國立臺灣師範大學

理學院資訊工程學系碩士論文

指導教授：方瓊瑤 博士

基於深度學習之鯨豚個體身分辨識系統

Cetacean Individual Identification System Based on Deep Learning

研究生 ： 蔡妤涓 撰

中華民國一百一十三年一月

摘要

本研究提出一個基於深度學習之鯨豚個體身分辨識系統，希望透過鯨豚個體身分辨識的技術，追蹤鯨豚遷徙路徑來估算鯨豚族群數量，進一步評估和保護海洋生態系統的健康。研究目標為辨識同一物種內不同鯨豚個體的生物特徵，以及同一隻鯨豚在不同拍攝環境下的影像特徵差異。由於鯨豚資料集中存在影像品質不穩定和個體影像數量極不平均的問題，故本研究著手解決這些問題，包含資料前處理(Data Preprocessing)、提出模型改良方法，及不同面向的測試方法。

本系統首先對鯨豚資料集進行資料前處理，接著進行鯨豚偵測，最後作鯨豚個體身分辨識。資料集前處理包括資料清理(Data Cleaning)和資料增強(Data Augmentation)，其目的在解決資料集中的潛在問題。在鯨豚偵測階段，採用 YOLOv5 定位鯨豚位置，過濾背景雜訊以增加模型訓練速度。在鯨豚個體身分辨識階段，利用骨幹模型(Backbone Model)從鯨豚影像中提取特徵，並使用頭部模型(Head Model)進行個體身分預測。本研究使用 EfficientNetV1-B4 作為骨幹模型，頭部模型使用附加角度邊界損失函數(ArcFace)。針對資料集問題對頭部模型進行改良，以提高鯨豚個體身分辨識的正確率。透過在 ArcFace 加入子中心(Sub-center)向量，解決同一隻鯨豚在不同拍攝環境下的影像特徵差異的問題，從而提升鯨豚個體身分辨識的正確率。此外，引入動態邊界(Dynamic Margin)解決在訓練階段鯨豚個體影像數量極不平均的問題，加快模型的收斂速度。

實驗結果顯示改良後的子中心附加角度邊界損失函數在三個面向的測試實際應用情況、多數合成資料庫(Synthetic Data)，和部分合成資料庫(影像數量 3 張以上的鯨豚個體)之 mAP 分別為 68.63%、81.60%和 35.70%。相較於原始的 ArcFace 提升 4.83%、6.08%和 8.19%。另外，將動態邊界應用於子中心附加角度邊界損失函數的改良方案，在維持相當正確率相當的情況下，減少 28%的訓練時間。由實驗結果發現，本研究所提出的改良方案能對資料集問題進行適當處理並提升鯨豚個體身分辨識的準確率。

關鍵詞: 鯨魚、海豚、個體身分辨識、深度學習、影像檢索、附加角度邊界損失函數、動態邊界應用於子中心附加角度邊界損失函數

Abstract

This research presents a system based on deep learning for individual cetacean identification. It aims to track cetacean migration paths and estimate their population numbers for assessing and maintaining the health of marine ecosystems. The study focuses on distinguishing individual biological characteristics from images of cetaceans of the same species and those of the same cetaceans captured in different environments. To address the issues with the dataset, such as unstable image quality and a highly uneven distribution of individual images, the research focuses on data preprocessing, model improvement, and comprehensive testing methods.

First, the cetacean dataset is preprocessed to achieve clean data. Subsequently, cetacean detection is performed using YOLOv5 to identify cetaceans and filter background noise, followed by cetacean individual identification. EfficientNetV1-B4 is chosen as the backbone model, and the Additive Angular Margin Loss (ArcFace) is adopted for the head model. Incorporating sub-centers into ArcFace addresses the problem of different image features of the same cetacean under varying environments, thus improving identification accuracy. Moreover, the introduction of dynamic margins in sub-center ArcFace deals with the uneven distribution of individual images during training, enhancing the model's convergence speed.

Experimental results show that the improved sub-center ArcFace achieves higher mAP scores across three testing scenarios: real-world application, majority synthetic dataset, and partial synthetic dataset (individuals with more than three images). Compared to the original ArcFace, mAP improves by 4.83%, 6.08%, and 8.19%, respectively. Additionally, applying sub-center ArcFace with dynamic margins maintains similar accuracy levels while reducing training time by 28%. The findings indicate the effectiveness of the proposed improvements in handling dataset issues and improving the accuracy of cetacean individual identification.

Keywords: Whale, Dolphin, Individual Identification, Deep Learning, Image Retrieval, Additive Angular Margin Loss, and Sub-center Additive Angular Margin Loss with Dynamic Margin.

致謝

在本篇論文的研究與撰寫過程中，我得到了許多人的幫助和支持，對此我深表感激。首先，我要感謝我的指導教授方瓊瑤教授，感謝她對於研究和未來的人生方向上一路的指導和支持。在研究過程中，教授獨特的見解和專業的建議都是我完成論文的最大助力。她不斷教導我從實驗失敗的結果中學習，並進行調整與改進。同理而言，對人生也是如此，她讓我理解到每次的挫折都是進步的養分。她對學術研究的堅持與嚴謹，以及對實驗結果保持正向開放的態度都是我學習的典範。接著我要感謝共同指導教授陳世旺教授，感謝他持續提供不同面向的研究探討與建議，身體力行地向我展示終身對研究探討與知識學習的熱忱。

我還要感謝 CVIU 實驗室的所有成員。感謝孟霖、家安、佑如學長姊對我在專題生時期的指導與啟蒙。感謝后玲、日棠、秉琛學長姐們為我樹立良好的研究榜樣。感謝雅雯、展兢、柏恩、哲緯、育德、柏丞、聖傑、鈺瑄同學們給予的研究建議。

感謝在羽汝、雨平、孟凡和桃園高中 306 的同學們，在遭遇研究瓶頸時給予我心靈上的安慰，提供我繼續做研究的力量。同時他們對待生活和工作認真的態度也提醒我需時刻努力。

非常感謝柏恩、佩恩、雅雯、后玲在碩士期間的互相對我的關照。無論是生活煩惱、研究內容、未來方向都給予我非常多實用的建議與鼓勵，促使我成為更好的人。

最後，我要對我的家人表示最深的感謝。感謝父母對我的學業和生活提供了無條件的支持和關愛，感謝弟弟總是默默的支持我的每個選擇，你們是我溫暖的避風港。感謝所有家人們的付出，讓我無後顧之憂專心做研究。特別感謝已在天上的致愛外公，你對我的期待和支持是我做研究的最大動力。

感謝所有在這段旅程中支持我的人，沒有你們就沒有今天的我。

蔡妤涓 謹致

國立臺灣師範大學 資訊工程學系研究所

中華民國 113 年 1 月

Contents

List of Figures

List of Tables

1 Introduction

1.1 Research Motivation

The ocean, which covers 70.9% of the earth's surface [1], plays a crucial role in climate regulation and supports a vast range of ecosystems and species. Cetaceans, including whales and dolphins, are important bioindicators to monitor ocean health. Our research utilizes deep learning techniques to automatically identify individual cetaceans to monitor cetacean migration paths and manage marine ecosystems. The identification of individual cetaceans can help predict cetacean populations. This information can further be used to assess ocean health, leading toward a sustainable future for the ocean and all living beings.

The Ocean Health Index (OHI) [2] is a comprehensive quantitative framework for measuring and monitoring the health of the human–ocean systems. Cetacean conservation contributes significantly to two of the public goals of OHI, namely carbon storage and biodiversity.

Cetaceans, recognized as crucial ecosystem engineers [3], significantly contribute to carbon storage and biodiversity, ensuring the health and stability of marine environments. Figure 1 illustrates the relationship between cetaceans, carbon storage, and biodiversity. In terms of carbon storage, cetaceans accumulate carbon in their bodies throughout their long lifespans. Once they die, the carcasses descend to the ocean floor. On average, each great whale sequesters approximately 33 tons of carbon [4]. The presence of cetaceans increases phytoplankton, in turn enhancing the biodiversity of the ocean. Cetaceans participate in the transportation of minerals to the ocean's surface through their vertical movements and migrations across vast ocean expanses. The nutrient transport contributes to the extension of phytoplankton blooms. The phytoplankton, supported by nutrient-rich whale pumps, serve as a vital food resource for a wide variety of species, ultimately promoting biodiversity in the region.

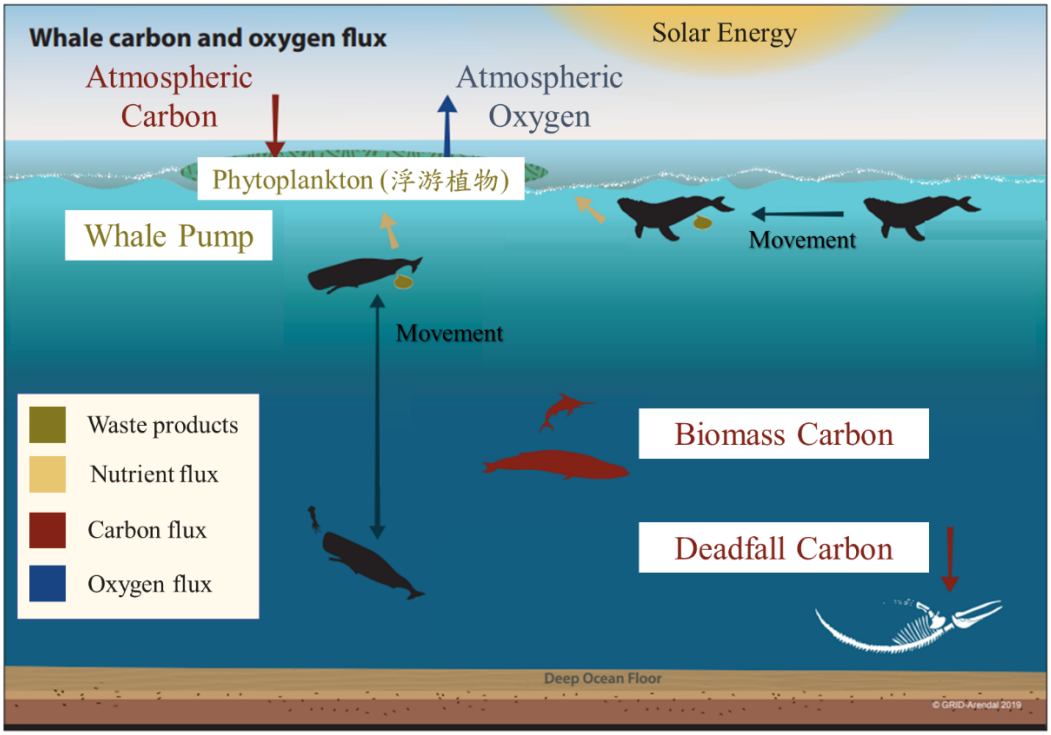


Figure 1. Whale carbon and oxygen flux. Adapted from [4].

Many scientists, recognizing the significance of cetaceans, track them to assess and safeguard ocean health. Biologists track the location and timing of each individual cetacean appearances in the ocean to estimate cetacean population. Cetacean tracking relies on individual identification through either the use of physical tags or image matching. Physical tagging is an invasive method that involves anchored tags, bolt-on tags, and consolidated electronic tags. However, the process of capturing and installing tags is time-consuming and laborious, and it can easily cause physical and mental harm to cetaceans. By contrast, the image matching approach for individual cetacean identification is a non-invasive method that minimizes the human involvement and potential disturbance to the animals.

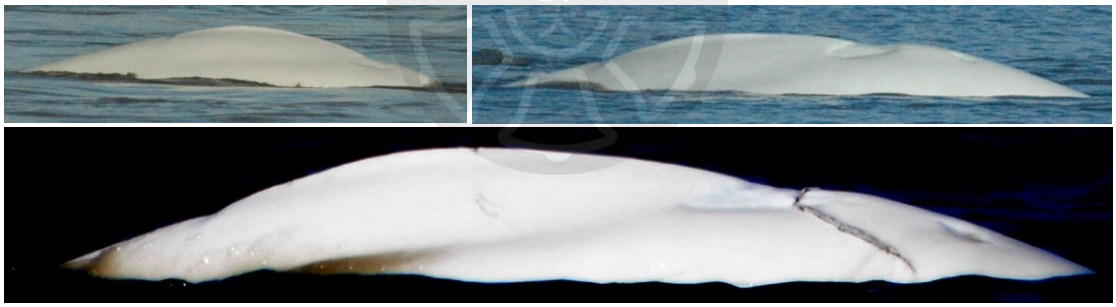
Our research aims to develop an automated cetacean individual identification system based on deep learning to track the location and timing of each individual cetacean. With non-invasive image identification, individual cetaceans can be tracked more efficiently while minimizing the risk of physical and psychological harm to cetaceans. The cetacean tracking system provides valuable data for natural scientists to study and protect marine life and promotes the protection and sustainable use of marine resources.

1.2 Research Difficulties

Our research utilizes the cetacean image dataset provided by Happywhale [5]. Cetacean images can be input into the Happywhale dataset to obtain their corresponding cetacean ID numbers.

Figure 2 shows some cetacean images in the Happywhale dataset, including a dusky dolphin, a beluga, and a gray whale. Figure 2 (a) shows the dorsal fin curves of the dusky dolphin with ID 0c0adcf55bf6. Figures 2 (b) and (c) show the body notches of the beluga with ID afb9b3978217 and the body markings of the gray whale with ID a8c9dfb8ac6f, respectively.

(a) Dusky Dolphin, ID: 0c0adcf55bf6



(b) Beluga, ID: afb9b3978217

(c) Gray Whale, ID: a8c9dfb8ac6f

Figure 2. Images in the Happywhale dataset.

Some of the challenges that should be addressed while developing a cetacean individual identification system include the following: (1) insufficient and imbalanced image count of individuals; (2) cetacean pixels account for a small portion of the entire image; (3) slight variations in the characteristics of the same species but different individuals; (4) variation in the image quality of the same cetacean in different image-capturing conditions; and (5) potential uncertainty in image labeling. The challenges are discussed in detail below.

(1) Rare and imbalanced image count of individuals

The Happywhale dataset contains 51,033 cetacean images, including a total of 15,587 unique cetacean identities. Among these identities, 9,258 cetaceans are associated with only one image. Additionally, 3,091 cetaceans are linked to two images. These statistics reveal that approximately 79% of the individual cetaceans have only one or two images available for identification.

In this study, the problem of uneven distribution of individual cetaceans is addressed through data augmentation and modifications to the head model. Data augmentation is performed for cetacean individuals with fewer than 15 images, ensuring that each individual has a total of 10 images for training and 5 images for validation and testing. Moreover, the head model is modified to incorporate dynamic margins in ArcFace. This allows for the imbalanced dataset to converge more effectively by adaptively learning the margins based on the number of identities.

(2) Cetacean pixels account for a small proportion of the entire image

Random sampling of the Happywhale dataset indicates that approximately 43% of the cetacean bounding boxes in the images occupy less than 80% of the entire image area. The non-cetacean information in the images may cause instability to the individual cetacean identification system. Figure 3 illustrates examples of the cetacean bounding box areas that are less than 10%, between 10% and 80%, and greater than 80%.

The proposed system employs YOLOv5 to detect the cetacean bounding boxes in the images. The objective is to enable the system to concentrate on identifying cetaceans while filtering out unnecessary background information, thereby enhancing training efficiency of the proposed system.

(a) (b) (c)

Figure 3. Examples of different ratios of the cetacean bounding box to the entire image area: (a) ratio = 2%, (b) ratio = 23%, (c) ratio = 82%

(3) Slight variations in the characteristics of the same species but different individuals

Cetaceans of the same species often exhibit similar body shapes. Consequently, during the cetacean identification process, it is crucial to carefully consider subtle differences in identification biometrics. These subtle changes, which might not be immediately noticeable to the human eye, encompass features such as body markings, notches, and the curvature of the dorsal fin (refer to Figures 4 and 5).

Our system solves this problem by employing EfficientNet as the backbone model for feature extraction and the Additive Angular Margin Loss Module, ArcFace, as the head model for individual cetacean identification. ArcFace utilizes the features extracted by EfficientNet, maps them onto the hypersphere, and adds a margin penalty to arrange features of cetaceans with the same identity closer and those of different identities farther apart. The process facilitates the identification of subtle biometric differences among individuals of the same cetacean species.

Figure 4. Subtle variations exist among different identities of spotted dolphin ID: 1f1913e886bb (left), 29a5948afd72 (middle), c9673ac23c88 (right)

Figure 5. Subtle variations exist among different identities of blue whales
ID: efb454619ccd (up), 349e26968b33 (middle), da6ec08d049e (down)

(4) Variation in the image quality of the same cetacean in different image-capturing conditions

Differences in the conditions of image capturing of the same cetacean can impede identification. Various factors can influence identification results, including differences in shooting angles, lighting conditions, image resolution of the parts and proportions of cetaceans exposed to the ocean surface, and the color and ripples of the waves (as shown in Figure 2(c)).

Applying the sub-center information in the loss function of ArcFace is crucial for preventing the dominance of specific features. For example, in cases where most images capture the side view and only a few are taken from the front, the frontal features are potentially ignored. Selecting multiple sub-center vectors as representatives for the same individuals can effectively recognize non-dominant features, while filtering noisy data.

(5) Potential uncertainty in image labeling

In the Happywhale dataset, some images contain more than two cetaceans (Figure 6), but only one ID label is provided for the entire image. Furthermore, the Happywhale dataset includes images that are too cluttered (Figure 7) or blurry (Figure 8) or comprising extremely small cetacean area (Figure 9). These types of

images are collectively referred to as defective images. The dataset includes 8,073 defective images, accounting to approximately 16% of the entire dataset.

Our system is highly sensitive to minor feature differences. Hence, in this study, dataset cleaning is performed to avoid using defective images to train and test the cetacean individual identification system.

Figure 6. Examples of images containing more than two cetaceans.

Figure 7. Examples of excessively cluttered images.

Figure 8. Examples of too blurry images.

Figure 9. Examples of images with extremely small cetacean area.

1.3 Research Contribution

This study develops a deep-learning system for individual cetacean identification. First, the study undertakes data cleaning and augmentation on the dataset to avoid using defective images for training and testing. Subsequently, the system employs YOLOv5 for cetacean object detection, leverages EfficientNet as the backbone model, and integrates ArcFace as the head model for individual cetacean identification. The contributions of this study are as follows.

(1) Individual cetacean identification

This study proposes a novel individual cetacean identification system that uses YOLOv5 to detect the cetacean displayed in an image. The system uses EfficientNet as the backbone model to extract cetacean features and ArcFace as the head model to identify the cetacean.

In this study, the performance of various backbone and head models were observed and measured. In terms of the backbone model, ConvNeXt and EfficientNet were used for testing, and the SE module in EfficientNet was replaced with the CBAM module and its feasibility was analyzed. Moreover, as the head model, we used ArcFace improved through a sub-center technique and a dynamic margin approach.

(2) Handling unstable quality and imbalanced dataset

To handle images of unstable quality, data cleaning was utilized to eliminate defective images from the dataset. YOLOv5 was used in cetacean detection to filter out the background. Sub-center ArcFace was employed for model regularization to distinguish subtle cetacean individual features. Data augmentation addresses the issue of imbalanced datasets and low counts of images of individual cetaceans. Sub-center ArcFace with dynamic margins was utilized during model training to set larger margins between different cetaceans in case of individuals with low image counts.

(3) Evaluation methods for rare and imbalanced dataset

To analyze the experimental results, cetaceans were divided into four categories based on the number of images corresponding to the cetaceans in the dataset. Three evaluation methods were proposed, namely evaluation of real-world applications, majority synthetic dataset (for all categories of individuals), and partial synthetic dataset (for individuals with three or more images). This approach allows for a comprehensive assessment under varying conditions, ensuring a thorough understanding of the model's performance across different data subsets.

2 Related Work

2.1 Research Background and Method

This research is inspired by the models and processes shared by participants in the Kaggle competition, with a particular focus on the techniques employed by the first-place winners [Pat23]. The individual cetacean identification system is designed with two stages: cetacean object detection from images and individual object identification. The study presents relevant literature related to the two stages in the following sections: (1) YOLO series for object detection; (2) Backbone series for feature extraction; (3) Head series for individual identification.

2.2 YOLO series for Object Detection

Cetacean object detection technique should be utilized before individual identification to filter the background environment in the images. Based on the successful approaches of most winning participants, YOLOv5 has shown effective results in the subsequent identification of cetaceans. Thus, we selected YOLOv5 in this study. A brief overview of YOLOv1 to YOLOv5 is presented below.

In 2015, Redmon et al. proposed You Only Look Once (YOLO) real-time object detection, which was later called YOLOv1 [Red16]. YOLOv1 stands out as a single-stage object detection model, approaching detection through a regression problem. Figure 10 illustrates the YOLOv1 system model, which predicts both the location and class of the bounding boxes for objects simultaneously. The final detections are determined through non-maximum suppression (NMS). This design allows the network to conduct global reasoning across the entire image, enabling real-time detection of all objects present.

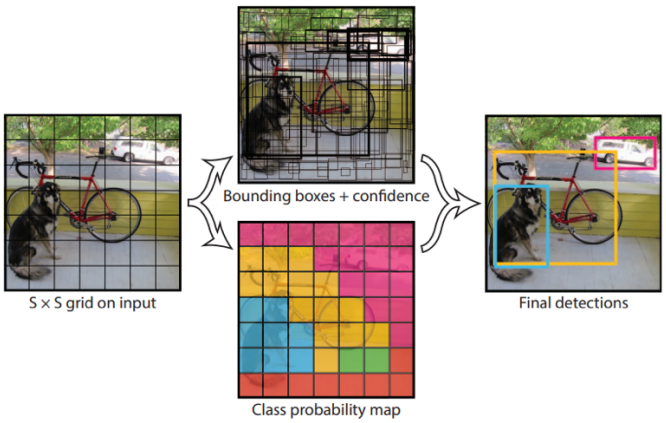


Figure 10. YOLOv1 system model [Red16].

YOLOv2 [Red17], also known as YOLO9000, is an improvement over YOLOv1 proposed by Redmon et al. in 2016. YOLOv2 retains the real-time detection capability and high efficiency of YOLOv1 while enhancing accuracy. Key improvements include adopting Darknet-19 as the backbone model and incorporating Batch Normalization (BN) [Lof15] and the Region Proposal Network (RPN) [Ren15] as multi-scale training methods.

In 2018, Redmon et al. introduced two key enhancements to YOLOv3 [Red18], namely Darknet-53 and the Feature Pyramid Network (FPN) [Lin17]. They enhance the accuracy of various size object detection by extracting multi-scale deep convolution features. Darknet-53 eliminates all max pooling layers, increasing the number of convolution layers to 53. In addition, the residual block concept from ResNet [He16] is incorporated into Darknet-53 to address the problem of gradient disappearance in deep neural networks. In the second improvement, FPN integrates feature maps obtained by down-sampling three scales to detect objects of various sizes.

In 2019, Bochkovskiy et al. introduced YOLOv4 [Boc20], enhancing the accuracy of object location detection and recognition by refining the backbone model, neck model, and loss function of the neural network. The YOLOv4 backbone model utilizes the Cross Stage Partial (CSP) blocks [Wan20] to reduce the neural network operation. The neck model incorporates Spatial Pyramid Pooling (SPP) blocks [He15] and Path Aggregation Network (PAN) blocks [Liu18] for multi-scale feature aggregation. The location loss component is modified from Mean Square Error to Complete Intersection over Union (CIoU), considering the distance and aspect ratio between the ground truth and the predicted bounding box, to enhance the accuracy of object positioning. In the NMS method, Intersection over Union (IoU) is replaced with Distance Intersection over Union (DIoU) to address overlapping objects. DIoU considers two types of distance, one is the distance between the center points of the ground truth and the predicted bounding boxes, and the other is the diagonal distance of the minimum closed area between them.

YOLOv5 [Joc20], developed by Jocher in 2020, is an enhancement and framework implementation based on YOLOv4. The primary enhancement lies in implementation, as YOLOv5 is ported to the PyTorch framework, enhancing the usability of YOLOv5. The incorporation of the CSP block in both the backbone and neck models improves feature learning by propagating unchanged feature maps. This not only reduces network parameters but also enhances computational speed.

Over model development from YOLOv1 to YOLOv5, there is a notable emphasis on handling multi-scale information, integrating features, and enhancing the versatility and efficiency of the model. In this study, given the specific focus on improving individual cetacean identification, we decided to utilize the YOLOv5 pretrained weights provided by the participants [6] [Pat23] of the competition.

2.3 Backbone series for Feature Extraction

In our system, the backbone model of identification plays a role in extracting meaningful features from cetacean images, represented as feature embeddings. The enhanced feature extraction method significantly improves the effectiveness of calculating feature similarity (the distance between feature embeddings) for individual identification in the subsequent step. This study conducts a literature review on various backbone series, including DenseNet, ConvNeXt series, and EfficientNet series. The aim is to understand and select appropriate backbone models for feature extraction, contributing to the enhancement of the effectiveness in individual identification.

2.3.1 Densely Connected Convolutional Network (DenseNet)

In 2017, Huang et al. proposed DenseNet [Hua17] in which the dense connection mechanism between layers is utilized through Dense Blocks to enhance the performance of deep neural networks. Dense connection implies that the outputs of each layer are connected to the inputs of all previous layers. This design has the advantage of alleviating the vanishing gradient problem. In addition, through the dense connection mechanism, models can effectively utilize features at different levels, contributing to the improvement of feature expressiveness. With the dense connection mechanism in DenseNet, the outputs of each layer are connected to the inputs of all previous layers, which requires a large number of parameters and high operational costs.

2.3.2 ConvNeXt series

In 2021, Liu et al. introduced the Swin Transformer [Liu21], a visual hierarchical Transformer model that utilizes shifted windows. The Swin Transformer is designed to provide flexibility for modeling at various scales. The shifted window mechanism enhances efficiency by constraining self-attention computation to non-overlapping local windows. The Swin Transformer exhibits computational complexity linearly related to the image size, making it suitable as a general backbone model for tasks such as image classification and object detection.

Liu et al. drew inspiration from the Swin Transformer and developed ConvNeXt [Liu22] in 2022, also called ConvNeXtV1. ConvNeXtV1 is constructed with a convolutional neural network (CNN), which competes robustly with state-of-the-art hierarchical vision transformers across various computer vision benchmarks. Importantly, ConvNeXtV1 retains the inherent simplicity and efficiency of the CNN.

In 2023, Woo et al. introduced ConvNeXtV2 [Woo23] by enhancing ConvNeXtV1 by including self-supervised learning technology and architectural improvements. ConvNeXtV2 incorporates a fully convolutional masked autoencoder framework and a global response normalization layer, augmenting inter-channel feature competition. These enhancements have resulted in performance improvements across various tasks, such as classification, object detection, and segmentation. Therefore, this study adopts ConvNeXtV2 as the backbone model for cetacean individual identification.

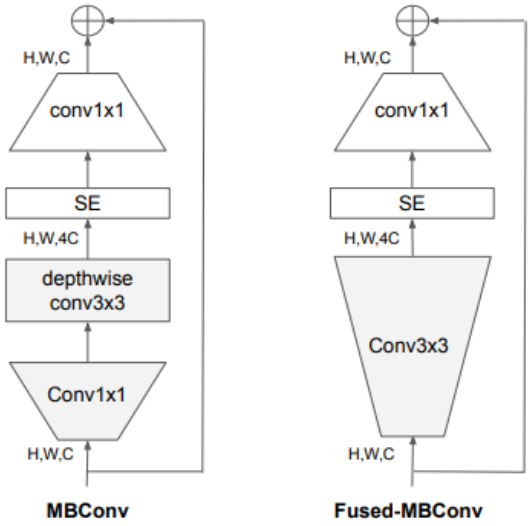
2.3.3 EfficientNet series

Hu et al. proposed the Squeeze-and-Excitation (SE) Network (SENet) [Hu18] in 2017 to enhance the performance of deep neural networks through the incorporation of a channel attention mechanism. The SE module in the SENet represents a general approach that can be seamlessly integrated into existing neural networks. EfficientNet leverages the SE module to improve the performance of neural networks, emphasizing the enhancement of channel-wise feature learning.

In 2019, Tan et al. introduced EfficientNet [Tan19], a neural network known for its high efficiency, lightweight design, and accuracy; it is also called EfficientNetV1. The EfficientNet baseline network (EfficientNet-B0) is composed using a Neural Architecture Search (NAS) method, which optimizes both accuracy and floating-point operations per second (FLOPS). Figure 11 (a) illustrates the main building block of EfficientNet-B0s; it comprises mobile inverted bottleneck MBConv, containing a depth-wise convolution layer [Cho17] and the SE module. The compound scaling was applied to EfficientNet-B0, which uniformly scales the network depth, width, and resolution of the model following the resource constraints, to obtain EfficientNet-B1 to B7.

In 2021, as an enhancement of EfficientNetV1, Tan et al. introduced EfficientNetV2 [Tan21]. With the enhancements of the training-aware NAS and progressive learning method, EfficientNetV2 featured faster training speed and improved parameter efficiency. Key modifications in EfficientNetV2 architecture involve transitioning from MBConv to fused-MBConv for early three stages (refer to

Figure 11 (b)). In addition, the progressive learning method allows adaptive adjustments to regularization based on image size.



(a) (b)

Figure 11. Main building blocks of EfficientNet. (a) MBConv [Tan19]; (b) Fused-MBConv [Tan21].

Utilizing the baseline network optimized by NAS with lightweight modules in EfficientNetV1 and training-aware improvements in EfficientNetV2, this study adopts EfficientNet series as the backbone model for the cetacean individual identification model, facilitating thorough performance observation. This study compares the training efficiency and the feature extraction performance of the backbone models, namely ConvNeXt series and the EfficientNet series.

2.4 Head series for Individual Identification

After cetacean feature extraction using the backbone model, our system utilizes ArcFace as the head model for individual identification. This study conducts a literature review on ArcFace and sub-center ArcFace with dynamic margins. The aim is to understand and select appropriate head model for cetacean individual identification.

2.4.1 Additive Angular Margin Loss (ArcFace)

In 2018, Deng et al. [Den22] proposed ArcFace, a margin-based mechanism for the face recognition task. ArcFace incorporates additive angular margins into the loss function to stabilize the training process and enhance the discriminative power of the Softmax loss. Softmax loss (L_1) is a well-known classification loss function, given by

$$L_1 = -\log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{N} e^{W_j^T x_i + b_j}}. \tag{2.1}$$

Here $x_i \in \mathbb{R}^d$ represents the cetacean feature of the i-th sample with its corresponding y_i-th identity, where i ranges from 1 to N. Symbol d denotes the dimension of the embedding feature and N represents the total number of cetacean identities, respectively. $W_j \in \mathbb{R}^d$ is the j-th column of the individual weight $W \in \mathbb{R}^{d \times N}$, and $b_j \in \mathbb{R}^N$ is the bias term, where j ranges from 1 to N. The limitation of the Softmax loss lies in its lack of specific optimization for feature embedding. This indicates that it does not effectively enhance the discriminative capability of feature embedding, causing difficulties in distinguishing the similarity among same identities and the diversity from different identities. The limitation of the Softmax loss reduces the recognition performance of individual cetacean identification when dealing with large variations in the appearance and background within samples of the same identity.

For simplicity, the bias is set to zero and the logit is transformed as $W_j^T x_i = \|W_j\| \|x_i\| \cos\theta_j$, where θ_j represents the angle between the individual weight W_j and cetacean feature x_i. θ_{y_i} represents the angle between the ground-truth weight W_{y_i} and cetacean feature x_i. Through ℓ_2 normalization, $\|W_j\| = 1$ and $\|x_i\| = 1$ are fixed to make the predictions focus on the angle between the model weight and the cetacean feature. The learned embedding features $\|x_i\|$ are re-scaled to s, implying that they are distributed on a hypersphere with a radius of s. Therefore, Equation (2.1) can be rewritten as Equation (2.2).

$$L_2 = -\log \frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{N} e^{s\cos\theta_j}}. \quad (2.2)$$

An additive angular margin penalty m is introduced between x_i and W_{y_i} to simultaneously enhance the similarity among same identities and the diversity from different identities. Since the additive angular margin penalty is equal to the geodesic distance margin penalty in the normalized hypersphere, the method is named ArcFace. The ArcFace loss (L_3) considering the angular margin penalty m can be shown as Equation (2.3).

$$L_3 = -log \frac{e^{s\,cos(\theta_{y_i}+m)}}{e^{s\,cos(\theta_{y_i}+m)} + \sum_{j=1, j\neq y_i}^{N} e^{s\,cos\,\theta_j}}. \quad (2.3)$$

Den et al. proposed sub-center ArcFace [Den22] to improve the robustness against label noise. In this study, the label noises of the images in the dataset include different parts of cetaceans (such as dorsal fins and bodies) as well as various image-capturing conditions (such as distance, angle, light source, and background). Using K sub-center vectors means each individual can be represented by more than one center feature vectors. The advantage of sub-center ArcFace lies in its ability to effectively differentiate features of the same individual across diverse image-capturing conditions, while preventing the dominance of specific features in the same individual.

Figure 12 illustrates the flowchart of ArcFace of K sub-centers for each identity. The cetacean feature embedding $x_i \in \mathbb{R}^{512}$ and sub-centers $W \in \mathbb{R}^{512\times N\times K}$ are input into the ArcFace system. An ℓ_2 normalization is used on both feature embedding x_i and all sub-centers W. The subclass-wise similarity scores $\mathcal{S} \in \mathbb{R}^{N\times K}$ can be obtained by matrix multiplication $W^T x_i$. Subsequently, a max pooling step is applied to the subclass-wise similarity scores $\mathcal{S}$ to obtain the class-wise similarity score $\mathcal{S}' \in \mathbb{R}^{N\times 1}$. The class-wise similarity score $\mathcal{S}'$ is then transformed through an arccosine function to get angle θ_{y_i} between the feature x_i and the ground-truth center. Then, a margin penalty is added on angle θ_{y_i}. $cos(\theta_{y_i} + m)$ is calculated and scaled by s. Finally, the prediction result is obtained after the Softmax function, followed by the ground-truth one-hot vector to compute cross-entropy loss.

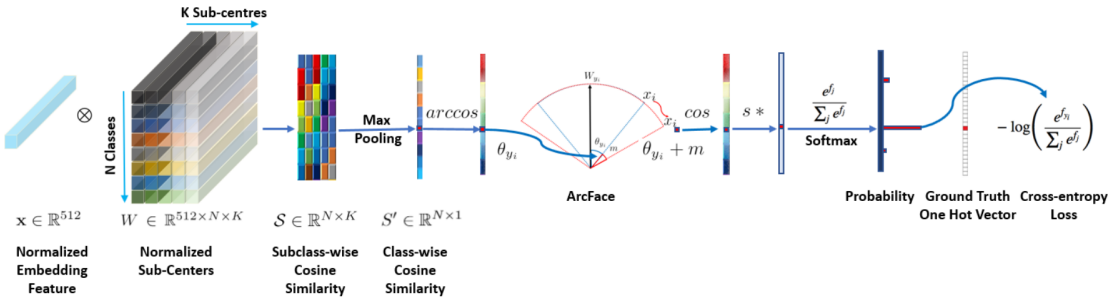


Figure 12. ArcFace system flowchart of sub-center ArcFace [Den22].

The proposed sub-center ArcFace loss (L_4) can be formulated as Equation (2.4).

$$L_4 = -log\frac{e^{s\cos\left(\theta_{y_i}+m\right)}}{e^{s\cos\left(\theta_{y_i}+m\right)}+\sum_{j=1,j\neq y_i}^{N} e^{s\cos\theta_j}}, \tag{2.4}$$

where

$$\theta_j = arccos\left(max_k\left(W_{jk}^T \mathrm{x}_i\right)\right), k \in \{1, \dots, K\}.$$

Sub-center ArcFace with dynamic margins was proposed by Ha et al. in 2020 [Ha20]. The aim is to let datasets with imbalanced data converge better by the adaptive learning of margins based on the number of identities. Since identities with a low count of images are more difficult to learn, larger margins are used for learning. On the contrary, smaller margins are utilized for identities with greater number of images. Equation (2.5) shows the formula to adjust margins according to the number of identities.

$$m_i = a \cdot {n_i}^{-\lambda} + b, \tag{2.5}$$

where m_i is the margin value for identity i, n_i is the number of images for identity i. Parameters a and b represent the upper and lower bounds of each margin function, and λ is the rate of decay to b as n_i grows.

An ablation study is conducted to compare the training efficiency and identification performance of the head models ArcFace and sub-center ArcFace with dynamic margins.

3 Research Method

The proposed cetacean individual identification system consists of two stages. The first stage is cetacean detection from the input images. The second stage is individual cetacean identification using a combination of backbone and head models.

Previous research [Pat23] has shown the effectiveness of YOLOv5 for cetacean detection, alongside EfficientNet as the backbone model and ArcFace as the head model. Together discriminating results can be obtained in identifying individual cetaceans. This study further enhances the identification results through three methods. First, data preprocessing, which involves cleaning defective images and data augmentation for cetacean identities with low image counts. Second, focusing on selection and architecture modification of both backbone and head models. Finally, proposing comprehensive evaluation methods for identification in an imbalanced dataset. This chapter presents an overview of the system and methodologies employed to improve the accuracy of cetacean individual identification.

3.1 System Overview

Figure 13 illustrates the flowchart of the proposed individual cetacean identification system. The system is divided into training and inference phases. In the training phase, the system first undergoes data cleaning and augmentation to prepare the cetacean dataset. The cetacean dataset is input for cetacean detection, and then cetacean identification is performed to output individual retrieval results. In the inference phase, a cetacean input image undergoes detection followed by identification to predict the identity of the cetaceans in the image.

More specifically, during the training phase, the cetacean dataset undergoes detection to filter out background and noise using YOLOv5, followed by individual cetacean identification. The identification process utilizes EfficientNet as the backbone model for feature extraction and ArcFace as the head model to project these features onto a hypersphere for identity retrieval results. Figure 14 illustrates the detailed identification process. The process involves feature extraction by EfficientNet from the cetacean images and subsequent calculation of the cosine similarity of these features by ArcFace. This similarity is then transformed through an arccosine function to determine the angular distance between the feature vector and the ground-truth center vector, integrating a margin penalty. ArcFace is designed to minimize this angular distance for

similar identities while maximizing it for different ones, thereby enhancing the accuracy of identification. The identification results, juxtaposed with the ground truth, are utilized to compute the ArcFace Loss. Subsequently, backpropagation is employed to adjust the weight of EfficientNet, further refining and improving the feature extraction and identification of cetaceans.

Figure 15 illustrates an example of cetacean identification results. The top-left corner depicts the cetacean image input into the system, while the top-middle section displays the cetacean detection results. In the top-right corner, detailed information about the input cetacean, including the image name, species, and ID, is presented. Following this, the top-5 cetacean identification results are provided, sorted by the image similarity score for image retrieval. Correctly identified images are indicated with a green line, while incorrectly identified ones are marked with a red line.

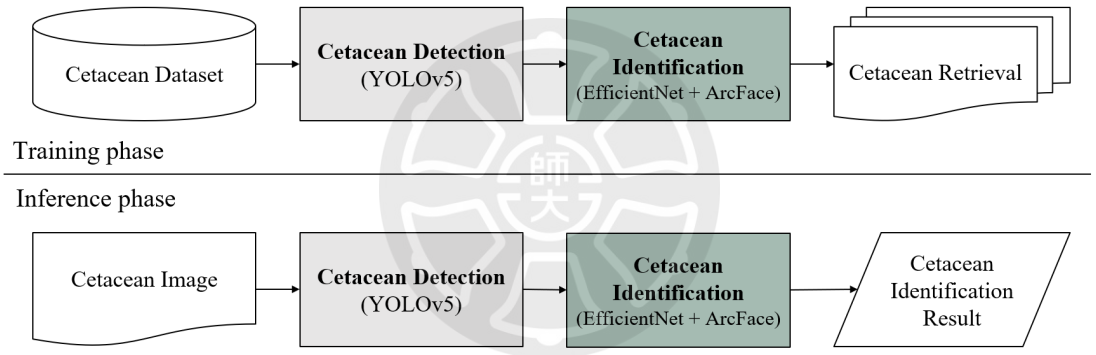


Figure 13. Flowchart of the individual cetacean identification system.

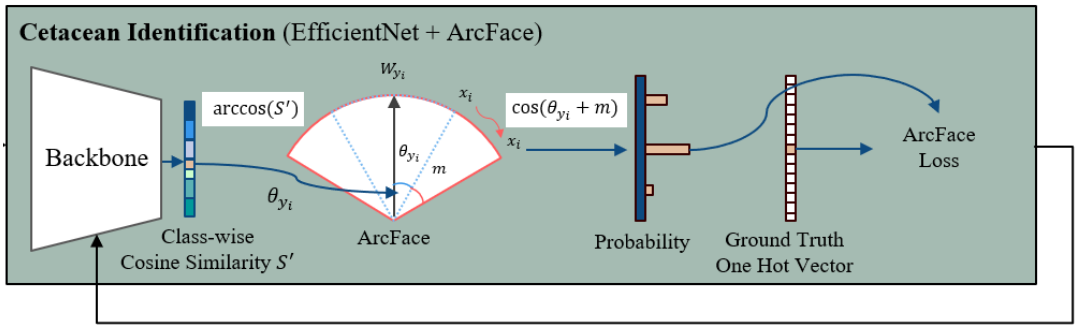


Figure 14. Detailed process of cetacean identification in the training phase.

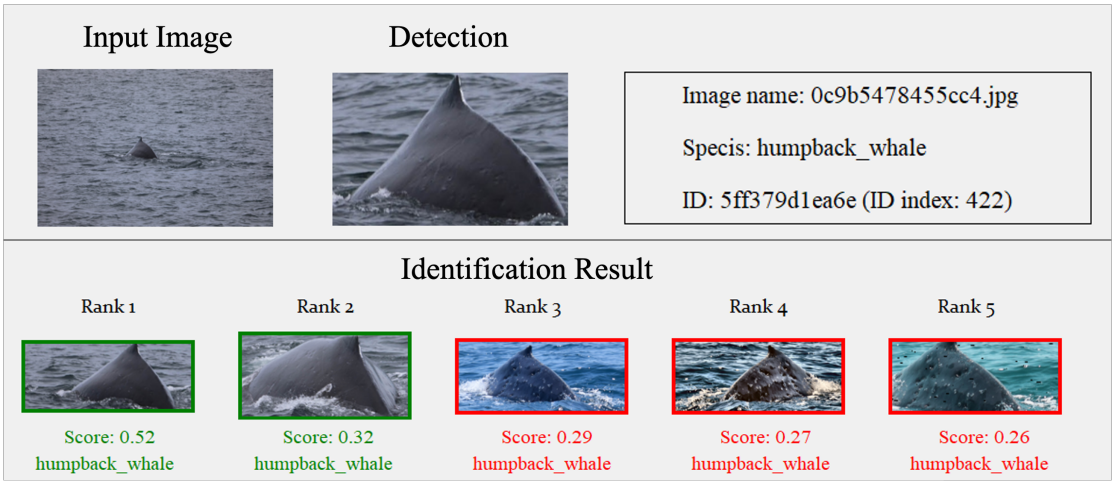


Figure 15. Example for cetacean identification results of humpback whale ID: 5ff379d1ea6e.

3.2 Data Preprocessing

3.2.1 Data Cleaning

As mentioned earlier, images collected in the dataset were categorized as either good or defective. Defective images include those with multiple identities incorrectly labeled as one, as well as images that were overly cluttered, extremely blurry, or had a cetacean area that was extremely small. This study manually identified and cleaned a total of 8,073 defective images, which represents approximately 16% of the total dataset. This ensures the usability of the data for accurate cetacean identification.

3.2.2 Data Augmentation

Owing to the highly uneven distribution of individual cetacean images in the dataset, with approximately 63% of the individuals having only one image, this study implements a data augmentation process for individuals with fewer than 15 images to increase their image count to 15. The data augmentation approach consists of a random combination of three categories from the Python library Albumentations [Bus20], namely noise addition, color transformation, and geometric transformation (refer to Figure 16).

Noise addition involves applying Gaussian blur and Gaussian noise to images, with randomly selected kernel sizes of 3x3, 5x5, or 7x7. Color transformation includes converting images to grayscale, randomly adjusting brightness and contrast, randomly adjusting HSV, applying Contrast Limited Adaptive Histogram Equalization (CLAHE), and performing random RGB shifts. Finally, geometric transformation entails affine transformation by rotating images from −15 to 15 degrees and scaling them from 0.75 to 1.25 times. Subsequently, coarse dropout is applied to rectangular

regions, and the images are cropped and padded by 10% to 20% of the image sizes in pixel amounts and horizontal flipped.

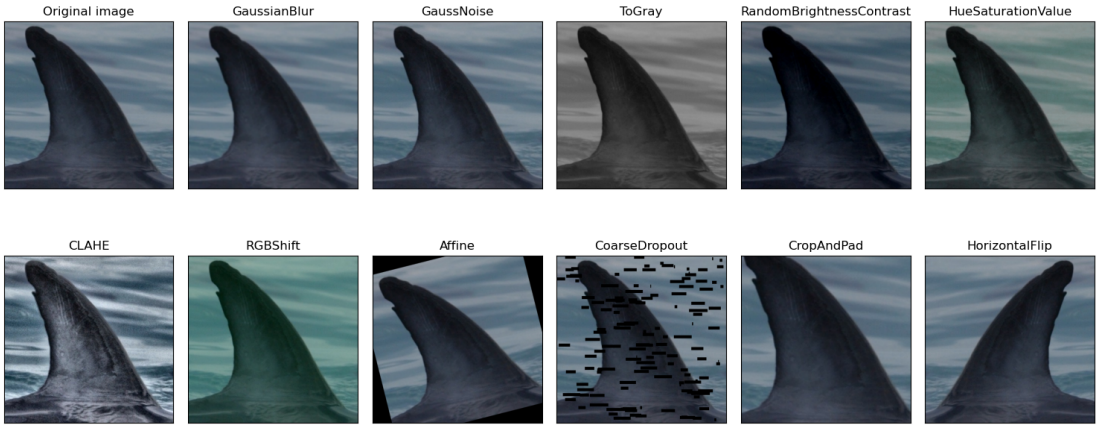


Figure 16. Examples of data augmentation results.

3.3 Examination of Backbone and Head Models

3.3.1 Examination of backbone model

In this study, we first assessed the accuracy of cetacean individual identification using four different backbone models, consistently applying the same preprocessing methods, cetacean detection technique, and head model. The backbone models ConvNextV1, ConvNextV2, EfficientNetV1, and EfficientNetV2 were selected considering that ConvNeXtV1 and EfficientNetV1 were backbone models used by the winners in Kaggle competitions. Meanwhile, ConvNeXtV2 and EfficientNetV2 are enhancements of the two models, both optimizing computational speed while improving accuracy. Experiments show that EfficientNetV1 had a higher accuracy in cetacean individual identification. Further, this study investigated the performance of EfficientNetV1 under varying computational resources and batch sizes. Given the constraints on device computational resources, models ranging from EfficientNetV1-B2 to B5 were selected for experimentation. These models include EfficientNetV1-B5 with a batch size of 8, EfficientNetV1-B4 and B3 both with a batch size of 16, and EfficientNetV1-B2 with a batch size of 32.

Given the variability in shooting angles and distances for cetacean images, we believe that incorporating spatial dimension learning in feature extraction will enhance the accuracy of cetacean individual identification. Therefore, in the EfficientNetV1's main MBConv block, the SE module is replaced with a Convolution Block Attention (CBAM) module [Woo18]. EfficientNetV1 primarily consists of MBConv blocks, each

structured with a 1x1 convolution layer for expanding channel dimensions, followed by a depth-wise convolution layer and an SE module, and concluding with a 1x1 convolution layer that compresses back to the original dimensions. The role of the SE module is to learn the relative importance of the channels in the feature map.

Figure 17 illustrates the CBAM module, which includes two sequential sub-modules, Channel Attention and Spatial Attention modules. The advantage of the CBAM module lies in its ability to learn not only the importance of channels within the feature map but also the significance of spatial dimensions. Moreover, compared to the SE module that uses only average pooling for Channel Attention, the CBAM module employs both average and max pooling for Channel and Spatial Attention (refer to Figure 18).

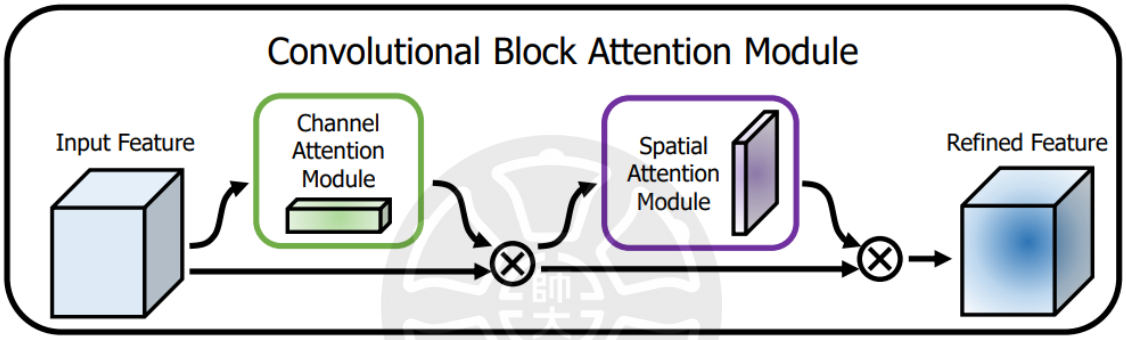


Figure 17. Overview of the CBAM module [Woo18].

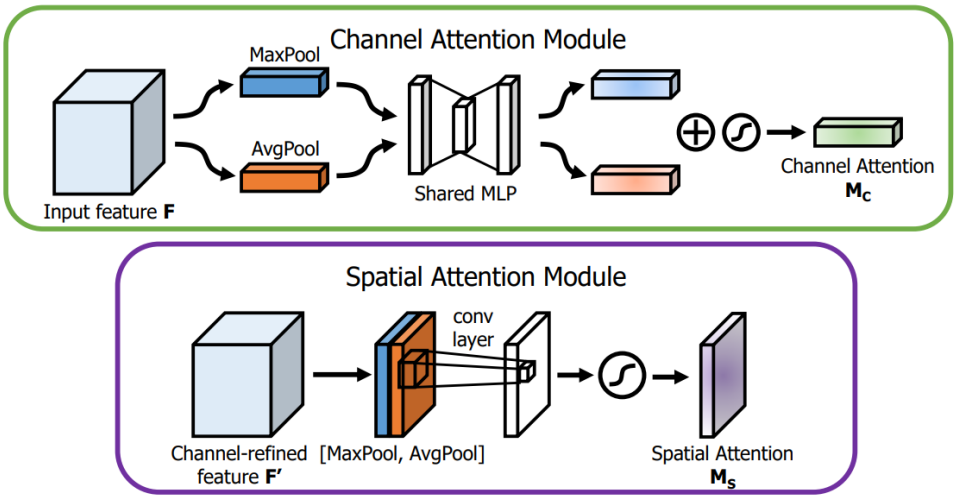
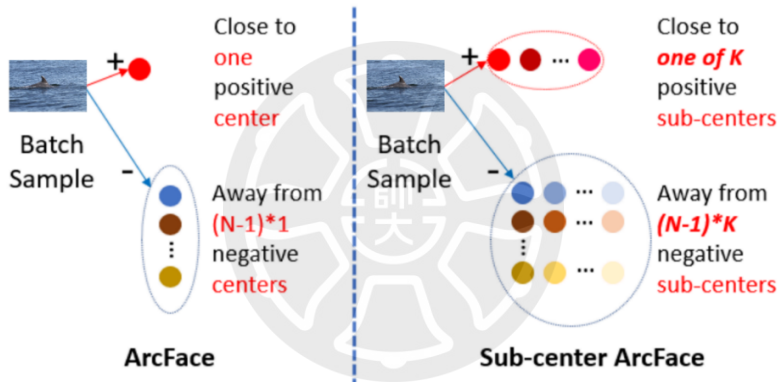


Figure 18. Diagram of channel attention and spatial attention sub-modules [Woo18].

3.3.2 Modification of head model

As mentioned earlier, one of the factors affecting identification is the differences in the image-capturing conditions of the same cetacean. Sub-center ArcFace can be used to recognize cetacean individuals with more than one representative vector. Figure 19 compares ArcFace and sub-center ArcFace. In ArcFace (Figure 19(a)), for each cetacean individual, a central representative vector is selected to represent all the features of that individual's images. This can lead to the problem of a majority of features dominating the identity recognition. To address the multiple representative problem, this study employs sub-center ArcFace (Figure 19(b)), which designs K representative sub-center vectors for each cetacean. During training, being close to any one of sub-center vectors is sufficient to identify the individual. At the same time, the training process also involves distancing from the K sub-center vectors of other cetacean individuals.



(a) (b)

Figure 19. Comparison of ArcFace and sub-center ArcFace. (a) ArcFace; (b) Sub-center ArcFace. Adapted from [Den22].

Figure 20 shows an example of the multiple representative problem. The images in the figure are of the same bottlenose dolphin. One can observe that most images are of the dorsal fin from the side view. This representative vector might neglect the dolphin's frontal features or body characteristics. The images shown in Figure 19 can be suitably represented by four sub-center vectors ($K = 4$), distinguishing between the dominant feature and other sub-features. The dominant feature includes the dorsal fin of a dolphin against a dark background and the gray ocean. The other sub-center vectors include the dolphin's body, the light-colored ocean with a blue background, and a light-

colored dolphin with a green ocean background. This figure illustrates that sub-center ArcFace reasonably solves the multiple representative problem.

Furthermore, to facilitate better convergence in datasets with imbalanced data, this study adopts a dynamic margin approach within sub-center ArcFace. This facilitates learning margin values adaptively based on the varying number of cetacean individuals. Since it is harder to identify distinct features of individuals with fewer images, larger margins are employed for learning. By contrast, for individuals with sufficient images, smaller margins are utilized. Figure 21 illustrates the dynamic margin as a function of the number of individuals, with the horizontal axis representing the number of cetacean individuals and the vertical axis showing the margin values adjusted according to Equation (2.5).

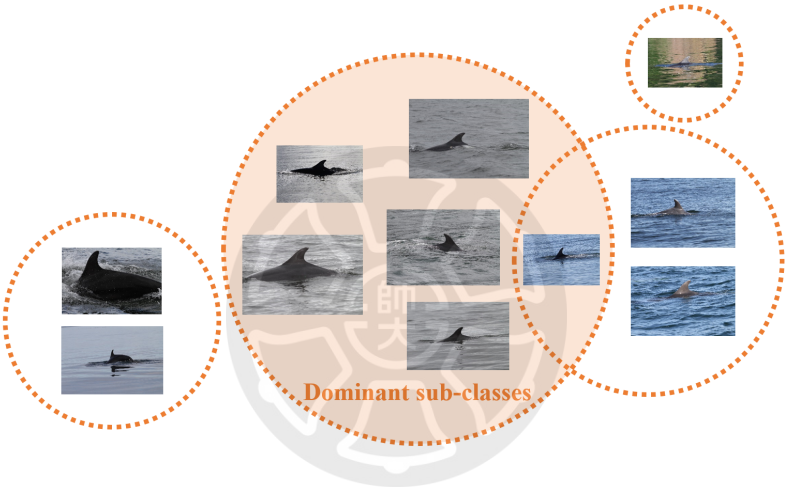


Figure 20. Schematic diagram of one identity of bottlenose dolphin after using the sub-center ArcFace Loss ($K = 4$).

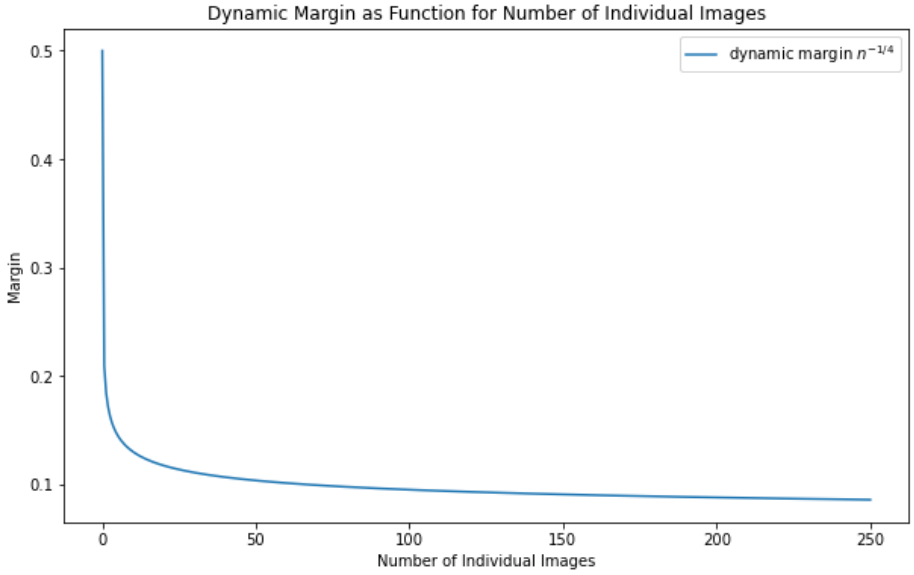


Figure 21. Dynamic margin as a function for the number of individuals.

3.4 Comprehensive Evaluation Methods for Imbalanced Data

This section first introduces the application of ArcFace in cetacean individual identification, a process also known as image retrieval. We focus on the training, validation, and testing phases of ArcFace. Moreover, a comprehensive evaluation method for the dataset is proposed owing to the highly uneven distribution of cetacean individual images in the dataset. This evaluation method involves classifying the number of images for cetacean individuals in the validation and test sets to suit both realistic and ideal evaluation scenarios and conducting cross-testing.

3.4.1 Training, validation, and testing phases of ArcFace

In the training, validation, and testing phases of ArcFace, the robustness of the model in handling images of the same individual with different qualities is considered. In situations where there is an adequate number of individual images, different images of the same cetacean individual are used for (1) training the model and for (2) validation and testing. The validation phase is conducted during the training to assess the model's performance on non-training data. The testing phase is executed after the model training is complete. The objective is to evaluate the model's performance on entirely unseen data, thus providing an assessment of the model's generalization capability.

The image retrieval task divides images into three sets, namely training, query, and gallery sets. The training set is designated for training the model, while the query set and gallery set are employed for model validation and testing, respectively. For each cetacean, five images are randomly selected for validation and testing, with the remaining ten (or more) for training. Specifically, out of the five images designated for validation and testing of each cetacean, one is placed in the query set, and the other four in the gallery set, while the ten (or more) training images are allocated to the training set. This implies that the query set includes images intended for search, encompassing one image of each individual, whereas the gallery set serves as the retrieval image database, containing four images of each individual.

The retrieval process involves taking a query image of a cetacean individual with ID number 1, followed by matching for the top-k (@k) similar cetacean individual images from the gallery set, which comprises four images of each individual. Subsequently, accuracy metrics, namely Mean Average Precision top-5 (mAP@5), Recall top-1 (R@1), Recall top-5 (R@5), and Recall top-10 (R@10), are computed to determine if the top-k ranked images of cetacean individuals match the identity of the

query image. The highest accuracy is achieved when images in the gallery set that share the same identity as the query image are correctly matched and ranked at the top. Different individual identities of cetaceans are used in the validation and testing phases. The classification approach in this study involves sorting cetaceans according to the number of individual images and species, followed by assigning 50% of the data each for validation and testing.

3.4.2 Evaluation methods of imbalanced data

A testing methodology is proposed to effectively utilize the limited quantity of data. As mentioned above, 81% of cetacean individuals have two or fewer images in the dataset. Hence, in situations where there is an extreme imbalance in the number of images per individual identity, an effective testing method is important. This method involves numbering each of the five images of all cetacean individuals in the test data. Then labels the five images of each cetacean individual with numbers 1–5, respectively. Subsequently, grouping the images of all individuals with the same number together, and each group alternating as the query set while the other four groups as the gallery set.

Finally, a test method on these five groups is carried out and averaged to calculate the overall accuracy for cetacean individual identification. To closely simulate real-world application scenarios of the individual cetacean identification system, testing methods for identification system are conducted on cetacean images that are not used in training. Images from group 1 (query set 1) are neither augmented nor included in the model training process, except the individuals with one image.

For the validation and testing, two evaluation methods are designed. Based on the reasons for the uneven distribution of individuals' images and data augmentation up to 15 images in the dataset, each cetacean is categorized into one of four groups based on the number of images: one, two, three to fourteen, and greater than or equal to fifteen, as detailed in Table 1. The first evaluation method considers the actual distribution in the dataset, using all four categories and weighting them according to the number of individuals. The second evaluation method focuses on cetacean individuals with more images, applying the third and fourth categories for weighted evaluation.

Table 1: Number of images for each individual.

Category	#Images per ID	#ID
1	1 image / ID	8,972
2	2 images / ID	2,578
3	3–14 images / ID	2,322
4	15↑images / ID	461

In conclusion, this study categorizes the evaluation methods into two scenarios: real-world application and synthetic data. While some data-augmented images are included in the model training for both scenarios, the evaluation methods are further categorized into three groups: (1) all categories of real-world application scenarios, (2) all categories of synthetic data scenarios, and (3) categories of individual cetaceans with three or more images in synthetic data scenarios.

4 Experimental Results

4.1 Research Environment and Equipment Settings

This study employed a desktop computer equipped with an Intel i5-13600K processor and Nvidia RTX 4090 graphics card. The system operated on the Ubuntu 20.04.05 operating system, with a software environment consisting of Python 3.8, CUDA 11.1, and PyTorch 2.0.1. The models were constructed for training and performance evaluation using the MMPreTrain library in OpenMMLab.

4.2 Cetacean Dataset

This study utilized the cetacean dataset provided by Happywhale for cetacean identification. The dataset included 25 species of cetaceans collected by 28 research organizations with 51,033 labeled training images and 27,956 unlabeled test images. The initial training images were 51,033 images, including a total of 15,587 unique cetacean identities. After the data-cleaning process, the dataset was refined to 42,893 images and 14,333 unique cetacean identities for analysis. Detailed statistics on the image and identity numbers for each whale and dolphin species are listed below in Tables 2 and 3.

In the training and testing process of the ArcFace requirement identification system, for identities with 15 or more images, 1 search (query) image and 4 gallery images were selected for model testing, while the remaining images were used for training. In cases of identities with fewer than 15 images, after applying data augmentation to increase the set to 15 images, 10 were used for training and 5 (including 1 search image and 4 gallery images) were allocated for testing.

Table 2: Dataset distribution of dolphin species.

No.	Species ID	Species of Dolphin	#Image	#ID
1	D1	Dusky Dolphin	2,990	2,627
2	D2	Bottlenose Dolphin	9,057	866
3	D3	Spinner Dolphin	1,315	756
4	D4	Pantropic Spotted Dolphin	574	301
5	D5	White sided Dolphin	172	134
6	D6	Commerson's Dolphin	87	69
7	D7	Rough-Toothed Dolphin	45	37
8	D8	Common Dolphin	310	35
9	D9	Frasier's Dolphin	9	9

Table 3: Dataset distribution of whale species.

No.	Species ID	Species of Whale	#Image	#ID
10	W1	Humpback Whale	6,043	2,545
11	W2	Blue Whale	3,366	1,851
12	W3	Melon Headed Whale	1,564	1,250
13	W4	Beluga	6,951	1,008
14	W5	Southern Right Whale	850	544
15	W6	Fin Whale	1,181	445
16	W7	Killer Whale	2,009	443
17	W8	Short-Finned Pilot Whale	697	435
18	W9	Sei Whale	426	197
19	W10	False Killer Whale	2,632	185
20	W11	Gray Whale	960	175
21	W12	Cuvier's Beaked Whale	197	128
22	W13	Long-Finned Pilot Whale	226	127
23	W14	Minke Whale	1,017	101
24	W15	Bryde's Whale	146	42
25	W16	Pygmy Killer Whale	69	23

4.3 Evaluation for Image Retrieval

The accuracy metrics for cetacean image retrieval are mainly Mean Average Precision (mAP) and Recall as supplementary metrics. mAP represents the mean of the average precisions across all search images in the query set. Let r_c be the number of correctly identified relevant images and r_t be the number of total retrieved images. Precision is defined as the ratio of correctly identified relevant images to the total retrieved images, as given by Equation (4.1).

$$Precision = \frac{|\{r_c\}|}{|\{r_t\}|}, \tag{4.1}$$

where r_c is the number of correctly identified relevant images and r_l is the number of total relevant images. Recall measures the proportion of relevant images that are correctly identified, as given by Equation (4.2).

$$Recall = \frac{|\{r_c\}|}{|\{r_l\}|}. \tag{4.2}$$

The study evaluated cetacean identification results using key accuracy metrics, namely mAP@5, Recall@1, Recall@5, and Recall@10. In the below formulas, symbol N donates the total number of search images, with i representing the i-th search image, ranging from 1 to N. The Mean Average Precision for the top-5 retrieval results ($mAP@5$) is formulated as Equation (4.3).

$$mAP@5 = \frac{1}{N}\sum_{i=1}^{N} AP_i = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{RD_i}\sum_{k=1}^{5} P_i(k) \times rel_i(k). \tag{4.3}$$

For the top-5 retrieval results of the i-th search image, the average precision AP_i can be calculated by parameters RD_i, $P_i(k)$, and $rel_i(k)$. Here RD_i is the number of relevant images in the gallery set. $P_i(k)$ is the precision at k, which can be calculated using Equation (4.1). $rel_i(k)$ is the relevance of the k-th retrieval image, where relevance is defined as 1 if the r-th retrieval image is the same identity as the search image, and 0 otherwise.

Recall for the top-k retrieval results ($Recall@k$) is given by Equation (4.4).

$$Recall@k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{RD_i}\sum_{r=1}^{k}R_i(r) \times rel_i(r)\,. \tag{4.4}$$

For the top-5 retrieval results of the i-th search image, Recall can be calculated using parameters RD_i, $R_i(r)$, and $rel_i(r)$. Here RD_i is the number of relevant images in the gallery set. $R_i(r)$ is the recall at r, which is calculated using Equation (4.2). $rel_i(r)$ is the relevance of the r-th retrieval image.

4.4 Backbone Examination Analysis

Based on the examinations of backbone models within ConvNeXt series and EfficientNet series, EfficientNetV1 exhibited the best performance. Then, backbone models with different number of FLOPS, EfficientNetV1-B2 to B5, were tested with different batch sizes. Finally, EfficientNetV1-B4 with a batch size of 16 was utilized as the backbone model for the identification system.

This study assumed that in the feature extraction phase of the cetacean individual identification model, extracting features from more aspects leads to better identification results. Therefore, the SE module in EfficientNet was replaced with the CBAM module, thereby adding a spatial attention mechanism to the original channel attention module in the backbone model. Tables 4, 5, and 6 show the comparison of two backbone models for cetacean individual identification through the three evaluation methods; one model used the EfficientNet-b4 model with ArcFace (baseline setting, and in the other, the SE module in EfficientNet-b4 was replaced with the CBAM module.

Across the three evaluation methods, the baseline setting yielded mAP@5 scores of 62.44%, 75.52%, and 27.51%. By contrast, EfficientNet-b4 integrated with the CBAM module yielded mAP@5 scores of 62.44%, 73.96%, and 17.00%. For the baseline setting, the training time was 24.69 h, whereas the training time for EfficientNet-b4 with the CBAM module extended to 47.19 h. These results indicate that substituting the SE module with the CBAM module did not enhance performance; instead, it prolonged the model's training duration. This increased training time can be

attributed to the complexity introduced by excessive spatial information in the images, such as waves, which could complicate the feature learning process and consequently reduce identification accuracy. Figure 22 shows the cetacean identification results in real-world application. Replacing the SE module with the CBAM module in EfficientNet led to the top five results more closely matching the shooting angles and environments of the input images.

Table 4: Comparison of the SE and CBAM modules in EfficientNet-b4 for Evaluation (1).

Real-world application							
No.	Model Settings	Best Epoch	Training Time (hr.)	mAP @5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace	41	24.69	**63.80**	**67.44**	**70.89**	**72.95**
2	EfficientNet – b4 (CBAM) + ArcFace	57	47.19	62.44	65.87	69.58	71.80

Table 5: Comparison of the SE and CBAM modules in EfficientNet-b4 for Evaluation (2).

Synthetic data for all individuals					
No.	Model Settings	mAP@5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace	**75.52**	**84.95**	87.12	88.33
2	EfficientNet – b4 (CBAM) + ArcFace	73.96	84.65	**87.22**	**88.63**

Table 6: Comparison of the SE and CBAM modules in EfficientNet-b4 for Evaluation (3).

Synthetic data for #image≥3 individuals					
No.	Model Settings	mAP@5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace	**27.51**	**50.12**	**59.89**	**65.07**
2	EfficientNet – b4 (CBAM) + ArcFace	17.00	41.49	51.14	57.18

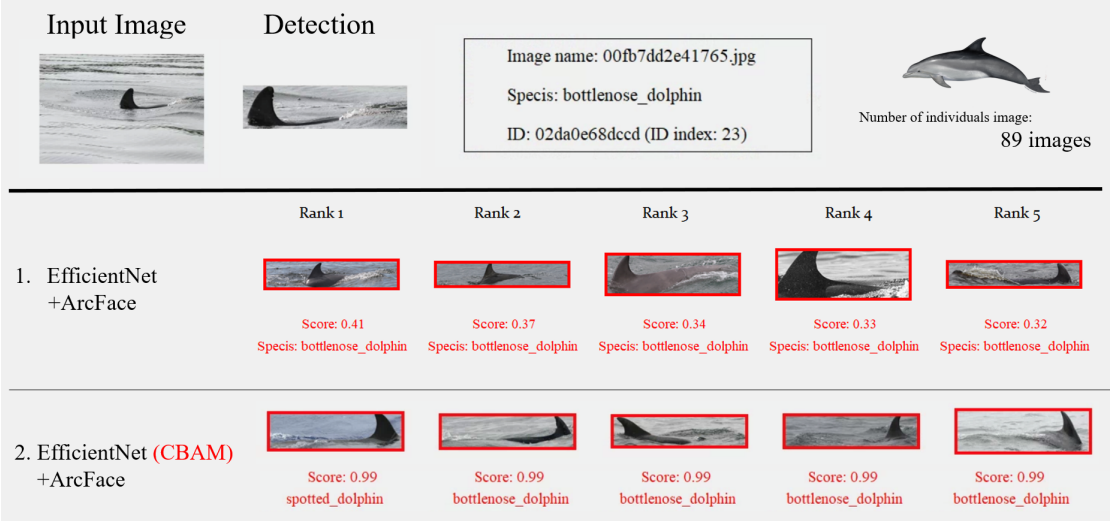


Figure 22. Comparison of SE and CBAM modules in EfficientNet-b4 for cetacean identification results of bottlenose dolphin ID: 02da0e68dccd in Evaluation (1).

4.5 Head Improvement Analysis

This study incorporates sub-centers in ArcFace, providing additional k centers for each cetacean to capture the subtle differences among individuals more efficiently. Experiments to improve the accuracy of cetacean individual identification were conducted, with the number of sub-centers set from 1 to 3. Tables 7, 8, and 9 show that setting the sub-center to 2 and 3 enhanced the accuracy of cetacean individual identification compared to the baseline. The sub-center set to 2 yielded the most significant improvements. Specifically, for the baseline setting with sub-center set to 1, the mAP@5 scores were 63.80%, 75.52%, and 27.51%. When the sub-center was set to 2, the mAP@5 scores increased to 68.63%, 81.60%, and 35.70%. However, with the sub-center set to 3, the mAP@5 scores were 65.92%, 78.42%, and 21.91%, indicating a varied impact on accuracy.

Logically, setting sub-center to 3 for individuals with a larger number of images should facilitate learning cetacean features across more varied image-capturing conditions, potentially boosting accuracy. However, in the lower right synthetic data, where individual image counts are more than three, setting sub-center to 3 actually decreased accuracy compared to the baseline. This decrease could be attributed to the uneven distribution of individual images in the dataset, where most cetaceans have fewer than two images, making it more challenging to learn features specific to a minority of individuals. Figure 23 shows the results of cetacean detection and individual identification for a humpback whale with 16 images, displaying the top five outcomes. Setting the sub-center number from 1 to 3 all successfully identified the humpback

whale. Notably, with sub-centers set at 1 to 3, the system recognized the same humpback whale captured from similar shooting angles. Particularly, at sub-center = 2, the system identified the same humpback whale even from different shooting angles. The experimental data and resulting images suggest that for this dataset, setting sub-center to 2 offers better identification performance for the same cetacean under different image-capturing conditions.

Table 7: Comparison of different sub-centers of ArcFace for Evaluation (1).

Real-world application							
No.	Model Settings	Best Epoch	Training Time (hr.)	mAP @5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace (sub-center = 1)	41	24.69	63.80	67.44	70.89	72.95
2	EfficientNet – b4 + ArcFace (sub-center = 2)	56	41.58	**68.63** (+4.83)	**71.94** (+4.5)	**76.29** (+5.4)	**78.58** (+5.63)
3	EfficientNet – b4 + ArcFace (sub-center = 3)	48	36.25	65.92	68.06	71.41	73.78

Table 8: Comparison of different sub-centers of ArcFace for Evaluation (2).

Synthetic data for all individuals					
No.	Model Settings	mAP@5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace (sub-center = 1)	75.52	84.95	87.12	88.33
2	EfficientNet – b4 + ArcFace (sub-center = 2)	**81.60** (+6.08)	**89.78**	**92.16**	**93.17**
3	EfficientNet – b4 + ArcFace (sub-center = 3)	78.42	86.81	88.97	90.27

Table 9: Comparison of different sub-centers of ArcFace for Evaluation (3).

Synthetic data for #image≥3 individuals					
No.	Model Settings	mAP@5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace (sub-center = 1)	27.51	50.12	59.89	65.07
2	EfficientNet – b4 + ArcFace (sub-center = 2)	**35.70** (+8.19)	**59.88**	**71.50**	**75.84**
3	EfficientNet – b4 + ArcFace (sub-center = 3)	21.91	46.17	56.34	62.82

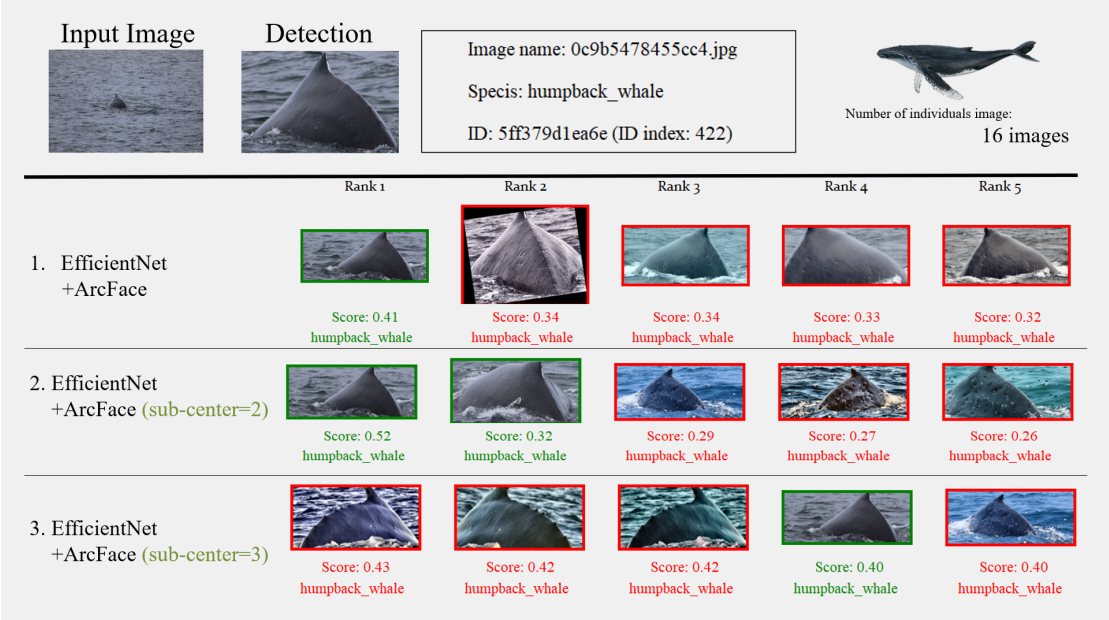


Figure 23. Comparison of different sub-centers of ArcFace for cetacean identification results of humpback whale ID: 5ff379d1ea6e in Evaluation (1).

To handle imbalanced dataset, sub-center ArcFace with dynamic margins can adaptively learn margin based on the number of identities. This approach aims to optimize the learning process, particularly for those cetaceans represented by fewer images, by effectively balancing the margin size. Larger margins are used to differentiate identities with fewer images, as their characteristics are harder to learn. The margin size is adjusted for each cetacean based on the corresponding number of images, following Equation (2.5). In the baseline setting, ArcFace uses a fixed margin value of 0.5. In this study, the parameters of Equation (2.5) are set with an upper bound of 0.5, a lower bound of 0.05, and λ set to 1/4.

Tables 10, 11, and 12 present the results of three evaluations for sub-center ArcFace, both with and without dynamic margins. The training times for sub-center ArcFace with dynamic margin and without it were 41.58 and 29.78 h, respectively. Notably, the mAP@5 scores for sub-center ArcFace with dynamic margins were 63.84%, 80.78%, and 29.44%. In comparison, without dynamic margin, the mAP@5 scores were 68.63%, 81.60%, and 35.70%. Sub-center ArcFace with dynamic margins maintained comparable accuracy to the version without dynamic margins, while reducing training time by 28%. Figure 24 illustrates the cetacean identification results for a short-finned pilot whale with three images, highlighting the top five outcomes. The results demonstrate that sub-center ArcFace, both with and without dynamic

margins, successfully identified the correct individual. Notably, the version with dynamic margins showed slightly higher confidence scores and accuracy compared to the one without dynamic margins.

Table 10: Comparison of sub-center ArcFace with and without dynamic margins in Evaluation (1).

Real-world application							
No.	Model Settings	Best Epoch	Training Time (hr.)	mAP @5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace (sub-center = 2)	56	41.58	**68.63**	**71.94**	**76.29**	**78.58**
2	EfficientNet – b4 + ArcFace (sub-center = 2, **dynamic margin**)	**50**	**29.78** (-28%)	68.34	71.30	76.04	78.55

Table 11: Comparison of sub-center ArcFace with and without dynamic margins in Evaluation (2).

Synthetic data for all individuals					
No.	Model Settings	mAP@5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace (sub-center = 2)	**81.60**	**89.78**	**92.16**	**93.17**
2	EfficientNet – b4 + ArcFace (sub-center = 2, **dynamic margin**)	80.78	89.04	91.46	92.64

Table 12: Comparison of sub-center ArcFace with and without dynamic margins in Evaluation (3).

Synthetic data for #image≥3 individuals					
No.	Model Settings	mAP@5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace (sub-center = 2)	**35.70**	**59.88**	**71.50**	**75.84**
2	EfficientNet – b4 + ArcFace (sub-center = 2, **dynamic margin**)	29.44	52.54	65.47	71.77

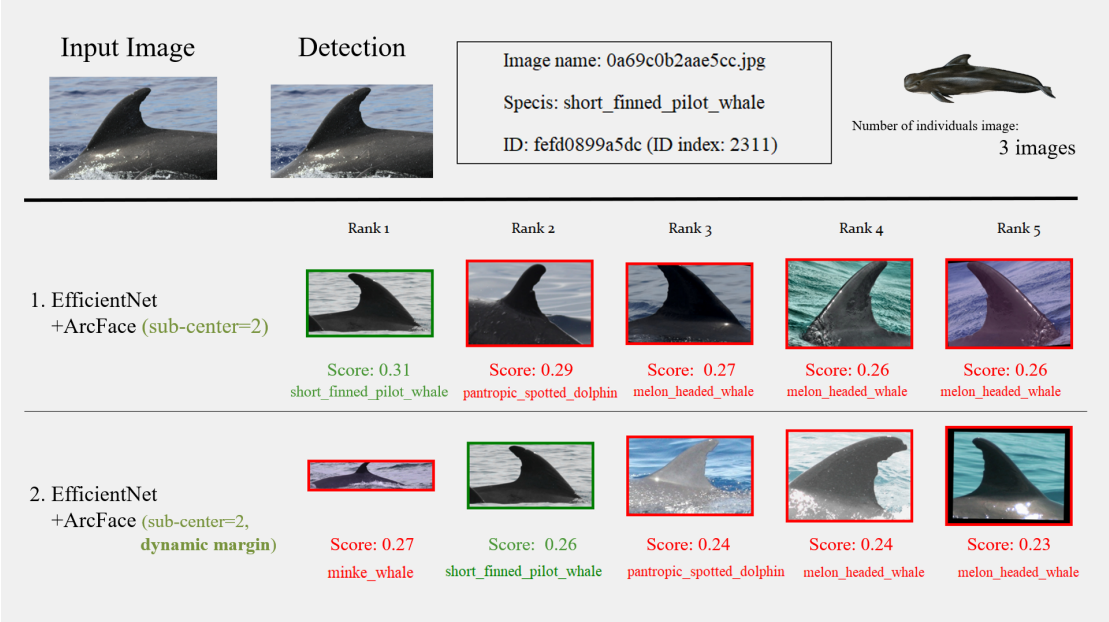


Figure 24. Comparison of sub-center ArcFace with and without dynamic margins for cetacean identification results of short-finned pilot whale ID: fefd0899a5dc in Evaluation (1).

Table 13 shows the ablation results to analyze the effect of model components on the task of cetacean individual identification. The baseline setting achieved mAP@5 accuracy of 63.80% and Recall@1, Recall@5, and Recall@10 scores of 67.44%, 70.89%, and 72.95%, respectively. The experiments demonstrate that adjusting the number of sub-centers to 2 significantly improved the model's mAP@5 accuracy to 68.63%. This adjustment, termed improvement version 1, also resulted in Recall@1, Recall@5, and Recall@10 scores of 71.94%, 76.29%, and 78.58%, respectively. Furthermore, the incorporation of dynamic margins in improvement version 2 not only preserved this accuracy level but also enhanced training efficiency by 28% compared to improvement version 1.

Table 13: Ablation results of identification system in Evaluation (1).

Real-world application									
No.	Model Settings	**Head** Sub-center	**Head** Dynamic margin	Best Epoch	Training Time (hr.)	mAP @5 (%)	R@1 (%)	R@5 (%)	R@10 (%)
1	EfficientNet – b4 + ArcFace			41	24.69	63.80	67.44	70.89	72.95
2	EfficientNet – b4 + ArcFace v1	✓		56	41.58	**68.63** (+4.83)	**71.94** (+4.5)	**76.29** (+5.4)	**78.58** (+5.63)
3	EfficientNet – b4 + ArcFace v2	✓	✓	**50**	**29.78** (-28%)	68.34	71.30	76.04	78.55

5 Conclusion and Future Works

5.1 Conclusion

This study established a cetacean individual identification system and addressed potential issues related to unstable image quality and imbalanced distribution of individual images in the dataset. The dataset was preprocessed, involving data cleaning, data augmentation, and then the cetacean detection task was performed. Then, for cetacean identification, EfficientNetV1-B4 was used as the backbone model to extract features from cetacean images, and ArcFace was employed as the head model for individual identity prediction. This study improved the head model ArcFace to address potential issues in the dataset, thereby enhancing the accuracy of cetacean individual identification. By incorporating sub-center vectors into ArcFace, the model effectively addressed the issue of the effect of different image-capturing conditions for the same cetacean, improving the accuracy of individual identification. In addition, the introduction of dynamic margins for sub-center ArcFace addressed the problem of highly uneven distribution of cetacean individual images during training, increasing the convergence speed of the model. Three evaluation methods were proposed to analyze the dataset, namely real-world application, majority synthetic dataset, and partial synthetic dataset (individuals with more than three images).

The experimental results showed that EfficientNet as the backbone model combined with sub-center ArcFace (sub-center $= 2$) as the head model achieved the highest accuracy across all three evaluation methods, with mAP scores of 68.63%, 81.60%, and 35.70%, respectively. Compared to the baseline EfficientNet and ArcFace, there was an increase in mAP by 4.83%, 6.08%, and 8.19%, respectively. The improved EfficientNet and sub-center ArcFace with dynamic margins maintained a high accuracy comparable to previous improvement, while reducing the training time by 28%.

5.2 Future Works

This study proposed a system for individual cetacean identification. Evaluation of its performance indicates room for improvement in this area. The future work is listed below in three goals: short-term, medium-term, and long-term.

The short-term goal will focus on enhancing the accuracy of cetacean detection and expanding the dataset to better identify individuals. The work will include improving the detection system to reduce detection errors in complex backgrounds and adding a function in the detection system to recognize cetacean body parts. In addition, expanding the dataset will be crucial to address the current issue of uneven distribution of individual identities. We will also improve the overall accuracy of the identification system.

For the medium-term goal, the plan is to integrate additional cetacean data such as video, voice, and location into the identification process, providing a more comprehensive method to identify individual cetaceans. Using different types of sensors complementing each other's functions can improve the recognition rate of cetacean individuals.

The long-term goal is to apply cetacean identification technology in developing a platform for cetacean route tracking. This platform aims to achieve marine life conservation by tracking cetacean movements, thereby supporting marine life conservation.

With these improvements, the study aims to continually enhance the cetacean identification system’s efficiency and contribute significantly to the conservation of marine ecosystems.

References

[1] B.W. Eakins and G.F. Sharman, “Volumes of the World's Oceans from ETOPO1,” *NOAA National Geophysical Data Center*, Boulder, CO, 2010.

[2] B.S. Halpern, C. Longo, D. Hardy, K.L. McLeod, J.F. Samhouri, S.K. Katona, K. Kleisner, S.E. Lester, J. O’Leary, M. Ranelletti, A.A. Rosenberg, C. Scarborough, E.R. Selig, B.D. Best, D.R. Brumbaugh, F.S. Chapin, L.B. Crowder, K.L. Daly, S.C. Doney, C. Elfes, M.J. Fogarty, S.D. Gaines, K.I. Jacobsen, L.B. Karrer, H.M. Leslie, E. Neeley, D. Pauly, S. Polasky, B. Ris, K.S. Martin, G.S. Stone, U.R. Sumaila, and D. Zeller, “An index to assess the health and benefits of the global ocean,” *Nature*, vol. 488, pp. 615-620, Aug. 2012.

[3] J. Roman, J. Estes, L. Morissette, C. Smith, D. Costa, J. McCarthy, J.B. Nation, S. Nicol, A. Pershing, and V. Smetacek, “Whales as Marine Ecosystem Engineers,” *Frontiers in Ecology and the Environment*, vol. 12, no. 7, pp. 377-85, Jul. 2014.

[4] R. Chami, T. Cosimano, C. Fullenkamp, S. Oztosun, R. Chami, T. Cosimano, C. Fullenkamp, and S. Oztosun, “Nature’s Solution to Climate Change,” *Finance and Development*, vol. 56, pp. 34-38, Dec. 2019.

[5] T. Cheeseman, K. Southerland, W. Reade, and A. Howard, “Happywhale - Whale and Dolphin Identification,” *Kaggle website* (2022). Available: https://kaggle.com/competitions/happy-whale-and-dolphin (Feb. 7, 2023).

[6] M.A. Rahman, “Happywhale: BoundingBox [YOLOv5],” *Kaggle website* (2022). Available: https://www.kaggle.com/code/awsaf49/happywhale-boundingbox-yolov5/notebook (Feb. 7, 2023).

[Pat23] P.T. Patton, T. Cheeseman, K. Abe, T. Yamaguchi, W. Reade, K. Southerland, A. Howard, E.M. Oleson, J.B. Allen, E. Ashe, A. Athayde, R.W. Baird, C. Basran, E. Cabrera, J. Calambokidis, J. Cardoso, E.L. Carroll, A. Cesario, B.J. Cheney, E. Corsi, J. Currie, J.W. Durban, E.A. Falcone, H. Fearnbach, K. Flynn, T. Franklin, W. Franklin, B. G. Vernazzani, T. Genov, M. Hill, D.R. Johnston, E.L. Keene, S.D. Mahaffy, T.L. McGuire, L. McPherson, C. Meyer, R. Michaud, A. Miliou, D.N. Orbach, H.C. Pearson, M.H. Rasmussen, W.J. Rayment, C. Rinaldi, R.

Rinaldi, S. Siciliano, S. Stack, B. Tintore, L.G. Torres, J.R. Towers, C. Trotter, R.T. Moore, C.R. Weir, R. Wellard, R. Wells, K.M. Yano, J.R. Zaeschmar, and L. Bejder, "A deep learning approach to photo–identification demonstrates high performance on two dozen cetacean species," *Methods in Ecology and Evolution*, vol. 14, pp. 2611-2625, Oct. 2023.

[Red16] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Las Vegas, Nevada, 2016, pp. 779-788.

[Red17] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Hawaii, Honolulu, 2017, pp. 6517-6525.

[Lof15] S. Loffe and C. Szegedy, (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint arXiv: 1502.03167.*

[Ren15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Proceedings of Advances in neural information processing systems 28* (NIPS), Montréal, Canada, 2015.

[Red18] J. Redmon and A. Farhadi, (2018). "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*.

[Lin17] T.Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Hawaii, Honolulu, 2017, pp. 2117-2125.

[He16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Las Vegas, Nevada, 2016, pp. 770-778.

[Boc20] A. Bochkovskiy, C. Wang, and H.M. Liao, (2020). "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv: 2004.10934.*

[Wan20] C.Y. Wang, H.Y. Mark Liao, Y.H. Wu, P.Y. Chen, J.W. Hsieh, and I.H. Yeh, "CSPNet: A New Backbone that can Enhance Learning Capability of

CNN," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW), 2020, Seattle, WA, USA, pp. 1571-1580.

[He15] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), vol. 37, no. 9, pp. 1904-1916, Sept. 2015.

[Liu18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia; "Path Aggregation Network for Instance Segmentation," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Salt Lake City, UT, USA, 2018, pp. 8759-8768.

[Joc20] G. Jocher, "YOLOv5 in Pytorch," Github (2020). Available: https://github.com/ultralytics/yolov5 (Oct. 23, 2023).

[Hu18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2018, Salt Lake City, UT, USA, pp. 7132-7141.

[Hua17] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, "Densely connected convolutional networks," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Hawaii, Honolulu, 2017, pp. 4700-4708.

[Liu21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin., and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of IEEE/CVF International Conference on Computer Vision* (ICCV), Montreal, BC, Canada, 2021, pp. 10012-10022.

[Liu22] Z. Liu, H. Mao, C. -Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), New Orleans, LA, USA, 2022, pp. 11966-11976.

[Woo23] S. Woo, S. Debnath, R. Hu, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Vancouver, BC, Canada, 2023, pp. 16133-16142.

[Tan19] M. Tan and Q.V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of 36th International*

Conference on Machine Learning (PMLR), Long Beach, California, USA, 2019, vol. 97, pp. 6105-6114.

[Cho17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Hawaii, Honolulu, 2017, pp. 1251-1258.

[Tan21] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," *Proceedings of International Conference on Machine Learning* (PMLR), online, 2021, pp. 10096-10106.

[Den22] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI), vol. 44, no. 10, pp. 5962-5979, Oct. 2022.

[Ha20] Q. Ha, B. Liu, F. Liu, and P. Liao, (2020). "Google Landmark Recognition 2020 Competition Third Place Solution," *arXiv preprint arXiv: 2010.05350*.

[Hua20] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Seattle, Washington virtual/online, 2020, pp. 5901-5910.

[Bus20] A. Buslaev, V.I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A.A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2: 125, Feb. 2020.

[Woo18] S. Woo, J. Park, J.Y. Lee, and I.S. Kweon, "Cbam: Convolutional block attention module," *Proceedings of the European conference on computer vision* (ECCV), Munich, Germany, 2018, pp. 3-19.

Appendix

A. Chinese and English whale species table:

- Monodontidae (一角鯨科)
- Ziphiidae (喙鯨科)
- Balaenidae (露脊鯨科)
- Eschrichtiidae (灰鯨科)
- Delphinidae (海豚科)
- Balaenopteridae (鬚鯨科)

Number ID	Species ID	Chinese	Species of Whale
1	W4	白鯨	Beluga
2	W5	南露脊鯨	Southern Right Whale
3	W12	科氏喙鯨	Cuvier's Beaked Whale
4	W11	灰鯨	Gray Whale
5	W3	瓜頭鯨	Melon Headed Whale
6	W8	虎鯨(殺人鯨)	Killer Whale
7	W10	偽虎鯨	False Killer Whale
8	W16	小虎鯨	Pygmy Killer Whale
9	W7	短肢領航鯨	Short-Finned Pilot Whale
10	W14	長肢領航鯨	Long-Finned Pilot Whale
11	W1	大翅鯨(座頭鯨)	Humpback Whale
12	W2	藍鯨	Blue Whale
13	W6	長須鯨	Fin Whale
14	W9	塞鯨(北鬚鯨)	Sei Whale
15	W13	小鬚鯨	Minke Whale
16	W15	布氏鯨	Bryde's Whale

B. Chinese and English dolphin species table:

Number ID	Species ID	Chinese	Species of Dolphin
1	D1	瓶鼻海豚	Bottlenose Dolphin
2	D2	暗色斑紋海豚	Dusky Dolphin
3	D3	長吻飛旋海豚	Spinner Dolphin
4	D4	熱帶點斑海豚	Pantropic Spotted Dolphin
5	D5	真海豚	Common Dolphin
6	D6	斑紋海豚	White Sided Dolphin
7	D7	康式矮海豚	Commerson’s Dolphin
8	D8	糙齒海豚	Rough-Toothed Dolphin
9	D9	弗氏海豚	Frasier’s Dolphin