft and unmanned aerial systems (UASs): A review. Rem. Sens. **11**(11): 1308 doi: 10.3390/rs11111308.

Williams, R., Trites, A.W., and Bain, D.E. 2002. Behavioural responses of killer whales (Orcinus orca) to whale-watching boats: Opportunistic observations and experimental approaches. J. Zool., **256**(2): 255–270. doi: 10.1017/S0952836902000298.

Wojke, N., Bewley, A., and Paulus, D. 2018. Simple online and realtime tracking with a deep association metric. *In* Proceedings of the International Conference on Image Processing, ICIP. pp. 3645–3649. doi: 10.1109/ICIP.2017.8296962.