# Marine Ecology Enhancement Fund (MEEF)

# Declaration

To: The Secretariat of the MEEF

**Reference No.**: MEEF2020006

**Project Title**: Automated CWD Detection with Machine Learning for Vessel-based Line-transect Survey

**Name of the Project Leader**: Prof. Antoni B. Chan

I hereby irrevocably declare to the MEEF Management Committee and the Steering Committee of the relevant Funds including the Top-up Fund, that all the dataset and information included in the completion report has been properly referenced, and necessary authorisation has been obtained in respect of information owned by third parties.

Signature: ________________
*Project Leader, Antoni B. Chan*

Date: ___2022/May/26______

# Marine Ecology Enhancement Fund (MEEF)

# Completion Report for Project MEEF2020006

Any opinions, findings, conclusions or recommendations expressed in this report do not necessarily reflect the views of the Marine Ecology Enhancement Fund or the Trustee.

## 1 Executive Summary

The project entitled "Automated CWD Detection with Machine Learning for Vessel-based Line-transect Survey" has been finished. The four primary objectives are:

1. To propose an automated detection framework for recognizing Chinese White Dolphins (CWD) captured by video cameras.
2. To build a prototype based on the proposed detection framework.
3. To embed human expertise into the prototype.
4. To improve the effectiveness of the survey by performing simultaneous and continuous detection on a 180-degree field of view.

The project started in 1 July 2020 and ended in 31 December 2021. Through the 18 months, the realization of the 4 project objectives is illustrated as follows.

*Objectives 1 & 2)* This is the first work specially targeted at automatic CWD detection in the sea using modern deep learning methodology. We propose to install video cameras on the survey boat and rely on deep learning-based object detection methods for automatic dolphin detection. To train the model for automatic dolphin detection, we first join boat survey trips from October 2020 to October 2021 to collect dolphin images. To prepare the dataset, we design an interactive dolphin box annotation strategy to annotate sparse dolphin instances in long videos efficiently. The strategy firstly adopts a coarse dolphin detector to generate candidate bounding-boxes and then display them to human annotators for further refinement. Equipped with the strategy, we construct a "Dolphin dataset" with more than 2.6k dolphin instances (surfacing dolphins). Thus, these objectives are completed.

We compare the performance and efficiency of three off-the-shelf object detection algorithms, including FasterRCNN [24], FCOS [27], and YoloV5 [4], on the Dolphin dataset. Considering both running speed and performance, we adopt the model using YoloV5 as the dolphin detection algorithm in the framework. At last, we incorporate the dolphin detector into the system prototype, which receives the video stream from a GoPro camera, detects dolphins in video frames at 100.99 FPS on a standard desktop GPU (Nvidia RTX2080Ti) with high accuracy (i.e., 90.95 mean average precision; mAP@0.5), and then notifies users of detection results in real-time.

*Objective 3)* We have embedded expert knowledge into our system in the following two ways. First, one of the key steps in building the prototype is to train the detector with samples in which sighted dolphins are labelled with the help of human experts so that, after training, the detector can automatically identify features from the examples for recognizing the dolphins. The detector implements a machine learning method based on artificial neural networks (NNs), such as convolutional neural networks (CNNs), which use multiple layers to progressively extract higher level features from input images. Second, since there are many distractors with similar appearance to CWD (e.g., white plastic garbage), we have also added a specific distractor class to the detector so that the NN can better discriminate between CWD and other objects. Thus, this objective is completed.

*Objective 4)* The main goal of this objective was to show that the system can achieve acceptable accuracy in the field while performing simultaneous and continuous detection over a large FOV. For the field tests, the prototype system is built on two notebook PCs, each analyzing one video camera stream in real-time. In the 2-camera prototype system, interactive operations are implemented to: i) record the dolphin appearing time when a dolphin expert spots a dolphin; ii) judge and record whether a detection made by the system is a true dolphin example or just a false positive. These operations help validate that the system output is consistent with human dolphin experts. In the field test, the two-camera system cover a 170-degree field of view (FOV) and perform simultaneous and continuous detection. The FOV could be increased by simply further rotating the cameras away from each other. Since each camera has an FOV of 140 degrees, the theoretical FOV for two-camera system is 280 degrees. We did not measure any associated data, e.g., angle to transect line or distance to camera. This associated data can be calculated in a straightforward way using the detection location and the camera geometry – a good detector is a prerequisite for extracting the associated data. See details in Section 16a3. Finally, note that we did not test embedding the detection system into an actual line-transect survey to see whether it can improve the effectiveness of the human team, as this was not our original intention in objective 4. This is interesting future work, and please see the details in Section 16a5.

In summary, we have fully met the first 3 objectives, and partially fulfilled the 4th objective as we did not measure the associate data (angle and distance). Future work can consider adding the ability to calculate associated data such as angle and camera distance. Moving forward, it will be interesting to incorporate the system into line-transect surveys to improve efficiency/effectiveness of the human survey team. We have outlined some possible future works in Section 16a.

## 2 Project Title

Automated CWD Detection with Machine Learning for Vessel-based Line-transect Survey

## 3 Project Period

From 1 July 2020 to 31 December 2021

## 4 Nature of the Project

□ Marine Habitat & Resource Conservation & Enhancement

☑ Scientific Research & Studies

□ Environmental Education & Eco-tourism

## 5 Brief description of the Purpose of the Project

For ecological protection of the ocean, biologists usually conduct line-transect vessel surveys to measure a species' population density within their habitat. Specifically, the survey area is rigidly divided by parallel lines, and one vessel sails along the line and records instances of the species for each line. A system for automatic detection of Chinese White Dolphins (CWDs) has potential applications in line-transect surveys to both reduces human experts' workload and improve overall accuracy. However, dolphin detection in the wild is much more challenging than detection of other common objects. The reasons lie in three aspects: (1) the scarcity of data in the wild, (2) varying outdoor conditions, and (3) tiny/blur object instances in images.

In this project, we develop a practical system to detect Chinese White Dolphins in the wild automatically, with an ultimate goal to assist human dolphin experts in vessel surveys. To prepare the dataset, we design an interactive dolphin box annotation strategy to annotate sparse dolphin instances in long videos efficiently. The strategy firstly adopts a coarse dolphin detector to generate candidate bounding-boxes and then displays them to human annotators for further refinement. Equipped with the strategy, we construct a "Dolphin dataset" with more than 2.6k dolphin instances. Later, we compare the performance and efficiency of three off-the-shelf object detection algorithms, including Faster-RCNN, FCOS, and YoloV5, on the Dolphin dataset and finally adopt YoloV5 as the detector. At last, we incorporate the dolphin detector into a system prototype, which receives the video stream from a GoPro camera, detects dolphins in video frames at 100.99 FPS per RTX2080Ti GPU with high accuracy (i.e., 90.95 mAP@0.5), and then notifies users of the detection results in real time. We ran 2 field tests on our system prototype and successfully show that the system can detect dolphins during surveys with a 170 degree field-of-view (FOV).

# 6 Investigator(s) and Academic Department/Units Involved

| **Research Team** | **Name / Post** | **Unit / Department / Institution** |
|---|---|---|
| Principal Investigator | Prof. Antoni B. Chan | Department of Computer Science, City University of Hong Kong |
| Co-investigator | Prof. Victor C. S. Lee | Department of Electrical and Electronic Engineering, The University of Hong Kong |
| Member | Dr. Qi Zhang | Department of Computer Science, City University of Hong Kong |
| Member | Dr. Hao Zhang | Department of Computer Science, City University of Hong Kong |
| Member | Dr. Phuong Anh Nguyen | Department of Computer Science, City University of Hong Kong |

Table 1. The Investigators and members of the project.

*Remark:* During the MEEF assessment meeting, a suggestion was made for us to include a dolphin expert on our team. *Please note that this suggestion was not shared with us as confirmed by the ERM Secretariat*. Thus the composition of the team remained the same throughout the project. Nonetheless, all the data collection and field test were conducted together with dolphin experts.

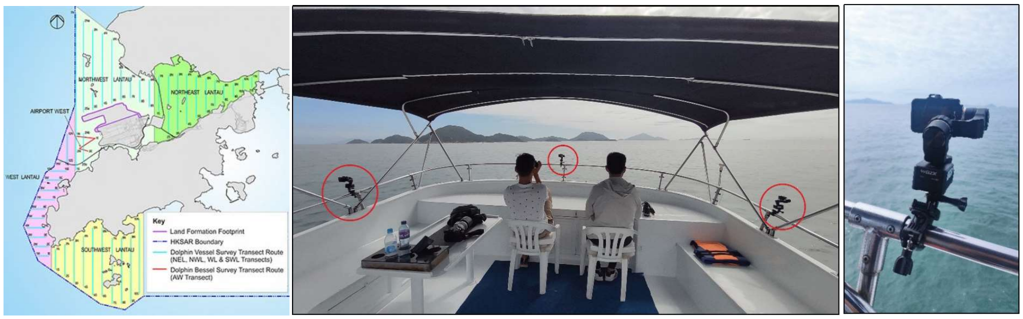
# 7 Completed Activities against the Proposed Work Schedule

| Activities | Proposed Period (Extended) | Actual Period | Progress | Explanation for Deviation |
|---|---|---|---|---|
| 1. Data collection | Nov 2020 – Jun 2021 | Oct 2020 – Nov 2021 | Completed | The data collection was period was extended to collect more data for training and testing the model in the experiments. |
| 2. Framework design | Jul 2020 – Aug 2020 | Jul 2020 – Aug 2020 | Completed | |
| 3. Prototype implementation | Sep 2020 – Jul 2021 | Sep 2020 – Nov 2021 | Completed | The implementation and experiment period were extended to improve the performance of the model and to thoroughly evaluate on the newly collected data. |
| 4. Experiments | Aug 2021 – Oct 2021 | Aug 2021 – Nov 2021 | Completed | |
| 5. Field tests | Nov 2021 – Dec 2021 | Dec 2021 | Completed | The field test was delayed until the prototype was implemented and UI designed for the field test. |
| 6. Seminar | Dec 2021 | Dec 2021 | Completed | |

Table 2. Completed activities against the proposed work schedule.

# 8 Detailed Project Introduction

Since 2016, the Hong Kong International Airport (HKIA) has been constructing a third runway to fill up the required capacity for future air traffic [1]. The construction plan is to add 640 hectares of artificial land to the north of Chek Lap Kok island in the Lantau area. This construction has been causing construction-related disturbances to the surrounding ecosystem. These disturbances - including noise, land reshaping, and increasing water traffics - affect the distribution and behavior of marine mammals [19]. HKIA has been actively conducting research projects to study the impact of the construction disturbances on marine mammals, particularly the Chinese White Dolphins (also known as the Indo-Pacific Humpback Dolphin).


NORTHWEST LANTAU
NORTHEAST LANTAU
AIRPORT WEST
WEST LANTAU
SOUTHWEST LANTAU
Key
Land Formation Footprint
HKSAR Boundary
Dolphin Vessel Survey Transect Route (NEL, NWL, WL & SWL Transects)
Dolphin Bessel Survey Transect Route (AW Transect)

(a) (b) (c)

Figure 1: Line-transect survey methodology and data collection system: (a) line-transect survey map, (b) the camera locations, and (c) camera setup for data collection.

Researching marine mammals requires fieldwork, in which the researchers can observe, approach, and collect data using vessels. The fieldwork follows the line transect survey methodology [5]. In brief, the survey area is split into parallel lines, and the vessel is instructed to follow these lines during the survey (see Fig. 1a). The survey team consists of 3 or 4 observers. A pair of observers watches the water surface to detect marine mammals and record the sighting experiences, while the other observers rest. Each observing split lasts 30 minutes and requires two observers, one using binocular and one using unaided eyes, to cover a 180° field of view (FOV) in front of the vessel (see Fig. 1b). This observation work is demanding, with observers rotating every 30 minutes to prevent fatigue. A survey trip takes an average 5-6 hours of non-stop observing, depending on the survey area.

The FOV of the human eyes is about 190 degrees, which contains both central vision (60 degrees FOV around the center line) and peripheral vision (the remaining 130 degrees on the outside). Peripheral vision has low visual acuity and is mainly sensitive to large motions. Dolphin appearing in peripheral vision are too small to be noticed, and thus the observer needs to move their heads around for scanning using central vision (60 degrees), and only the central 5 degrees has very high visual acuity for perceiving small objects (e.g. reading text). Thus, a limitation of the human observer is that they need to actively scan with central vision, and thus may miss some dolphins that appear outside of the central vision FOV. Although the line transect methodology based on distance sampling has assumed that some dolphins will be missed, the consistency of the density estimate (with respect to modeling assumptions) will improve when fewer dolphins are missed during the survey [44].

To this end we propose developing a marine mammal detection system that can fill the gaps by analyzing all areas of the survey FOV simultaneously. A potential application of the proposed system is its use in conjunction with line transect surveys, where it could help to reduce the chance of missing sightings when the observer is not focusing on a particular area, and also reduce human fatigue by reducing the amount of active scanning required by human observers. This study focuses on detecting the most unique and precious marine mammals of Hong Kong, the Chinese White Dolphins (CWDs or dolphins in short).

Different from common object detection problem, detecting dolphins in the wild encounters several challenges:

- *Scarcity of dolphin surfacing.* When continuously capturing video at sea, the appearance of dolphins is rare, i.e., there are many more video frames without dolphin than with dolphin. In addition, the appearing duration of a dolphin on the water surface is only around 1-2 seconds on each occasion.
- *Small size of sighted dolphins.* The detector should be able to detect dolphins that are as small as possible so as to maximize the detection distance. For example, in our dataset, the smallest recorded dolphins have a size of 30×30 pixels (1080p videos captured by GoPro Hero 8 with 140° field of view).
- *Partially visible body part of dolphins.* Dolphins naturally surface to breathe through blowholes on top of their head. As a result, only a few parts of their body are exposed on the water surface while the remaining parts are underwater (see Figure 1a).
- *Distracter objects.* Distant objects, such as waves, sun glare, debris, are visually similar to dolphins and should be distinguished to reduce false alarm. These objects are regarded

as Distracter samples (or false positives).

- *Low efficiency in data annotation.* Standard annotation of dolphins on long-term recorded videos is analogous to human monitoring on board. Annotators need to check each video frame and draw boxes of tiny dolphins in a wide sea view. This process bears low annotation speed and requires intensive human labor.

We aim to detect dolphins in the wild automatically by preparing a Dolphin dataset for training and testing, utilizing state-of-the-art object detection approaches, and developing a system prototype. Specifically, to tackle data scarcity, we joined the line-transect surveys to record videos of dolphin sightings. To improve annotation efficiency, we develop an interactive annotation strategy with two steps: (1) applying a coarse dolphin detector to detect candidate dolphins, and (2) displaying these candidates to the annotator for relabeling and refining bounding boxes. As a result, we build up a large-scale Dolphin detection dataset. We further fine-tune this dataset and compare three off-the-shelf object detection approaches (i.e., Faster-RCNN, FCOS, and YoloV5). We then select the YoloV5 algorithm considering the trade-off between accuracy and speed. Finally, we implement a system prototype, which uses a GoPro camera as an input source and processes the video stream using the selected dolphin detector. The system is designed to alert the users when detecting dolphins and log the detection records.

We summarize our research novelties as below:

- We collect and annotate a large-scale Dolphin detection dataset. Unlike existing datasets for common object detection (e.g. MSCOCO [16]), our dataset is collected from the “real world” (with unconstrained environment and without data pre-processing or curation) and reflects the nature of data from the wild. The dataset contains 2.6k dolphin instances and 16k distracter samples.
- We train three state-of-the-art object detection algorithms on the Dolphin detection dataset and compare their performance and efficiency. We conclude that the YoloV5 algorithm is the best choice for a dolphin detector, considering both speed and accuracy.
- We develop a system prototype to detect CWDs in real time under real-world scenarios. The system has potential application in line-transect surveys to reduce human labor or improve data quality.

## 9 Related Research in the Area

Our works are closely related to the applications for marine mammals and general object detection task; thus, we also review the task from the existing dataset and techniques aspects as below.

Existing marine animals’ detection systems and datasets. Marine biologists have been studying marine animal detection problems for a long time [3, 35, 36]. Traditional automatic detection methods include passive acoustic monitoring (PAM), radio detection and ranging (RADAR) [30], thermal infrared and multi-spectral cameras [37]. We refer the readers to recent surveys [26, 30] to have a better understanding of these methods. Using a thermal IR camera to capture thermal images on the water surface [34, 38] for detection is the closest to our work. Commercially available IR cameras have limited resolution (max around 320x240) which is due to the fundamental limitation of the IR sensor. This resolution is not adequate for monitoring a wide 180-degree FOV since it requires using many cameras or a rotating platform with high mount point. One recent work utilizes aerial and satellite images for whale detection and counting using deep learning approaches [11].

The experiment results show that CNNs based models work well in both tasks on images where the whales appears small in the image. Using Google Earth is not applicable for our study because the photos are updated monthly, and the estimate from a photo at a single time instance will miss CWD that are not surfaced at that instance. The second way to capture high-quality aerial images [39] is by using flights with specialized aircraft or using Unmanned Aerial Vehicles (UAVs) [40]. [42] proposes to detect whales via audio recordings with data-driven techniques. To reduce the burden of collecting data, [39] proposes weakly supervised method to train the detection model with less marine mammal annotations. [41] uses deep neural networks to detect sea cucumbers in underwater video captured by ROVs.

**Marine Mammal Datasets.** Recently, the research community has an interest in detecting marine animals underwater [18, 33] because of the complex environment. Datasets like Brackish Dataset [18] and SEACLEF- 2017 [13] have been released to study this problem. The task of marine mammals' identification [25, 20] (postsurvey task) captures images above the water surface. Marine mammal identification uses high-quality closeup images captured by DSLR with telephoto lens (e.g., the NDD20 dataset [28] and the FinBase dataset [2]), which is easy for the detection task. Additionally, telephoto lens has a narrow field-of-view, making it impossible to detect multiple dolphins appearing simultaneously in a wider view. These datasets do not capture distant images of marine mammals on the water surface as our task of CWD detection requires. Furthermore, there is no released dataset containing the CWDs species yet.

**Common Objects Datasets**. To facilitate the development of general detection algorithms, common object datasets evolve from few categories, and small-scale [8] to thousands of categories, and hyper-scale [16, 14]. Specifically, the pilot Pascal VOC dataset covers 20 daily object categories and contains 11k images, with 2.7 object instances per image. Then, the Microsoft COCO [16] enlarges the number of object categories to 80 and it is densely annotated (7.2 objects per image) on 200k images. The OpenImage [14] further extends the number of object categories to 600 and provides 16,000,000 bounding-boxes on 190,000 images (8.4 objects per image).

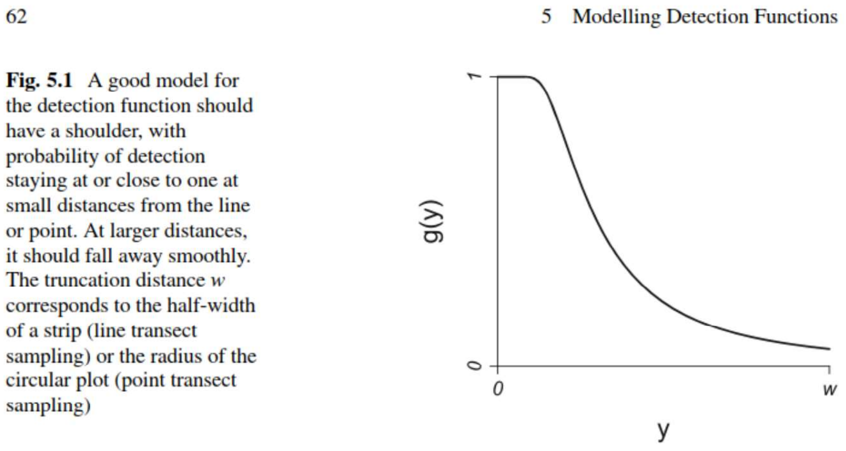
These datasets focus on frequently appearing daily-life objects, so it is easier to construct large-scale datasets. In contrast, the object dataset regarding wild marine mammals, such as CWDs, is not yet available before our work.

**Object Detection Algorithms**. Recently, the applicability of deep convolutional neural networks (CNNs) in computer vision tasks ushered object detection algorithms to rapid development. Current deep-based object detection algorithms can be divided into two categories: two-stage [10, 9, 24] and one-stage [22, 27, 17] object detectors. Their major difference is that: the former contains an extra phase to firstly generate candidate region of interests (RoI), whereas the latter directly regresses bounding-boxes on feature-maps according to rigidly divided grids or anchors. Consequently, the former is less computational efficient and slower in speed compared with the latter. Meanwhile, the former tends to achieve better accuracy than the latter. We recap their structures as follows.

Early works, such as RCNN [10] and Fast-RCNN [9], develop a two-stage object detector that adopts classical Selective Search [29] to generate candidate boxes in the first stage and then refines these candidates with CNNs. Moreover, Faster-RCNN [24] is the first pure convolutional neural network that replaces selective search with a small region proposal network. A drawback of two-stage detectors lies in extra computational cost in the region proposal generation; thereby, researchers turn to develop proposal-free, one-stage object detectors. Specifically, SSD [17] rigidly defines few anchor boxes on convolutional feature maps and regresses boxes on these anchors. A similar spirit is shared with Yolo paradigm [22, 23, 4]. FCOS [27] assigns a centerness score for each pixel according to its distance to box boundaries and regresses this score during training.

Our study selects three object detection algorithms, including Faster-RCNN, FCOS, and YoloV5, and compares them on our Dolphin detection dataset in terms of computational efficiency and detection accuracy.

**Line Transect Survey and Missed Dolphin Sightings.** The line-transect survey methodology based on distance sampling is able to correct for missed dolphin sightings along the transect line. Specifically, a *detection function* is fit to the observed distances to the transect line, which is then used to estimate the proportion of missed dolphins during the survey. The detection function $g(y)$ is the probability that the dolphin is detected at distance $y$ from the transect line. The detection function is decreasing and assumed to start at $g(0)$=1. An example is given here:



62 5 Modelling Detection Functions

**Fig. 5.1** A good model for the detection function should have a shoulder, with probability of detection staying at or close to one at small distances from the line or point. At larger distances, it should fall away smoothly. The truncation distance *w* corresponds to the half-width of a strip (line transect sampling) or the radius of the circular plot (point transect sampling)

Figure: Detection function for distance sampling method (from [42])

The portion of the detection function with value 1 close to $y$=0 is called the *shoulder*. A wider shoulder will give more consistent estimate of the dolphin density regardless of the assumed form of the detection function:

> *More pragmatically, if the true detection function has a wide shoulder, then different models for the function will tend to give similar estimates of density, while if the detection function has no, or only a narrow, shoulder, different models can give rise to very different estimates of density, even if they fit the data equally well. Thus field methods should be adopted that ensure that probability of detection stays close to one for some distance from the line or point.* [42, page 61]

Hence, reducing the number of missed dolphins during the survey (i.e., increasing the width of the shoulder) will make the resulting density estimate more reliable.

## 10 Data collection and preprocessing

The collection and annotation of the dataset for the wild Chinese White Dolphins are different from those of other common objects. Below we introduce the data collection and annotation processes for creating the Dolphin detection dataset.

**a) Collection**

To collect a dataset of CWD for our task, we joined line-transect vessel surveys in the Lantau area to record videos capturing dolphin samples in the wild from Oct. 2020 to Nov. 2021. These videos are recorded under different weather conditions with a total of about 40 hours duration. Figure 1b and 1c show the setup of the GoPro cameras for video recording. Specifically, the three cameras are separately mounted on the left/center/right handrail of the surveying vessel's second deck, fully covering a 180-degree field of view of the dolphin observers. The data collection time, area, team members who accompanied the trip, number of dolphin sightings (unique individuals), and video length can be found in the following table:

| Date | Location | Trip Duration | Team member | No. Dolphin Sightings | Video length |
|---|---|---|---|---|---|
| *Survey dataset:* | | | | | |
| 21-Oct-20 | SWL | 8:30-18:00 | Anh | 3 | 1.1h |
| 19-Nov-20 | SWL | 8:30-18:00 | Anh | 2 | 1.3h |
| 15-Jan-21 | SWL | 8:30-18:00 | Hao | 3 | 3.3h |
| 27-Jan-21 | WL | 8:30-18:00 | Hao | 5 | 1.5h |
| 5-Feb-21 | WL | 8:30-18:00 | Anh | 7 | 3h |
| 23-Feb-21 | SWL | 8:30-18:00 | Anh | 5 | 3.4h |
| 13-Apr-21 | SWL | 8:30-18:00 | Hao | 7 | 3.5h |
| 21-Apr-21 | SWL | 8:30-18:00 | Hao | 1 | 1.2h |
| 26-May-21 | SWL | 8:30-18:00 | Qi | 3 | 3h |
| 26-Jul-21 | NWL | 8:30-18:00 | Qi | 2 | 3.5h |
| 16-Sep-21 | WL | 8:30-18:00 | Qi | 7 | 4h |
| 20-Sep-21 | NWL | 8:30-18:00 | Boey | 1 | 3h |
| 3-Nov-21 | WL | 8:30-18:00 | Boey | 3 | 3h |
| 4-Nov-21 | WL | 8:30-18:00 | Boey | 4 | 3.8h |
| *Field Test:* | | | | | |
| 29-Nov-21 | WL | 8:30-18:00 | Qi and Boey | n/a | 4.5h |
| 10-Dec-21 | WL | 8:30-18:00 | Qi and Boey | n/a | 4.5h |

Table 3. Data collection time, area, video length team members, and number of dolphin sightings (unique individuals).

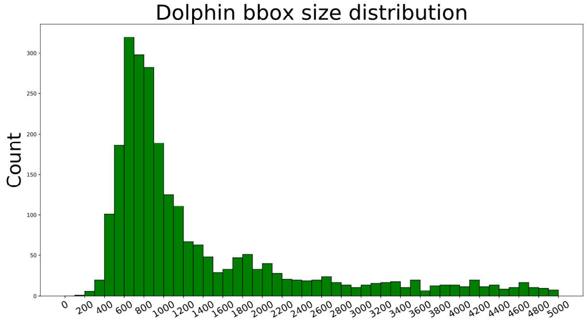
**b) Annotation**

To improve annotation efficiency, we separate annotation into two phases: initial and interactive Dolphin Box Annotation. Specifically, we gather videos from the first few boat trips (October 2020 to January 2021) and manually go through a few sampled videos with dolphins and annotate dolphins' boxes in the initial phase. Then, we use the annotated boxes to train a coarse dolphin detector. For the following surveys, we employ the interactive annotation phase. We first employ the coarse dolphin detector for dolphin detection, then manually correct wrong boxes or add missing boxes. Details of the two phases are elaborated below.

*Initial Dolphin Box Annotation*: In the initial phase, the sighting times from the dolphin experts during the line-transect surveys were used to select short video clips that contained dolphins. Annotators then label a box around the dolphins in each frame, knowing that there exists dolphin in the video clip. The annotators were the postdocs/RAs who went on the line-transect surveys to collect video, and thus they have knowledge of what to look for. The annotators can examine the video frames in sequence, in slow motion, frame-by-frame, and zoomed in, which makes the process more accurate. Annotation is performed using the VGG Video Annotator toolkit [7]. After the initial annotation process, we obtain 571 frames containing dolphins. Then, we randomly split 571 frames into 471 and 100 frames for training/validation purposes. Finally, we train a coarse dolphin detector on this initial subset for interactive dolphin box annotation.

*Interactive Dolphin Box Annotation*: We design an interactive dolphin box annotation strategy to improve the efficiency of annotation. The strategy involves machine prediction and human refinement. Specifically, we firstly apply the coarse dolphin detector on newly collected videos to generate candidate boxes. It is worthy of noticing that these candidates can either be true positives (i.e., dolphin) or near-miss objects (e.g., waves, flares, debris). Besides, the problem of false negatives (missing dolphin) also exists. To tackle this, the annotators then go through the predicted results and mark the true positives and distractors (false positives). Sighting times from the dolphin experts are used to confirm the true positives and to identify clips that contain false negatives (missing dolphins). The false negatives (missing dolphins) are then labelled, similar to in the first annotation stage. This strategy enables us to annotate newly collected video quickly and provides near-miss objects (or distractor examples) for improved detector training.

*Remark*: Although dolphin experts did not directly label each bounding box of each dolphin in the dataset, we think that our annotation process is sufficient for our purposes – the dolphin expert's sighting time is sufficient to find clips where dolphins are present, and the annotators are the same project members who have knowledge of dolphin appearance from the survey trips. Furthermore, the methods for training deep learning models are moderately robust to training with label noise (e.g., missing annotations, mislabeled classes, misaligned bounding boxes). Thus, a perfectly accurate dataset is not required for training the detector. Finally, building large datasets using a semi-automated approach is widely adopted in computer vision and deep learning research.



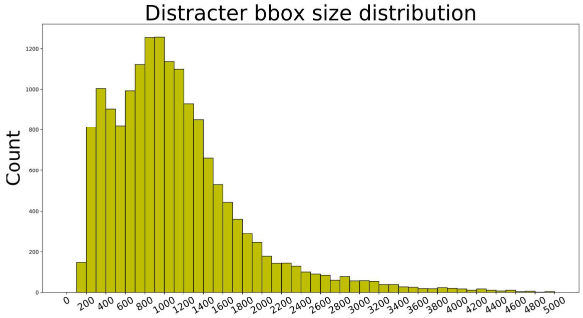


Figure 2: Size distributions of Dolphins and Distracter Instances.

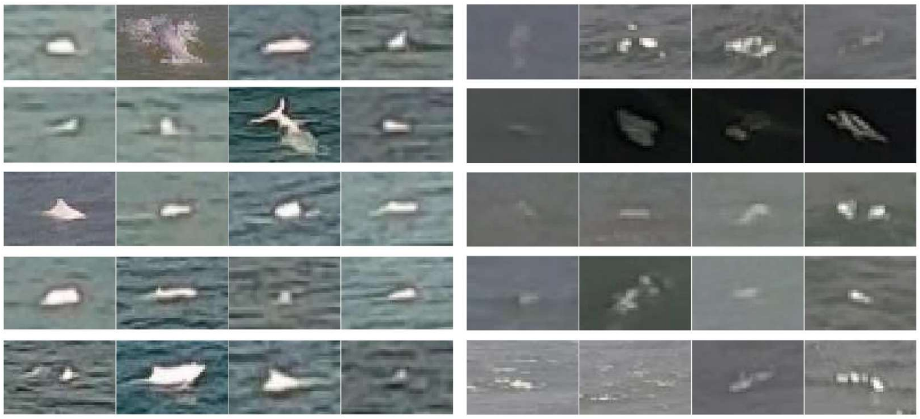
**c) Dataset Statistics**

Except for the first few vessel surveys (for initial annotations), we launch interactive dolphin box annotation after each vessel survey, and finally accumulate 20,404 manually annotated images. They are split into training set and validation set, containing 17,426 and 2,978 images, respectively. Specifically, the training set contains 2,667 dolphins and 22,661 distractor instances; whereas, the validation set contains 109 dolphin instances and 2,573 distractor instances. The training set images are used for training the detector (calibrating the parameters of the neural network), and the validation set is used to test its performance. The training and validation sets are distinct sets (no overlapping images), so as to test how well the detector generalizes to images unseen during training. We notice that only 2,376 images contain dolphins (an image may contain one or multiple dolphins), occupying 11% of the total number of images, also indicate wild dolphins' rare appearances.

| | **No. of images** | **No. of dolphins** | **No. of distracters** |
|---|---|---|---|
| **Training set** | 17,426 | 2,667 | 22,661 |
| **Validation set** | 2,978 | 109 | 2,573 |
| **Total** | 20,404 | 2,776 | 25,234 |

Table 4. Dataset statistics.

We further view distributions of dolphin's and distractor's bounding box sizes in Figure 2. We assess that the median size of dolphin and distractor objects are 800~1000 pixels (H×W≈ 30×30 pixels). This indicates that dolphin detection is a challenging task since the size of dolphins captured by the cameras is rather small, and distractors have similar sizes.



(a) Dolphin examples (b) Distracter examples

Figure 3: Visualization of Dolphin and Distracter samples.

We randomly sample a few dolphin and distracter instances and visualize them in Figure 3. We summarize three main visual characteristics from the figure: (1) Dolphins are quite vague when viewing from a long distance; (2). Only part of the dolphin's body is visible in most cases; (3). Dolphins have discriminative visual patterns such as the shape of fin and fishtail.

## 11 Detection Algorithms

We adopt off-the-shelf general object detection algorithms for dolphin detection. Specifically, we fine-tune three popular object detectors, including Faster-RCNN [24], FCOS [27] and YoloV5 [22], on Dolphin detection dataset. Existing algorithms can be divided into two-stage (e.g. Faster-RCNN) and one-stage (e.g. FCOS, YoloV5) models from the aspect of structural complexity. We separately recap the framework of each detector below.

**a) Faster-RCNN**

Faster-RCNN is the first end-to-end convolutional neural network for object detection and serves as the base framework for the latter improved object detectors. Specifically, before Faster-RCNN, RCNN [10], and Fast RCNN [9] relies on an external manually designed algorithm named Selective Search [29] to generate box candidates for objects, and then refine these boxes with convolutional neural networks. Faster-RCNN replaces the Selective Search module with a small region proposal sub-network; this subnetwork generates initial coarse box candidates for a latter network to regress. Since the Faster-RCNN generates and refines candidate boxes from coarser to a finer level, it is considered a two-stage object detector. To capture objects of multiple sizes, FPN [15] is introduced in Faster-RCNN. Specifically, the FPN module firstly refines feature maps of prior layers by that of the latter with the up-sampling operation, then detects objects on feature maps of multiple layers. Though promising in its performance, the FPN brings extra computational cost. Since the FPN-Faster-RCNN generates and refines candidate boxes from coarser to finer level in a two-stage manner and involves an extra computational cost of the FPN module, these object detectors are more appropriate for cloud computing rather than edge computing tasks.

**b) FCOS and YoloV5**

Like FCOS and YoloV5, one-stage object detectors drop region proposal sub-networks and directly regress objects' boxes on rigidly defined pixels, grids, or anchor boxes. Consequently, one-stage object detectors have better computational efficiency than the two-stage one. FCOS is a proposal-free, anchor-free one-stage object detector. Its main characteristic is the introduction of the center-ness score for each pixel. Specifically, the centerness score is larger when the pixel is far from the box's boundaries and lower when the pixel is near the box's boundaries. Then, FCOS regresses the center-ness score for each pixel and groups heat pixels to predict objects' boxes.

YoloV5 is a variant of the Yolo paradigm. The Yolo [22] paradigm targets improving the efficiency of object detection on both cloud/edge computing, eliminating operations with a heavy computational burden. The Yolo divides images into several grids; if the center of objects falls into a grid, the grid is treated as ground truth and utilized to regress objects' boxes. Hereby, we adopt Yolo version 5, which introduces extra modules and tricks to enhance detection accuracy. Specifically, YoloV5 adopts mature techniques such as mosaic data augmentation, auto anchor adjustment, and Cross Stage Partial Module (CSP) [31]. Moreover, YoloV5 also proposes a new operation named "Focus" which performs down-sampling in the following manner: reshape a $H * W * C$ feature map into $H/2 * W/2 * 4C$ one.

In summary, the dolphin detection prototype system demands both high efficiency and effectiveness. We carefully compare these methods and select YoloV5 as the trade-off algorithm.

**c) Training and Inference:**

Since these detectors have different training/inference strategies, we separately elaborate their settings as follows. Faster-RCNN and FCOS are enhanced by FPN module and adopt ResNet-50 as backbones. For the training phase, we adopt data augmentations, such as random horizontal flip and resizing as in [32]. Using data augmentation increases the size of the dataset and makes the trained network to be invariant to the augmentations used. The model is trained for 30 epochs. For inference, we set long-side to 1333 and resize input images according to the original aspect ratio.

YoloV5 is enhanced by Focus and CSP modules and adopts DarkNet [21] style backbones. Since the YoloV5 paradigm contains four variants (i.e., YoloV5-S/M/L/X with increasing layer-depth), we conduct a test to select an optimal one. All variants share the same training/inferencing scheme unless particularly specified. Specifically, we train 30 epochs with Adam optimizer. Moreover, we adopt data augmentation, including mosaic augmentation, random resizing, horizontal mirror flipping, and additive Gaussian pixel noise. The long-side is set to 1024 during both training and testing.

To handle the distractor objects (e.g., debris) that have similar appearance to CWD, we train with two object classes, CWD and distractor. The distractor examples were added incrementally to the dataset and comprise the false positives of the previously trained detector as described Section 10b.

Finally, we also attempted to train several variants of the model using consecutive frames, but did not see improved test results compared to the image-based model that used the distractor class. Better use of temporal information could be future work.

## 12 Prototype System

We present our prototype system for dolphin detection in Figure 4. We implement two prototype systems for different computation resources and application scenarios: Desktop computer and notebook computers.

**a) The prototype system for desktop computer**

We utilize a GoPro Hero 8 camera to capture live video of the sea surface and one Nvidia 2080Ti GPU to detect the dolphins in the captured video. The pipeline of our prototype system is designed as follows. The system firstly receives a live video stream from a GoPro Hero 8 through a wire connection. The video stream has a resolution of 1080p with 30 frames per second. These video frames are then fed into the dolphin detection module to find possible dolphin locations. Finally, the detected dolphins are visualized and displayed on the application interface for the project team to verify the functioning of the detector in the lab and in the field test.

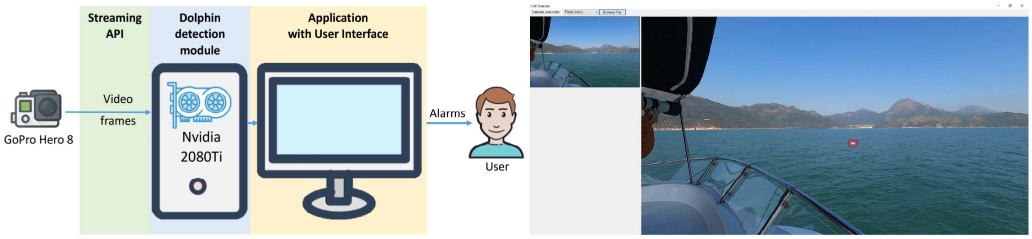


Figure 4. Prototype system for desktop computer.

From the perspective of software development, our system contains two layers:

- *Front-end:* The front-end is a Windows-based application implemented in C#. This application communicates with the GoPro camera, receives the real-time video stream, interacts with the back-end by sending the video stream and receiving detection results. At the same time, the application displays the current video stream together with the visualization of detection results for the user.
- *Back-end:* The back-end is also a Windows-based application implemented in Python. At first, the application loads a dolphin detection model, then waits for a video stream from

the front-end.

**b) The prototype system for notebook computers**

To satisfy the portability requirements of field tests, we also implement a 2-GoPro cameras prototype system with 2 notebook PCs. The whole system is shown in the Figure 5, which is built based on the OpenCV and GoPro python library. The 2 GoPro cameras are mounted on the 2 notebook PCs (with Nvidia GeForce MX450 GPU), respectively. Each camera streams the frames to the corresponding notebook PC in 30 FPS with resolution 1920*1080. To reduce the computation bottleneck, the frames from the two cameras are processed by each notebook PC separately. The frontend and the backend of the system are based on python and are all running on the Ubuntu operating system. In the final system, due to the limitation of the GPU computation power, the system is running at 5 FPS for each camera.  Note that 5 fps still exceeds the target of 2 fps that we specified in the proposal.

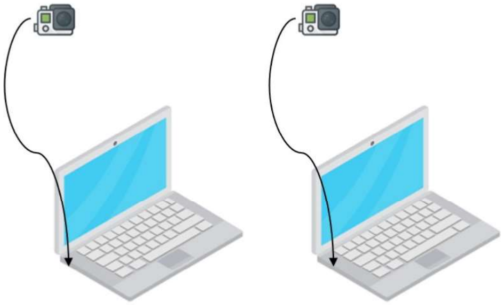


Figure 5. 2-camera prototype system for notebook computers.

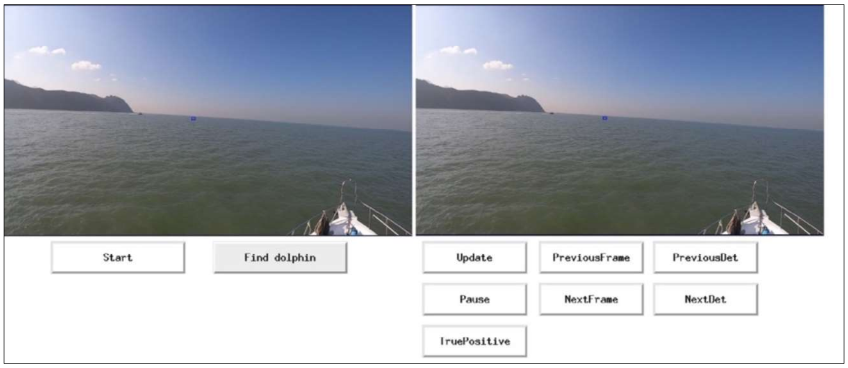


Figure 6. The 2-camera system prototype with the notebook computers.

Fig. 6 shows the GUI for the system on 1 notebook PC, where the left canvas (Canvas 1) shows the real-time input image and results, and the right canvas (Canvas 2) shows the latest image with positive detection results. The functions of the buttons of the GUI are as follows:

- *Start*: connect with the GoPro camera and start frame streaming to the PC;
- *Find dolphin*: the button should be pressed when the human expert spots a dolphin (possibly with binoculars). A screen shot of the frame in Canvas 1 is saved, indicating the human-annotated detection time.
- *Update*: continuously update Canvas 2 with the latest image with positive detection results (normal operating status).
- *Pause*: stop updating Canvas 2 and show the previous Canvas 2 image.
- *PreviousFrame*: when Canvas 2 is in Pause status, show the previous frame in Canvas 2.

- *NextFrame*: when Canvas 2 is in Pause status, show the next frame in Canvas 2.
- *PreviousDet*: when Canvas 2 is in Pause status, show the previous frame with positive detection in Canvas 2.
- *NextDet*: when Canvas 2 is in Pause status, show the next positive detection in Canvas 2.
- *TruePositive*: save a screenshot of Canvas 2 to record a true positive surfacing.

## 13 System Evaluation and Results

We conduct extensive experiments on the Dolphin dataset with three off-the-shelf object detectors, including Faster-RCNN, FCOS, and YoloV5. For simplicity, we select the efficient YoloV5 to study the influence of hyperparameters like resolutions.

**a) Evaluation metrics**

For the evaluation metric, we adopt precision, recall, and mAP@0.5 (mean Average Precision) [16] to report detection performance. Intuitively, Precision is the percentage of detections that are true dolphins (i.e., correct), and Recall is the fraction of true dolphins that were detected. The definitions of the Precision and Recall are:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}),$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}),$$

where 'TP' means the number of correctly detected dolphins, 'FP' means the number of non-dolphin objects detected as dolphins, and 'FN' is the number of dolphins wrongly detected as other objects. A confidence score is associated with each detection, and different Precision and Recall values can be computed by changing the threshold on the confidence score for accepting a detection. The various operating points of the detector is then characterized by plotting a Precision-Recall curve by varying the detection threshold. The area of the region under Precision-Recall curve is called the Average Precision (AP). AP summarizes the performance of the model, where the larger the area, the better the model performance. Mean average precision (mAP) is the mean over the class-wise APs (in our detector there are two classes, dolphin and distractors). A detection is considered correct if the intersection over union (IoU) of its predicted bounding box and the ground-truth box is over 0.5. Additionally, we report processing speed with FPS using a RTX2080Ti GPU.

**b) Object detectors for Dolphin**

We compare Faster-RCNN, FCOS, and YoloV5 on the Dolphin detection dataset regarding detection performance and processing speed.

| Model | Resolution (long-side) | mAP@0.5 (%) | FPS |
|---|---|---|---|
| Faster-RCNN | 1333 | 86.91 | 9.73 |
| FCOS | 1333 | 78.94 | 2.28 |
| YoloV5-S | 1024 | 90.95 | 100.99 |
| YoloV5-M | 1024 | 91.35 | 63.15 |
| YoloV5-L | 1024 | 91.46 | 38.94 |
| YoloV5-X | 1024 | 90.25 | 23.94 |

Table 5: Comparisons of three object detectors: Faster-RCNN, FCOS, and YoloV5-S/L/M/X.

As shown in Table 5, we observe that the YoloV5 models and Faster-RCNN perform much better than FCOS under metrics of mAP@0.5 and FPS. A possible reason is that the center-ness mechanism might not work well for tiny objects, as the central regions of the tiny box are even smaller for regression. We also observe that the FCOS has slow processing speed, which is because the Keras-TensorFlow toolkit requires extra "warm-up" time than the PyTorch toolkit.

We further find that the YoloV5 models perform better than Faster-RCNN under both metrics. The reason is because the Yolo models adopts a lighter computational-friendly backbone (i.e., DarkNet) than the Faster-RCNN does (ResNet50), and YoloV5 also introduces extra lightweight modules, including "Focus" and "CSP" to refine feature representations.

We also evaluate the trained YoloV5 model with Precision, Recall and Precision-Recall curve. We use confidence threshold = 0.45 to select detections. According to the Precision-Recall curve, the model performance of precision and recall are 0.89 and 0.92, respectively.

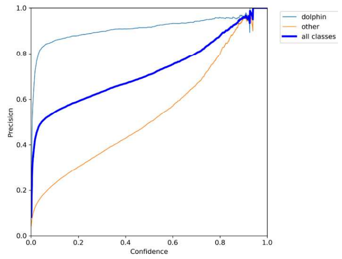


Figure 7. The Precision curve of the model. The confidence threshold is 0.45 and the Precision=0.89.

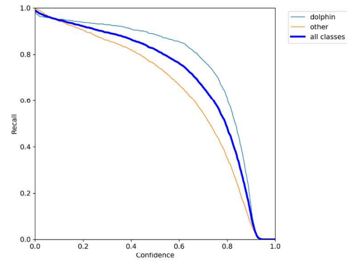


Figure 8. The Recall curve of the model. The confidence threshold is 0.45 and the Recall=0.92.

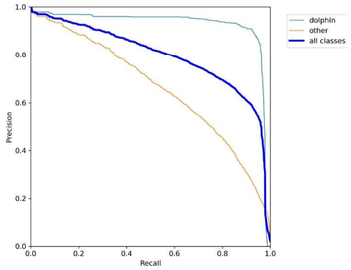


Figure 9. The Precision-Recall curve of the model.

We visualize six of the detected dolphin instances by zooming the corresponding regions in Figure 10. We observe that: (1) Dolphins are tiny compared to broad sea backgrounds; manually finding them by experts requires careful inspections; (2). Dolphin detectors manage to search them out of the background through blurry visual cues like fin shape. Again, we can infer that automatically finding the dolphin in the wild is challenging, considering the visual cues of dolphins are very small. More examples are available here.

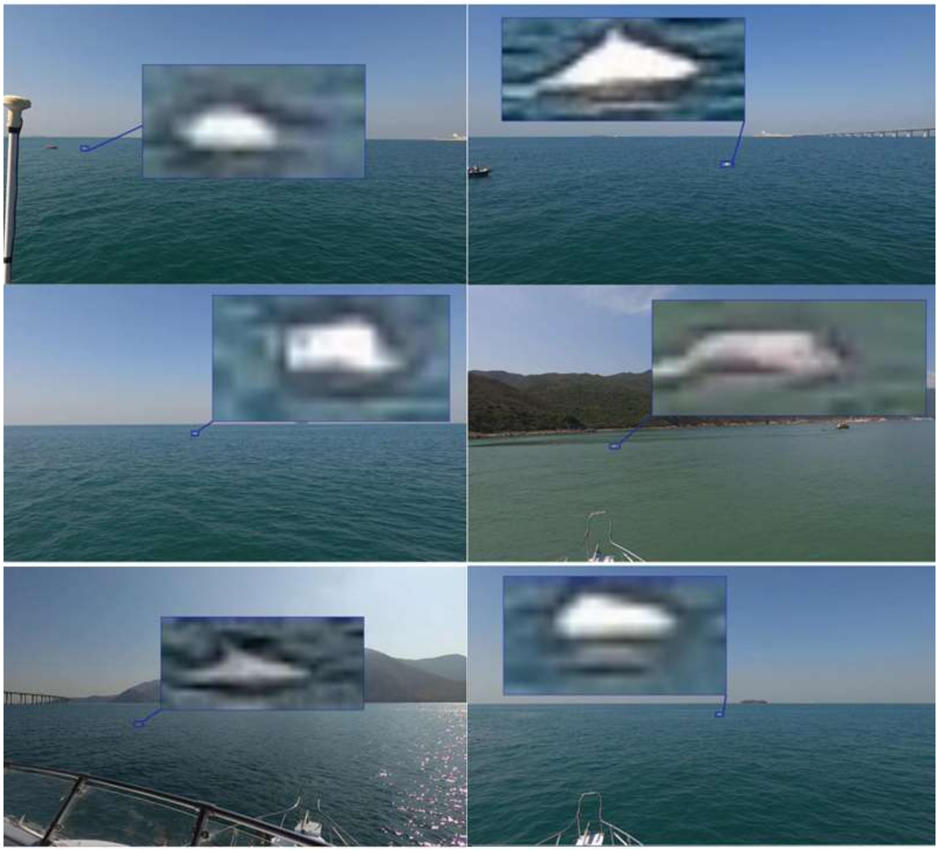


Figure 10: Visualization of detected dolphins in the dataset.

**c) Enhancing initial with interactive annotations:**

As aforementioned in Section 10 b) dolphin box annotation is performed in two phases: initial and interactive annotation. We study the enhancement of interactive annotations on the initial dataset by training the dolphin detector on both annotations.

| | Annotation phase | |
|---|---|---|
| | Initial | + Interactive |
| YoloV5-S | 83.91 | 90.95 |

Table 6: Impacts of two-stage annotation process on detection performance and mAP@0.5 (%) is used as metric.

As shown in Table 6 our first-round dolphin detector is trained with the initial dataset can achieve a reasonable and good initial detection performance: 83.91 mAP@0.5, considering the intensive human labor on hundreds of images. We then use this dolphin detector to go through newly collected videos to generate candidate boxes. We group false positives to another near-miss category and remedy missed true positives. Finally, we combine the interactive with initial annotations. We observe that the dolphin detector benefits from more training images and the distractor class.

**d) Image Resolution**

We experimentally verify the impact of the resolution of input images on detection performance with the most lightweight YoloV5-S detector. Hereby, we pick the long-side of the image in the range [256, 512, 1024] and compare their performances.

|  | Image resolution (long-side $L$) |  |  |
|---|---|---|---|
|  | $L$=256 | $L$=512 | $L$=1024 |
| YoloV5-S | 56.01 | 79.27 | 90.95 |

Table 7: Impact of input image resolution on dolphin detection in Dolphin detection dataset and mAP@0.5 (%) is used as metric.

As shown in Table 7, we observe that performance increases before L=1024. We fix this setting for the rest of the YoloV5 models. For Faster-RCNN/FCOS models, they predefine long-side to 1333, a similar scale as 1024, and we follow their original settings.

**e) Field Test performance**

We conduct field tests on Nov. 29 2021 and Dec. 10 2021. The field tests are used to evaluate detection performance of the system as compared with the human dolphin expert, i.e., we would like to see whether the detection system can detect dolphins when human expert finds a dolphin *with binoculars*. Figure 11 (top) shows the actual system in the field tests: 2 notebook PCs are connected with camera 1 and camera 2 separately, and camera 3 is used to record high quality videos for offline confirmation of the detection results.

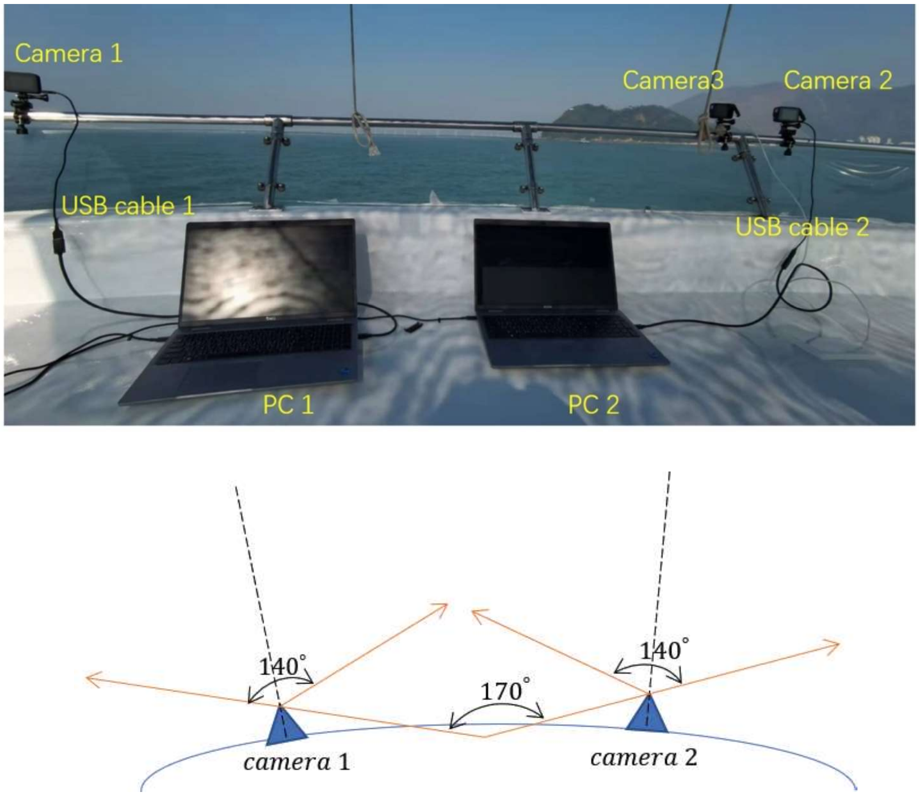


Figure 11: (top) The 2-camera detection system used in the field test. (bottom) the combined FOV of the 2 cameras.

In the field test, the combined FOV of the 2-camera system is about 170 degrees, as shown in Figure 11 (bottom). Each GoPro camera has an FOV of 140 degrees, so the theoretical largest FOV for the two camera system is 280 degrees. Since the cameras are processed separately, this

280-degree FOV could be achieved by simply rotating the cameras away from each other.

During the field test, the detection system scans the 170-degree FOV continuously and records detections in real-time. The human observer (dolphin expert) scans the same FOV with either binoculars or unaided eye and notifies the team when a dolphin surfacing is spotted. Immediately after notification, the surfacing record is entered into the system. Since there is a short 1-2 second delay for entering a human observer's spotting into the system, we define "on-time" detection as within 2 seconds of the spotting by human observer (between 2 seconds before or 2 seconds after the spotting). We define "early" detection as the system detecting a dolphin more than 2 seconds before the human, and likewise "delayed" detection as the system detecting a dolphin more than 2 seconds after the human. Missed detections are dolphins that the system did not match.

| | **Day 1**<br>**(Nov. 29 2021)** | **Day 2**<br>**(Dec. 10 2021)** | **Total** | |
|---|---|---|---|---|
| **No. of Dolphins (Surfacing)** | 19 | 224 | 243 | |
| **On-time detections** | 9 | 160 | 169 | (69.5%) |
| **Early detections** | 3 | 21 | 24 | (9.9%) |
| **Delayed detections** | 5 | 5 | 10 | (4.1%) |
| **Missed detections** | 2 | 38 | 40 | (16.5%) |
| **Reason for misses** | too far (1)<br>missed (1) | too far (25)<br>partial body (5)<br>occluded (1)<br>too dark (1)<br>missed (6) | too far (26)<br>partial body (5)<br>occluded (1)<br>too dark (1)<br>missed (7) | (10.7%)<br>(2.1%)<br>(0.4%)<br>(0.4%)<br>(2.9%) |

Table 8: The field test results: comparison of detection of surfacing dolphins by the system and a human observer.

Table 8 shows the performance of the field tests in terms of on-time, early, delayed, and missed detections. Overall, 83.5% of surfacing dolphins seen by the human were also detected by the system. Most dolphins can be detected by the system immediately (on-time, 69.5%), while a small percentage are detected before the human observer (9.9% early). The delayed detections (4.1%) are caused by the dolphin initially being too far away, and then moving closer to the vessel where it is found by the system. Finally, most of the missed detections are due to the dolphin being too far away (10.7% of surfacing dolphins). Note that the human observer was aided by binoculars, so it is possible for the human to see further than the camera system. If we exclude the missed detections due to distance, the detection recall for the field tests is 93.5%.

We further analyze the timeliness of the automatic detections compared to the human in the below Figure 12. There are 5 detections from the system that occurred more than 30 seconds before being observed by the human. These very early detections are promising because they show that, by continuously analyzing the whole FOV, the system can detect dolphins that may have been missed by the human observer at first, possibly due to their inattention to the particular region. Finally, note that in this analysis we are using the human observations as the ground-truth, and thus we cannot determine if the system finds dolphins that were not spotted by the observer.

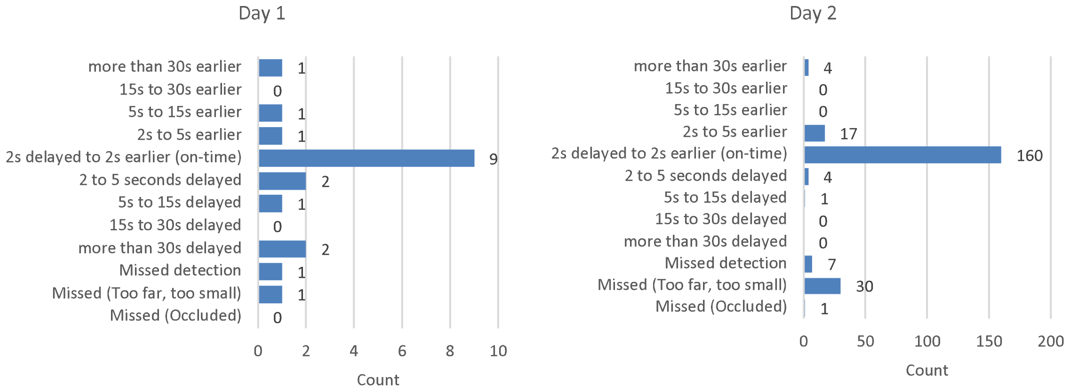


Figure 12: Timeliness of automatic detections compared to human observers for Day 1 and Day 2 of the field test.

We also show the visualization results in the field test in Figure 13, which shows the model can detect small dolphins. Several video demos showing the detection results of the proposed system during the field tests can be found here: demo_videos

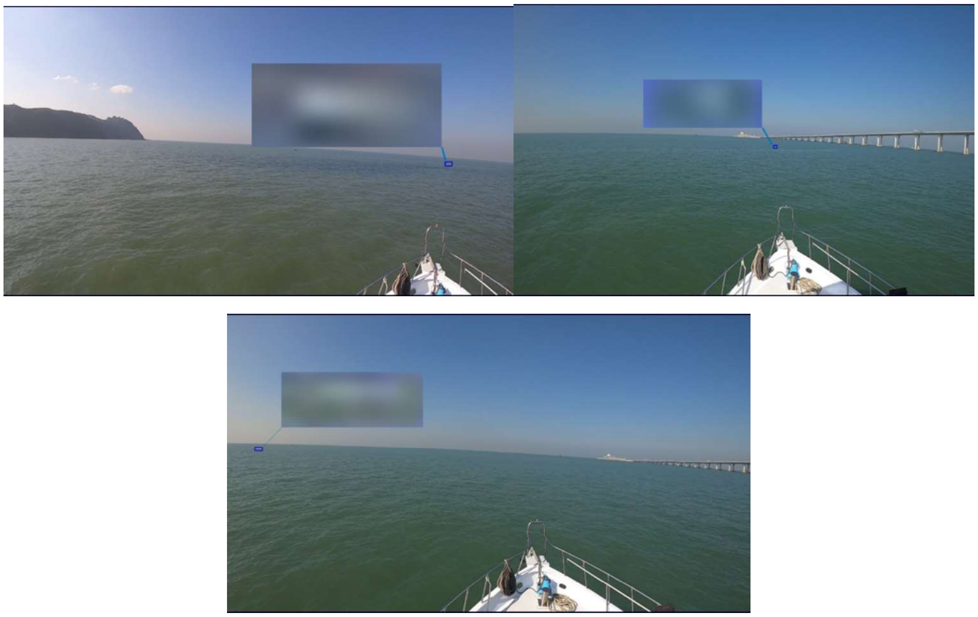


Figure 13: Visualization of detected dolphins in the field test.

**f) Analysis of detection distance**

We next analyze the performance of the system to detect at different distances from the camera. For the survey and field test data, we estimate the distance of each detection or ground-truth box to the camera in the 3D world. The distance is computed using the 2D-image to 3D-world mapping that is calculated mathematically using the GoPro camera’s calibration parameters. The camera calibration parameters consist of intrinsic parameters (the focal length, lens distortion) and the extrinsic parameters (the location and orientation of the camera in the 3D world). The mathematical equations are standard formulas developed in computer vision using 3D geometry [43]. In distant regions, the estimated distances have larger errors, so we clip the estimated distance to be no more than 500m. The following Figure 14 shows the estimated distances of detections in an example image. The length of the above-surface portion of Dolphin "3" is estimated as 1.2m by our system, which is a reasonable estimate (the average CWD length is 220cm, and the visible portion is usually 2/3 of the length). We have also verified that the estimate lengths of several landmarks (e.g., gap between pillars of the bridge) are reasonable. Thus, the estimated distance should be good enough for our post-hoc analysis.

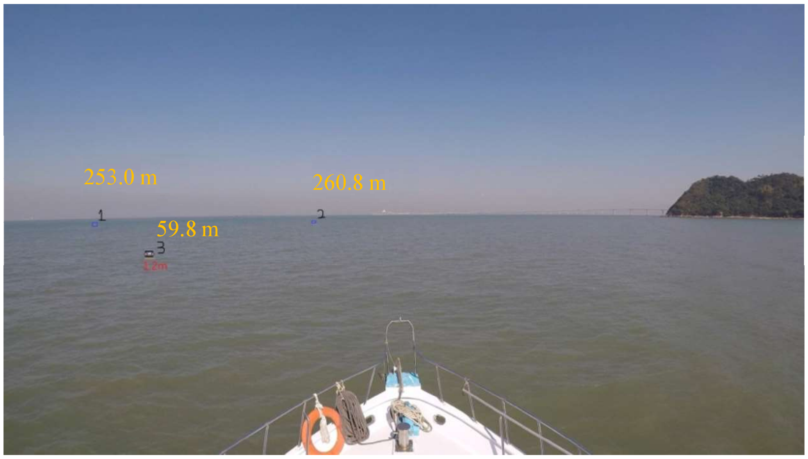


Figure 14: example of estimated distances calculated via 3D camera geometry. The distance to camera of the dolphins are in orange. The estimated length of the visible portion of Dolphin 3 is in red.

Table 9 shows the statistics for distances of dolphins in the training set ground-truth (survey data), true detections in the validation set (survey data), and the field test.

| | 0m - 100m | 100m - 150m | 150m - 200m | 200m - 250m | 250m - 300m | 300m - 350m | 350m - 400m | 400m - 450m | 450m - 500m |
|---|---|---|---|---|---|---|---|---|---|
| Training Set Ground-truth | 18588 | 1360 | 709 | 517 | 314 | 226 | 153 | 145 | 100 |
| Validation Set TP detections | 3481 | 209 | 131 | 20 | 33 | 9 | 4 | 1 | 3 |
| Field Test TP detections | 183 | 20 | 9 | 4 | 4 | 3 | 2 | 0 | 0 |

Table 9: Numbers of ground-truth or true-positive (TP) detections at different distances in the survey data (training and validation sets) and the field test.

The system is trained on examples that range up to 500m from the camera, while most training samples are between 0-200m. On the validation set and in the field test, the system can successfully detect dolphins up to 500m away. According to recent statistics from line-transect surveys (see Figure 15), about 94% of sightings by humans are at 500m or below, and sightings over 500m are rare (only 3/49=6% are over 500m). Thus, it is fair to say that our system has similar capability to normal human observers in terms of detection distance.

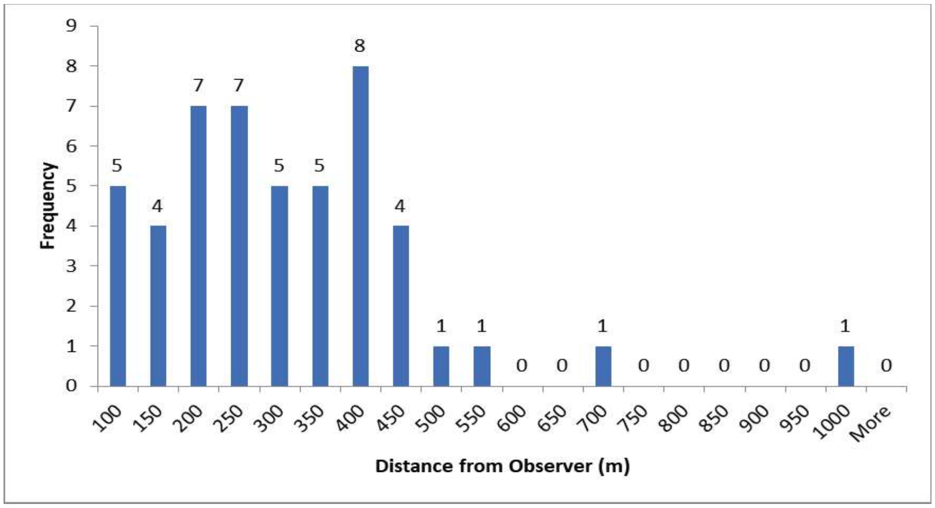
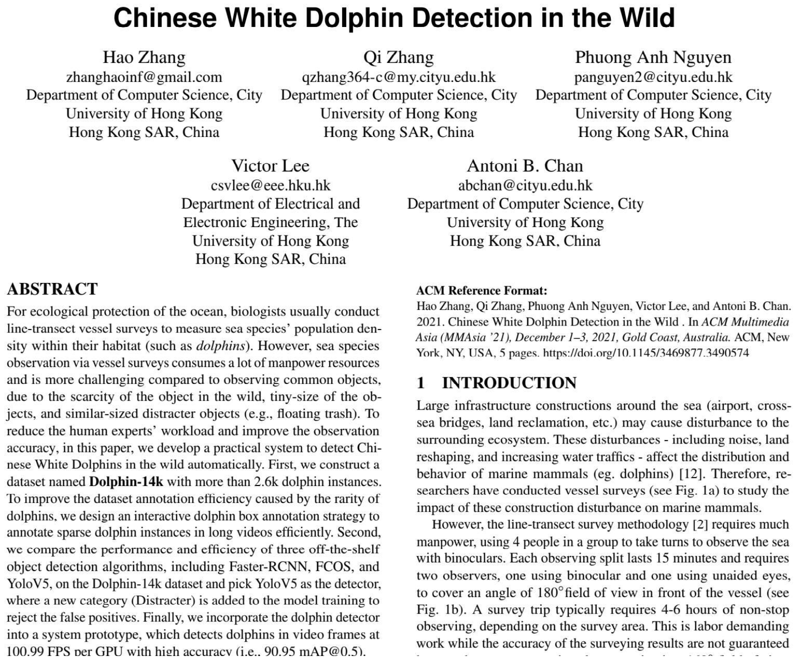


Figure 15: Distances of 50 dolphin sightings during line-transect surveys in NEL/NWL (Beaufort Sea State 3 or better, human eye detection). Figure provided courtesy of MEEF.

## 14 Research Outputs

Our research work is accepted by the 2021 ACM Multimedia Asia conference (Dec. 1-3, 2021), entitled “Chinese White Dolphin Detection in the Wild”. Our team member, Dr. Qi Zhang has given a talk in the conference together with a poster presentation. We also have made the Dolphin dataset publicly available for research purposes here: dataset.zip. This will extend the influence of the research work to more people.



**Chinese White Dolphin Detection in the Wild**

Hao Zhang
zhanghaoinf@gmail.com
Department of Computer Science, City University of Hong Kong
Hong Kong SAR, China

Qi Zhang
qzhang364-c@my.cityu.edu.hk
Department of Computer Science, City University of Hong Kong
Hong Kong SAR, China

Phuong Anh Nguyen
panguyen2@cityu.edu.hk
Department of Computer Science, City University of Hong Kong
Hong Kong SAR, China

Victor Lee
csvlee@eee.hku.hk
Department of Electrical and Electronic Engineering, The University of Hong Kong
Hong Kong SAR, China

Antoni B. Chan
abchan@cityu.edu.hk
Department of Computer Science, City University of Hong Kong
Hong Kong SAR, China

**ABSTRACT**

For ecological protection of the ocean, biologists usually conduct line-transect vessel surveys to measure sea species’ population density within their habitat (such as *dolphins*). However, sea species observation via vessel surveys consumes a lot of manpower resources and is more challenging compared to observing common objects, due to the scarcity of the object in the wild, tiny-size of the objects, and similar-sized distracter objects (e.g., floating trash). To reduce the human experts’ workload and improve the observation accuracy, in this paper, we develop a practical system to detect Chinese White Dolphins in the wild automatically. First, we construct a dataset named **Dolphin-14k** with more than 2.6k dolphin instances. To improve the dataset annotation efficiency caused by the rarity of dolphins, we design an interactive dolphin box annotation strategy to annotate sparse dolphin instances in long videos efficiently. Second, we compare the performance and efficiency of three off-the-shelf object detection algorithms, including Faster-RCNN, FCOS, and YoloV5, on the Dolphin-14k dataset and pick YoloV5 as the detector, where a new category (Distracter) is added to the model training to reject the false positives. Finally, we incorporate the dolphin detector into a system prototype, which detects dolphins in video frames at 100.99 FPS per GPU with high accuracy (i.e., 90.95 mAP@0.5).

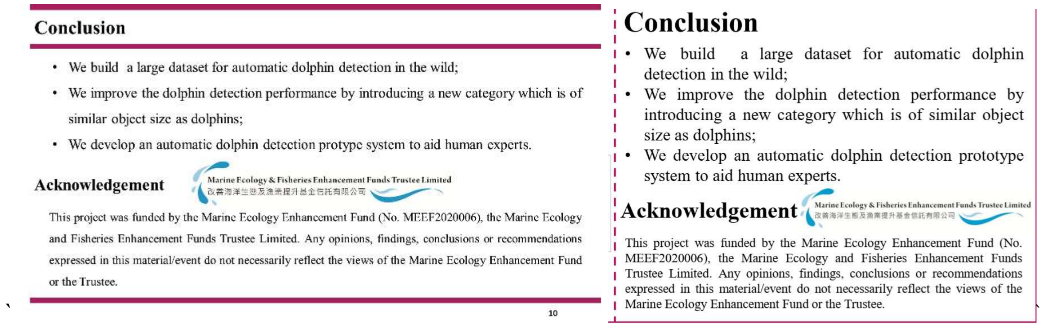
**ACM Reference Format:**
Hao Zhang, Qi Zhang, Phuong Anh Nguyen, Victor Lee, and Antoni B. Chan. 2021. Chinese White Dolphin Detection in the Wild . In *ACM Multimedia Asia (MMAsia ’21), December 1–3, 2021, Gold Coast, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3469877.3490574

**1 INTRODUCTION**

Large infrastructure constructions around the sea (airport, cross-sea bridges, land reclamation, etc.) may cause disturbance to the surrounding ecosystem. These disturbances - including noise, land reshaping, and increasing water traffics - affect the distribution and behavior of marine mammals (eg. dolphins) [12]. Therefore, researchers have conducted vessel surveys (see Fig. 1a) to study the impact of these construction disturbance on marine mammals.

However, the line-transect survey methodology [2] requires much manpower, using 4 people in a group to take turns to observe the sea with binoculars. Each observing split lasts 15 minutes and requires two observers, one using binocular and one using unaided eyes, to cover an angle of 180°field of view in front of the vessel (see Fig. 1b). A survey trip typically requires 4-6 hours of non-stop observing, depending on the survey area. This is labor demanding work while the accuracy of the surveying results are not guaranteed

Figure 16. The abstract page of the paper.



**Conclusion**

- We build  a large dataset for automatic dolphin detection in the wild;
- We improve the dolphin detection performance by introducing a new category which is of similar object size as dolphins;
- We develop an automatic dolphin detection protype system to aid human experts.

**Acknowledgement** Marine Ecology & Fisheries Enhancement Funds Trustee Limited 改善海洋生態及漁業提升基金信託有限公司

This project was funded by the Marine Ecology Enhancement Fund (No. MEEF2020006), the Marine Ecology and Fisheries Enhancement Funds Trustee Limited. Any opinions, findings, conclusions or recommendations expressed in this material/event do not necessarily reflect the views of the Marine Ecology Enhancement Fund or the Trustee.

10

**Conclusion**

- We build  a large dataset for automatic dolphin detection in the wild;
- We improve the dolphin detection performance by introducing a new category which is of similar object size as dolphins;
- We develop an automatic dolphin detection prototype system to aid human experts.

**Acknowledgement** Marine Ecology & Fisheries Enhancement Funds Trustee Limited 改善海洋生態及漁業提升基金信託有限公司

This project was funded by the Marine Ecology Enhancement Fund (No. MEEF2020006), the Marine Ecology and Fisheries Enhancement Funds Trustee Limited. Any opinions, findings, conclusions or recommendations expressed in this material/event do not necessarily reflect the views of the Marine Ecology Enhancement Fund or the Trustee.

Figure 17. The presentation slide and poster.

## 15 Discussion of Outcomes and Positive Impact

In this section we discuss the project outcomes and positive (potential) impact as listed in the original proposal.

**a) Project Outcomes**

There were 4 expected outcomes stated in the proposal. Here we list each expected outcome and discuss its fulfilment in turn.

1) *Detect CWDs automatically: the system can automatically detect CWDs appeared in the video frames captured by the cameras.*

The CWD detection system has been developed and tested in the field. Using the human observations as ground-truth, the system can achieve 92% recall of surfacing dolphins (seen by humans) with 89% precision (see Section 13b for details).  The system can detect dolphins up to 500m away, which is similar to human capability with unaided eyes (see Section 13f). The accuracy of the detector could still be improved through increasing training data, but we may think that its current performance is similar to a normal human observer, and thus the outcome is achieved.

2) *Detect CWDs in "real-time": the system can be connected to the cameras to perform detection at a rate of not less than 2 video frames per second such that it can detect CWDs as they move, disappear under the water surface, reappear above the water surface, or even change in appearance due to specular reflection or orientation of the CWDs. Once a CWD is detected, the system would display the result to the observers on board so that they can perform timely close monitoring of the detected CWDs.*

The system prototype runs at 5 fps and processes the video in real-time. The system can produce timely detections (see Section 13e). In summary, a majority of the time, the system can detect dolphins at the same time as the human observers (69.5% of surfacing dolphins). There are also cases where the system can detect dolphins before the human observers (9.9%), as well as vice versa, the system was delayed (4.1%). Most missed detections are due to the distance being too far (10.7%), as the human observer used binoculars but the system did not use any zoom feature. Thus, this outcome is achieved.

3) *Detect CWDs on a 180-degree field of view (FOV) simultaneously and continuously: unlike human observers who scan the FOV from one side to another periodically to search for dolphins, the prototype can continuously capture snapshots of the whole FOV by video*

*cameras and stream the frames to the detector for processing.*

This FOV of the two-camera system is about 170 degrees during the field test, which is close to the 180-degree target. The FOV can be increased by simply rotating the cameras away from each other, since the cameras are processed independently. Thus, this outcome is achieved.

4) *Record and compute data such as sighting time, position, angle and distance of the detected dolphins in real time: with the use of other sensors such as GPS and the captured video frames, the associated data can be computed, recorded and stored together with the images of the detected dolphins.*

The prototype system records the images, spotting time, and position in the image in real-time. However, we did not calculate the associated data, such as angle to transect line or distance to camera. Since a good detector is a prerequisite to this task, we therefore mainly focused on improving the detector. In Section 16a3, we discuss how the angle and distance can be calculated from the scene geometry as future work. Thus, this outcome is partially achieved.

**b) Positive Impacts**

Please note that the "positive impacts" written in the proposal mainly refer to *potential* medium and long-term benefits of the research, and do not mean directly measurable outcomes (short-term impact). Thus, many of these potential impacts have not been directly tested, since we have not intended to measure them from the outset of the project. Here we can only speculate from the system performance whether the medium- or long-term impact is possible. In Section 16a, we detail future work that could further realize these impacts. There are 7 potential impacts, which now discuss one-by-one.

1) *Relieve human effort from tedious, tiring and continuous search for dolphins by automating the detection process so that the human observers can spare their effort from the on-effort search to the off-effort state to track the sighted dolphins.*

The detection system has capability close to normal human observers, and thus we think that the system could be used as an extra observer to lighten the workload of the human team during line-transect surveys. How to do this effectively is a research topic in itself in human-computer interaction. The screen-viewing paradigm presented in this report is mainly for conducting the field test of the detection system. For the actual line-transect survey, instead of viewing the screens, audio prompts could be used to notify the observers of potential sightings (both angle and distance) and observers can then pay attention in those areas. Audio notification has been used previously in [26].

2) *Avoid human limitations due to a variety of factors such as fatigue, handshake, or even eye-blinking.*

A limitation of human observers is that they need to scan the survey FOV, and thus it is possible for them to miss dolphins that appear outside of their central FOV. Our system processes the full survey FOV simultaneously and thus can avoid this human limitation. This is evinced by the timeliness results (Section 13e), where in 5 cases the system detected dolphins more than 30 seconds before the observer. As for human fatigue and its associated problems, we speculate that the system could be used as an extra observer during line-transect survey, and thus the human observers could spare some effort during the search. How this could be achieved without affecting the quality of the line-transect survey would need to be further investigated.

3) *Overcome the human constraint of scanning the 180 degree observation range at the front from one side to another.*

Although the human FOV is large (~190 degrees), the central FOV that contains high visual acuity for detecting dolphins is only 60 degrees. Thus, the human must scan from side to side in the 180-degree survey FOV, and it is possible for them to miss dolphins not appearing in central vision. In the field test, our system processes large FOV (170-degrees or even higher) simultaneously, and thus can overcome this human constraint (see details in Section 13e). For the two-camera system, the largest combined FOV is 280 degrees.

4) *Improve the effectiveness of detection by learning from the increasing number of samples.*

The initial training phase used 571 images, while the interactive phase used an additional 16,855 images. Using more images (interactive phase) improves the mAP performance from 83.91 to 90.95, as illustrated in Table 6 (Section 13c). The detector still misses some occurences of surfacing dolphins, which could be caused by various factors such as appearance or illumination changes. Thus, we anticipate that training with more data collected from a variety of conditions should further improve the detection accuracy.

5) *Improve the timeliness and accuracy of the associated data to be recorded with the sighted dolphins.*

The summary of timeliness and accuracy of automated detection is presented in the discussion in Sections 15a1 and 15a2. For 5 dolphin surfacings in the field test, the automated system detected the dolphin more than 30 seconds before the human. Thus, there is potential for the system to help improve the timeliness of detection when used with human observers. Further investigation is of course needed. In the current project, we only test whether the system can achieve detections that are consistent with the human observer (the observations of the human observer are the ground-truth). The current system achieves good recall (about 92%) compared to the human ground-truth, and this could likely be improved by increasing the training data to include more variety of conditions. To test whether the system improves over human capability would require further investigation and experiments. Collection of associated data, such as angle and distance, has not been implemented in our system (see discussion in Section 15a4). Thus it is unclear whether such measurements by camera would be more accurate than those by the existing methods.

6) *Detect other objects of interest that may appear in the video frames such as other marine mammal and floating debris in relation to other environmental applications.*

Our detector already has a “distractor” class that comprises debris and other objects with similar appearance to CWD. It should be possible to add more classes, such as other marine mammals or types of floating debris, for other applications.

7) *Monitor a 24-hour ‘dolphin exclusion zone’ in relation to certain marine works for dolphin presence.*

Since the automated system does not “get tired”, it can monitor a set of video streams on a continuous basis. However, the current system is only tested during daytime conditions, whereas

24 hour monitoring requires usage in both daytime and nighttime conditions. Sunrise and sunset times might be particularly challenging due to the specular reflections of the sun on the water, which could either confuse the detector or obfuscate the CWD. Standard cameras could be used for detection at night time if there exists sufficient lighting. If there is not sufficient lighting, then infrared camera or low-light (night vision) camera would be required. For all these cases, additional video data would need to be collected and annotated to train the detection model. Most likely, separate detection models would be learned for each condition (daytime vs. nightitme) to enhance the overall accuracy.

## 16 Summary and Way Forward

In this project, we develop a practical application to detect Chinese White Dolphins in video cameras at sea. Towards the target, we conduct three sub-tasks: collection and annotation of the dataset, the experimental study on detection algorithms, and prototype system development. Specifically, we design an initial-interaction annotation strategy to accumulate a dolphin detection dataset efficiently. Then, we conduct experiments to compare and select an optimal object detection algorithm balancing both efficiency and performance. Finally, we design and implemented a user-friendly prototype system to conduct field tests of the detector. Through our work, we hope to promote general deep learning techniques to solve needs in real-world scenarios and help reduce intensive human workloads.

The project could bring the following positive impacts on the current survey methodology and the possibility of other applications. It relieves human effort from tedious, tiring and continuous search for dolphins by automating the detection process so that the human observers can spare their effort from the on-effort search to the off-effort state to track the sighted dolphins. The system overcomes the human constraint of scanning the 180-degree observation range at the front from one side to another, and could improves the timeliness and accuracy when recording sighted dolphins.

**a) Future Work:**

We next describe some future work to build on the current project with the aim of using it in line transect surveys.

***1. Increasing detection distance*:** The current system can detect dolphins up to 500m in 1080p video. Due to the low computing power of our laptops used in the field test, we did not use the full GoPro camera resolution of 4k video, but instead we used 1080p video. Thus, dolphins that are very far away will appear too small in the 1080p image for the detector to find. If we increase the resolution to 4k video (2x the resolution as 1080p), then the effective detection distance will potentially double (from 500m to 1000m). Detection at distance can be further improved by adding a telephoto lens on the GoPro video cameras. Adding a telephoto lens will decrease the field-of-view of the camera, so additional cameras need to be used to ensure 180-degree FOV capture. For example, 2x zoom will double the effective detection distance, but also require twice as many cameras since the FOV is halved for each camera.

***2. Study the impact of weather:*** When collecting data, we joined the surveys as much as possible without considering the weather conditions. Thus, the overall results reported in Table 5 are representative of performance in various weather conditions during the surveys. As shown in Table 4, the number of distractors (e.g., white caps caused by windy conditions) is almost 10 times more than the number of dolphins. The first day of the field test was windy, and the system could handle the roll and pitch of the vessel. Despite the windy conditions, the system still managed detect dolphins within its detection distance. A more detailed anaysis about the impact of weather (sunny vs. cloudy, windy vs. calm) could be undertaken to determine the reliability of the detector under various weather conditions. Additional training data could then be collected to improve the detector when operating in these less reliable conditions.

***3. Estimation of Distance and Angle***: In this report, we have computed the distance offline (posthoc) after data collection. The accuracy and robustness of the distance estimate can be improved by using more accurate calibration of the camera system and location on the vessel. To improve the accuracy of the estimates, a camera rig should be developed (see next improvement) so that the camera position is consistent across surveys. The distance estimate could also be improved by calculating the correspondence between detections from two cameras if they have overlapping FOV. The distance estimates would need to be compared with the current measurement method, such as laser range finder, to judge its efficacy. Finally, the system can be further enhanced to estimate the distance and angle to camera in real-time.

***4. Improved prototype system*:** The current prototype system could be improved in several ways to increase its robustness and consistency across deployments in preparation for usage in line-transect surveys. The updated prototype system could comprise a portable workstation computer with GPU, a camera array mount, and field software. The detection distance of the system can be improved by adjusting the video resolution and adding zoom lenses on the cameras. Furthermore, the field software will be updated to also calculate the angle and distance of the sighting.

- The camera array mount will comprise a platform for mounting the cameras in a consistent way, which can then be mounted on the viewing deck of the vessel. This will make the setup of the system easier during surveys, and ensure the proper 180-degree FOV is consistently covered for each survey. The rig and mount point will also ensure that the camera extrinsic parameters do not deviate, which improves the distance and angle estimation. Cameras can also be equipped with Zoom lenses to increase the detection distance.
- The workstation should have fast enough GPU for processing all videos in real time. The existing workstation can probably be reused, but the GPU and hard disk need to be updated, so as to handle increased video resolution or cameras needed for longer detection distance. An uninterruptible power supply (UPS) should also be included to mitigate power failure problems while operating on the vessel.
- The field software could be updated to give real-time notification of the results, which could include the following: a) audio prompts of detections (distance and angle) to notify human observers; b) a display map to show the locations of recent detections relative to the vessel, in case observers would like to review the recent detections. The software will also have a real-time data entry function for the observers to both help record the sightings from the observers and synchronize the human observations with the system. The data entry function could be based on speech recognition in order to minimize the need to look away from the survey task.

***5. Integration of prototype into line-transect survey*:** We believe that the automatic detection system could be used to supplement (rather than replace) the human scanning effort, since the system can analyze the whole 180-degree FOV continuously, thus monitoring regions where the human is not actively focusing on. To this end, an interesting question is how to best integrate the automatic detector into the line-transect survey so as to improve the efficacy of the whole line-transect survey and to reduce human fatigue. The former goal could provide better data quality from the survey, while the latter could provide reduction in human labor cost for each survey (e.g., allowing for more frequent surveys given the same budget). Several research questions are raised:

- What is the best interface for the human observers to seamlessly interact with the detection system, in terms of both improving usefulness and reducing cognitive load?
- How to best utilize the new system within the current line-transect workflow? How to modify the protocol to increase the survey's effectiveness, in terms of both accuracy and human labor?
- Given the benefits brought by the system, how does the existing line-transect methodology need to be modified when computing the dolphin abundance and other metrics?

Answering these questions would require close collaboration with dolphin experts who are familiar with line-transect surveys. We are open to such collaborations and look forward to any opportunities that are available.

**b) Comments from Dolphin Experts**

We have previously consulted with the survey team during the project, and they were interested in a system that can detect all dolphins (no false negatives), as compared to reducing false positives. Thus, in the last stage of the project, we have tuned the system to detect all dolphins. The dolphin experts were generally very supportive and cooperative during the project, and we hope to develop the project further with them to incorporate the system into the existing line-transect survey. We have recently asked the dolphin experts for more feedback, and their reply is copied here:

> **Expert 1:** *We are always interested to work with your team on developing the technology which aims the dolphin monitoring works. Regarding your approach to improve the system, we agree with you suggestion to improve the camera array. The vessel line transect distance sampling (i.e. vessel survey) for dolphin monitoring in Hong Kong adopted by both AFCD and Mott use double surveyor method which mean we have two surveyors one using naked eyes for a wider angle observation while another using marine binocular to aim observation far distance. Therefore, your camera array should have both wide angle cameras and telecameras, and both wild angle and tele cameras should cover the front 180 degree observation range. In vessel line transect distance sampling, we need to record the sighting angle of the dolphin (i.e. angle in between the sighted dolphin and transect) and also the sighting distance (i.e. the distance between the dolphin and the survey boat). These information are required for calculating the perpendicular distance of the dolphin toward the transect line, which will be used to abundance and density analysis of the dolphin. So you may need to think about how the system could obtain these information automatically or relay on human surveyor or system operator to obtain these information by human force. In addition to above, we know there are ecologists using drone for dolphin survey.*

> *This will be an interesting area that your automatic detection system can further explore.*
>
> **Expert 2:** *Further to Expert 1's feedback for vessel line transect survey, I think your automatic dolphin detection system could be also explored for applying to onshore survey/tracking of dolphins. Currently our land-based dolphin survey by theodolite tracking is facing an issue of using a software (theodolite tracking program) that can only work with old model of digital theodolite. We were alert to this issue years ago, but there has been no further development of available software compatible with new theodolite models, and it is very hard to find the old theodolite model now. The land-based dolphin survey by theodolite tracking also requires 3 surveyors for each time. So, it would be of great interest for having more advanced and cost-effective technology for onshore survey. I think the improvement in dolphin detection and cost effectiveness by automatic detection system would be even more significant for onshore survey/monitoring than vessel survey.*
>
> *Other than dolphin monitoring, maybe it can be broaden for automatic detection of other animals in similar habitat (e.g. shorebirds, seabirds) and monitoring in the future?*

We would like thank the dolphin experts for their help and continued support on our project. Our discussions with them have been very helpful for understanding the challenges and way forward in our project, as well as for improving this report.

## 17 Completion statement of accounts:

I hereby irrevocably declare, warrant and undertake to the MEEF Management Committee and the Steering Committee of the relevant Funds including the Top-up Fund, that I myself, and the Organisation:

1. do not deal with, and are not in any way associated with, any country or organisation or activity which is or may potentially be relevant to, or targeted by, sanctions administered by the United Nations Security Council, the European Union, Her Majesty's Treasury-United Kingdom, the United States Department of the Treasury's Office of Foreign Assets Control, or the Hong Kong Monetary Authority, or any sanctions law applicable;
2. have not used any money obtained from the Marine Ecology Enhancement Fund or the related Top-up Fund (and any derived surplus), in any unlawful manner, whether involving bribery, money laundering, terrorism or infringement of any international or local law; and
3. have used the funds received (and any derived surplus) solely for the studies or projects which further the MEEF Objectives and have not distributed any portion of such funds (including any derived surplus) to members of the Recipient Organisation or the public.

Signature: ___________________
*Project Leader, Antoni B. Chan*

Date: ___2022/Feb/05____

Office Chop:______________

# References

[1] Construction of three-runway system kicks off at hkia. https://www.threerunwaysystem.com/en/information-centre/press-release/01082016/, Aug 2016.

[2] Jeffrey Adams, Todd Speakman, Eric Zolman, and Lori Schwacke. Automating image matching, cataloging, and analysis for photoidentification research. Aquatic Mammals, 32:374–384, 09 2006.

[3] Jay Barlow, Megan Ferguson, William Errin, Lisa I, Tim Gerrodette, Gerald Joyce, Colin Macleod, Keith Mullin, Debi Palka, and Gordon Waring. Abundance and densities of beaked and bottlenose whales 32anada32c32iphiidae). Journal of Cetacean Research and Management, 7, 01 2005.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and HongYuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[5] S. Buckland, D. Anderson, K. Burnham, Jeffrey Laake, David Borchers, and Len Thomas. Introduction to Distance Sampling: Estimating Abundance of Biological Populations, volume xv. 01 2001.

[6] Joana Castro, Francisco O. Borges, Andre Cid, Marina I. ´Laborde, Rui Rosa, and Heidi C. Pearson. Assessing the behavioural responses of small cetaceans to unmanned aerial vehicles. Remote Sensing, 13(1), 2021.

[7] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In Proceedings o$^f$ the 27th ACM International Conference on Multimedia, MM'19, New York, NY, USA, 2019. ACM.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303–338, 2010.

[9] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

[11] Emilio Guirado, Siham Tabik, Marga L. Rivas, Domingo Alcaraz-Segura, and Francisco Herrera. Whale counting in satellite and aerial images with deep learning. Scientific Reports, 9(1):14259, Oct 2019.

[12] T. Jefferson. Hong kong's indo-pacific humpback dolphins (sousa chinensis): Assessing past and future anthropogenic impacts and working toward sustainability. Aquatic Mammals, 44:711–728, 2018.

[13] Alexis Joly, Herve Go ´ eau, Herv ¨ e Glotin, Concetto Spamp- ´inato, Pierre Bonnet, Willem-Pier Vellinga, Jean-Christophe Lombardo, Robert Planque, Simone Palazzo, and Henning ´Muller. Lifeclef 2017 lab overview: Multimedia species ¨identification challenges. In Gareth J.F. Jones, Seamus Law- ´less, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, editors, Experimental IR Meets Multilinguality, Multimodality, and Interaction, pages 255–274, Cham, 2017. Springer International Publishing.

[14] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. International Journal of Computer Vision, pages 1–26, 2020.

[15] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie.

Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016.

[18] Malte Pedersen, Joakim Bruslund Haurum, Rikke Gade, and Thomas B. Moeslund. Detection of marine animals in a new underwater dataset with varying visibility. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

[19] Sarah Piwetz, Thomas A. Jefferson, and Bernd Wursig. Effects of coastal construction on indo-pacific humpback dolphin (sousa chinensis) behavior and habitat-use off hong kong. Frontiers in Marine Science, 8:196, 2021.

[20] Debora Pollicelli, Mariano Coscarella, and Claudio Delrieux. Roi detection and segmentation algorithms for marine mammals photo-identification. Ecological Informatics, 56:101038, 2020.

[21] Joseph Redmon. Darknet: Open source neural networks in c. http://pjreddie.com/darknet/, 2013–2016.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[23] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015.

[25] Vito Reno, Gianvito Losapio, Flavio Forenza, Tiziano Politi, Ettore Stella, Carmelo Fanizza, Karin Hartman, Roberto Carlucci, Giovanni Dimauro, and Rosalia Maglietta. Combined color semantics and deep learning for the automatic detection of dolphin dorsal fins. Electronics, 9(5), 2020.

[26] Heather R. Smith, Daniel P. Zitterbart, Thomas F. Norris, Michael Flau, Elizabeth L. Ferguson, Colin G. Jones, Olaf Boebel, and Valerie D. Moulton. A field comparison of marine mammal detections via visual, acoustic, and infrared (ir) imaging methods o33anada33ca33anadac canada. Marine Pollution Bulletin, 154:111026, 2020.

[27] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9627–9636, 2019.

[28] Cameron Trotter, Georgia Atkinson, Matthew Sharpe, Kirsten Richardson, A. Stephen McGough, Nick Wright, Ben Burville, and Per Berggren. NDD20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorization. CoRR, abs/2005.13359, 2020.

[29] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. International journal of computer vision, 104(2):154–171, 2013.

[30] Ursula K. Verfuss, Douglas Gillespie, Jonathan Gordon, Tiago A. Marques, Brianne Miller, Rachael Plunkett, James A. Theriault, Dominic J. Tollit, Daniel P. Zitterbart, Philippe Hubert, and Len Thomas. Comparing methods suitable for monitoring marine mammals in low visibility conditions

during seismic surveys. Marine Pollution Bulletin, 126:1–18, 2018.

[31] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 390–391, 2020.

[32] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[33] Peiqin Zhuang, Linjie Xing, Yanlin Liu, Sheng Guo, and Yu Qiao. Marine animal detection and recognition with advanced deep learning models. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, Working Notes of CL–F 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017, volume 1866 of CEUR Workshop Proceedings. CEUR-WS.org, 2017.

[34] Daniel P. Zitterbart, Heather R. Smith, Michael Flau, Sebastian Richter, Elke Burkhardt, Joe Beland, Louise Bennett, Alejandro Cammareri, Andrew Davis, Meike Holst, Caterina Lanfredi, Hanna Michel, Michael Noad, Kylie Owen, Aude Pacini, and Olaf Boebel. Scaling the laws of thermal imaging–based whale detection. Journal of Atmospheric and Oceanic Technology, 37(5):807 – 824, 2020.

[35] Gabaldon, J., Zhang, D., Barton, K., Johnson-Roberson, M. and Shorter, K.A., 2017, September. A framework for enhanced localization of marine mammals using auto-detected video and wearable sensor data fusion. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2505-2510). IEEE.

[36] Hoekendijk, J.P., de Vries, J., van der Bolt, K., Greinert, J., Brasseur, S., Camphuysen, K.C. and Aarts, G., 2015. Estimating the spatial position of marine mammals based on digital camera recordings. Ecology and evolution, 5(3), pp.578-589.

[37] Schoonmaker, J., Dirbas, J., Podobna, Y., Wells, T., Boucher, C. and Oakley, D., 2008, October. Multispectral observations of marine mammals. In Electro-Optical and Infrared Systems: Technology and Applications V (Vol. 7113, p. 711311). International Society for Optics and Photonics.

[38] Weissenberger, J., Blees, M., Christensen, J., Hartin, K., Ireland, D. and Zitterbart, D.P., 2011. Monitoring for marine mammals in Alaska using a 360 infrared camera system.

[39] Berg, Paul, Deise Santana Maia, Minh-Tan Pham, and Sébastien Lefèvre“. "Weakly Supervised Detection of Marine Animals in High Resolution Aerial”Images." Remote Sensing 14, no. 2 (2022): 339.

[40] Aniceto, Ana S., Martin Biuw, Ulf Lindstrøm, Stian A. Solbø, Fredrik Broms, and JoLynn C“rroll. "Monitoring marine mammals using unmanned aerial vehicles: quantifying detection ce”tainty." Ecosphere 9, no. 3 (2018): e02122.

[41] Han, Fenglei, Jingzheng Yao, Haitao Zhu, and Chunhu“ Wang. "Marine organism detection and classification from underwater vision based on the deep CNN”method." Mathematical Problems in Engineering, 2020.

[42] Shiu, Yu, K. J. Palmer, Marie A. Roch, Erica Fleishman, Xiaobai Liu, Eva-Marie Nosal, Tyler Helble, Danielle Cholewiak, Douglas Gillespie, and Holger “linck. "Deep neural networks for automated detection of marine mammal ”pecies." Scientific reports 10, no. 1 (2020): 1-12.

[43] Richard Szeliski. *Computer Vision: Algorithms and Applications (Texts in Computer Science)*. 2nd edition. Springer 2022.

[44] Buckland, S.T., Rexstad, E., Marques, T.A. and Oedekoven, C.S. *Distance Sampling: Methods and Applications.* Springer, Heidelberg, 2015.

# Appendix

## 1. Work logs

Dr. Anh Nguyen (Postdoc)

| Week No. | Period | Duties | No. of hours | Total no. of hours |
|---|---|---|---|---|
| 1 | 30-NOV-2020 to 4-DEC-2020 | Data annotation for motion detection | 44 | 44 |
| 2 | 7-DEC-2020 to 11-DEC-2020 | Data annotation for motion detection | 44 | 44 |
| 3 | 14-DEC-2020 to 18-DEC-2020 | Develop GoPro 8 streaming API | 44 | 44 |
| 4 | 21-DEC-2020 to 25-DEC-2020 | Develop GoPro 8 streaming API | 44 | 44 |
| 5 | 28-DEC-2020 to 1-JAN-2021 | Data annotation for motion detection | 44 | 44 |
| 6 | 4-JAN-2021 to 8-JAN-2021 | Data annotation for motion detection | 44 | 44 |
| 7 | 11-JAN-2021 to 15-JAN-2021 | Develop system prototype | 44 | 44 |
| 8 | 18-JAN-2021 to 22-JAN-2021 | Develop system prototype | 44 | 44 |
| 9 | 25-JAN-2021 to 29-JAN-2021 | Develop system prototype | 44 | 44 |
| 10 | 1-FEB-2021 to 5-FEB-2021 | Develop system prototype<br>Boat trip for data collection | 44 | 44 |
| 11 | 8-FEB-2021 to 12-FEB-2021 | Develop system prototype for data collection | 44 | 44 |
| 12 | 15-FEB-2021 to 19-FEB-2021 | Data annotation for motion detection | 44 | 44 |
| 13 | 22-FEB-2021 to 26-FEB-2021 | Data annotation for motion detection<br>Boat trip for data collection | 44 | 44 |
| 14 | 1-MAR-2021 to 5-MAR-2021 | Data annotation for motion detection | 44 | 44 |
| 15 | 8-MAR-2021 to 12-MAR-2021 | Data annotation for motion detection | 44 | 44 |
| 16 | 15-MAR-2021 to 19-MAR-2021 | Implement 3D-CNN approach for motion detection | 44 | 44 |
| 17 | 22-MAR-2021 to 26-MAR-2021 | Implement 3D-CNN approach for motion detection | 44 | 44 |

Dr. Hao Zhang (Postdoc)

| Week No. | Period | Duties | No. of hours | Total no. of hours |
|---|---|---|---|---|
| 1 | 4-JAN-2021 to 8-JAN-2021 | 1. Survey potential literature about extension application of object detection, such as dolphin photo-ID.<br>2. Prepare a survey report ("A Survey of Artificial Intelligence aided Dolphin Photos Grouping and Identification").<br>3. Cooperate to propose a joint-framework of dolphin detection and photo re-identification. | 16<br>16<br>12 | 44 |
| 2 | 11-JAN-2021 to 15-JAN-2021 | 1. Join boat trip for data collection<br>2. Help to prepare Chinese version of Abstract for Dolphin Photo-ID proposal.<br>3. Research problems of false positives of dolphin detection (waves etc.,) and propose to adopt motion patterns to filter waves from real detected dolphin. | 8<br>8<br>28 | 44 |
| 3 | 18-JAN-2021 to 22-JAN-2021 | 1. Setup torch environments on server<br>2. Modifies codes of motion classifier to distinguish dolphin and wave motions. (cont) | 8<br>36 | 44 |
| 4 | 25-JAN-2021 to 29-JAN-2021 | 1. Join boat trip for data collection<br>2. Watch and prepare video data for motion annotation. (cont) | 8<br>36 | 44 |

| | | | | |
|---|---|---|---|---|
| 5 | 1-Mar-2021 to 5-Mar-2021 | 1. Label false positives' motions (cont). | 44 | 44 |
| 6 | 08-Mar-2021 to 12-Mar-2021 | 1. Label false positives' motions (cont). | 44 | 44 |
| 7 | 15-Mar-2021 to 22-Mar-2021 | 1. Get a trained detector from Anh, and write corresponding visualization codes.<br>2. Set up a Yolov5 alternative detector for computationally speeding up. | 20<br>24 | 44 |
| 8 | 22-Mar-2021 to 27-Mar-2021 | 1. Use the accumulated false positives as extra categories for detection training.<br>2. Write visualization code for the Yolov5 detector. | 44 | 44 |
| 9 | 29-Mar-2021 to 31-Mar-2021 | 1. Visually compare detectors trained with/without false positives. | 26 | 26 |
| 10 | 1-April-2021 to 2-April-2021 | 1. Label false positives (cont). | 16 | 16 |
| 11 | 05-April-2021 to 09-April-2021 | 1. Label false positives (cont).<br>2. Re-Train Yolov5 detector with newly added data. | 44 | 44 |
| 12 | 12-April-2021 to 16-April-2021 | 1. Join boat trip for data collection<br>2. Feed collected video into Yolov5 detector to obtain dolphin detections (including both true/false positives).<br>3. Continue manually annotate false positives on top of the result in step 2. | 12<br>16<br>16 | 44 |
| 13 | 19-April-2021 to 23-April-2021 | 1. Continue label newly collected data.<br>2. Retrain Yolov5 detector, the new detector's robustness benefits from more data. | 40 | 44 |
| 14 | 26-April-2021 to 30-April-2021 | 1. Join a boat trip for data collection.<br>2. Continue to label newly collected data | 12<br>32 | 44 |
| 15 | 3-May-2021 to 7-May-2021 | 1. Join boat trip to collect dolphin data<br>2. Label false positives (cont). | 12<br>36 | 44 |
| 16 | 10-May-2021 to 14-May-2021 | 1. Label false positives (cont).<br>2. Re-Train Yolov5 detector with newly added data. | 44 | 44 |
| 17 | 17-May-2021 to 21-May-2021 | 1. Join boat trip for data collection<br>2. Write and edit WACV manuscript "Find Dolphin in the Wild: Dataset, Algorithms and Prototype System" | 12<br>32 | 44 |
| 18 | 24-May-2021 to 28-May-2021 | 1. Write and edit WACV manuscript "Find Dolphin in the Wild: Dataset, Algorithms and Prototype System" | 44 | 44 |
| 19 | 31-May-2021 to 31-May-2021 | 1. Write and edit WACV manuscript "Find Dolphin in the Wild: Dataset, Algorithms and Prototype System" | 8 | 8 |
| 20 | 1-June-2021 to 4-June-2021 | 1. Write and edit WACV manuscript "Find Dolphin in the Wild: Dataset, Algorithms and Prototype System" | 44 | 44 |
| 21 | 7-June-2021 to 11-June-2021 | 1. Write and edit WACV manuscript "Find Dolphin in the Wild: Dataset, Algorithms and Prototype System"<br>Summarize finished jobs and handover to Dr. Zhang Qi | 44 | 44 |
| 22 | 14-June-2021 to 18-June-2021 | Summarize finished jobs and handover to Dr. Zhang Qi | 44 | 44 |

Dr. Qi Zhang (Postdoc)

| Week No. | Period | Duties | No. of hours | Total no. of hours |
|---|---|---|---|---|
| 1 | 13-May-2021 to 15-May-2021 | 1. Get familiar with the project, especially the working flow of the dataset collection and labeling system. | 20 | 20 |

| | | | | |
|---|---|---|---|---|
| 2 | 17-May-2021 to 22-May-2021 | 1. Take over the rest work of Anh: system prototype design, system deployment. | 44 | 44 |
| 3 | 24-May-2021 to 29-May-2021 | 1. Join boat trip for data collection and preprocess the collected data.<br>2. Re-train and evaluate the Faster-RCNN network on the large dataset dolphin-14k (the previous model was trained on the small dataset and the evaluation metric was not suitable). | 44 | 44 |
| 4 | 31-May-2021 to 5-June-2021 | 1. Work for the WACV deadline, including dataset analysis, figure plotting, and experiments. | 44 | 44 |
| 5 | 7-June-2021 to 12-June-2021 | 1. Work for the WACV deadline: train the FCOS on the large dataset, report the performance on the same metric and running speed.<br>2. Take over the rest work from Dr. Hao Zhang. | 44 | 44 |
| 6 | 14-June-2021 to 19-June-2021 | 1. Take over the rest work from Dr. Hao Zhang. | 44 | 44 |
| 7 | 21-June-2021 to 26-June-2021 | 1. Summarize and clean the collected videos from the previous surveys. | 44 | 44 |
| 8 | 28-June-2021 to 3-July-2021 | 1. Develop the new system interface under the same platform as the model. | 44 | 44 |
| 9 | 5-July-2021 to 10-July-2021 | Develop the new system interface under the same platform as the model for single camera system. | 44 | 44 |
| 10 | 12-July-2021 to 17-July-2021 | Develop the new system interface under the same platform as the model for single camera system. | 44 | 44 |
| 11 | 19-July-2021 to 24-July-2021 | GUI development and work on the multi-GoPro camera connection issue. | 44 | 44 |
| 12 | 26-July-2021 to 31-July-2021 | Join the data collection survey, GUI development and work on the multi-GoPro camera connection issue. | 44 | 44 |
| 13 | 2-Aug-2021 to 7-Aug-2021 | Test several possible ways to stream the frames from 2 GoPro cameras to the same PC. | 44 | 44 |
| 14 | 9-Aug-2021 to 14-Aug-2021 | Test multi-camera streaming methods and update the paper for ACM MM Asia 2021. | 44 | 44 |
| 15 | 16-Aug-2021 to 21-Aug-2021 | Use another PC to capture and forward the 2nd camera frames for multi-camera system. | 44 | 44 |
| 16 | 23-Aug-2021 to 28-Aug-2021 | Go through the videos for comparison with human detection time. | 44 | 44 |
| 17 | 30-Aug-2021 to 4-Sep-2021 | Add more evaluation metrics of the results of the method (precision, recall, etc.) | 44 | 44 |
| 18 | 6-Sep-2021 to 11-Sep-2021 | Deal with the category balance issue in the current dataset distribution. | 44 | 44 |
| 19 | 13-Sep-2021 to 18-Sep-2021 | Collect data and implement on 2-stage based method. | 44 | 44 |
| 20 | 20-Sep-2021 to 25-Sep-2021 | Collect data and implement on 2-stage based method. | 44 | 44 |
| 21 | 20-Sep-2021 to 25-Sep-2021 | Test and compare the 2-stage based method. | 44 | 44 |
| 22 | 27-Sep-2021 to 2-Oct-2021 | Test and compare the 2-stage based method. | 44 | 44 |
| 23 | 4-Oct-2021 to 9-Oct-2021 | Go through the latest data for comparison with the human records. | 44 | 44 |
| 24 | 11-Oct-2021 to 16-Oct-2021 | Update the 2-stage based method and adjust the training method. | 44 | 44 |
| 25 | 18-Oct-2021 to 23-Oct-2021 | Develop and test the system on the Notebook PCs. | 44 | 44 |
| 26 | 25-Oct-2021 to 30-Oct-2021 | Develop and test the system on the Notebook PCs, and compare the running speed with different streaming strategies. | 44 | 44 |
| 27 | 1-Nov-2021 to 6-Nov-2021 | Collect data and adjust the model setting on the notebook PCs for better testing performance. | 44 | 44 |

| | | | | |
|---|---|---|---|---|
| 28 | 8-Nov-2021 to 13-Nov-2021 | Combine the current GUI with the model together and test on the notebook PC. | 44 | 44 |
| 29 | 15-Nov-2021 to 20-Nov-2021 | Develop the system interface and modify the paper for camera-ready submission. | 44 | 44 |
| 30 | 22-Nov-2021 to 27-Nov-2021 | Develop the system interface and add more functions, and prepare for the field test. | 44 | 44 |
| 31 | 29-Nov-2021 to 4-Dec-2021 | The first field test, and process the data and evaluate the performance. | 44 | 44 |
| 32 | 6-Dec-2021 to 11-Dec-2021 | Evaluate the performance compared with human detection time and the conduct the second field test. | 44 | 44 |
| 33 | 13-Dec-2021 to 18-Dec-2021 | Process the data and evaluate the performance. | 44 | 44 |
| 34 | 20-Dec-2021 to 25-Dec-2021 | Evaluate the field test performance, and write the report draft for the project. | 44 | 44 |
| 35 | 27-Dec-2021 to 31-Dec-2021 | Write the final report for the project. | 44 | 44 |

## 2. List of Project Assets

List of project assets is not disclosed due to confidentiality reasons.

# 3. Recruitment Records

## Job advertisements

Recruitment records are not disclosed due to confidentiality reasons.

# 4. Financial Audits

Details of financial audits are not disclosed due to confidentiality reasons.

Project Title: Automated CWD Detection with Machine Learning for Vessel-based Line-transect Survey
MEEF Project No.: MEEF2020006
Period of the Completion Report: 1 July 2020 to 31 December 2021

I hereby irrevocably declare to the MEEF Management Committee and the Steering Committee of the relevant Funds including the Top-up Fund, that I myself as the person in charge of the ~~accounting~~ / finance* department of the Recipient Organisation, and confirms that:

(i) the books and records of the Recipient Organisation has been properly kept for the reporting period, and

(ii) the financial statement enclosed in the completion report has been prepared in accordance with the financial reporting requirements prescribed by the MEEF.

________________________________

Signed by Mr. Alfred Chau
as the Associate Director of Finance
For and on behalf of City University of Hong Kong

Date: 23 FEB 2022

City University of Hong Kong * 香港城市大學 * FO

Official Chop: __________________________

** Please delete as appropriate*