Whale and Dolphin Classification and Photo Identification

Mauricio Salim Gómez Chicre
Universidad de Los Andes
Cra. 1 18a-12, Bogotá, Cundinamarca
ms.gomezc@uniandes.edu.co

Sebastian Ojeda Alonso
Universidad de Los Andes
Cra. 1 18a-12, Bogotá, Cundinamarca
s.ojedaa@uniandes.edu.co

Abstract

Photo-ID facilitates marine mammal life investigations and population assessment as researchers manually identify individuals through shape and markings on tails, dorsal fins, heads and other body parts of individuals, bringing in the possibility of knowing species population status and trends over time. It is the most common and popular way to recognize individual cetaceans since its non-invasive, it plays an important role in the field of conservation science, but is very time consuming. Researchers need to manually compare photo through photo, and as the era of digital photography exponentially increases the volume of photos submitted to catalogs around the world, this task becomes more unaffordable and impossible to do by manual human eye matching. That is why in this article we a develop a cetacean classification method that obtains an F1-score of 86% and 89% in both tested folds, and experiments a first approach to the photo identification problem with a method that obtains a Top-5 Accuracy of 23%.

1. Introduction

To conduct marine mammal life investigations and population assessments, researchers manually identify individuals, this activity, known as Photo-ID, allows to know the species population status and trends over time. Photo-ID is the most common and popular way to recognize individual cetaceans since its non-invasive (See Figure 1) [8]. This assessment method plays an important role in the field of conservation science, as managing the recovery of endangered species relies on estimating population abundance and monitoring trends over time [22]. Photo-ID recognizes the shape and markings on their tails, dorsal fins, heads and other body parts. Being able to detect the increasing or decreasing abundance of endangered species such as the Sei Whale or Blue Whale [9], can improve and inform the effectiveness of species conservation.

But why is it important to study and monitor them? Well, there are several reasons that speak up for cetacean's importance in our environment. First, they are indicators for the state of marine ecosystems and drivers of economic activities related to tourism, such as the whale-watching industry worth over a billion dollars a year. Not only that but, these economic activities are the main income and source of employment of several coastal communities. Additionally, they contribute to their habitats productivity and richness by modifying it, as a result, cetaceans are a crucial species to the conservation science domain as they improve the health and stability of the environment they are living in. On the other hand, they soften the effect of fossil fuels to the atmosphere by reducing the amount of carbon dioxide (CO2) in their environment, and they are great subjects of study as they live in complex social societies and have incredible cognitive skills. Finally, cetaceans require special international protection given their nature, every year they do long migrations between the poles and the tropics. It is so that even though they provide a wide range of benefits to their surroundings, from the 88 species known to science, the International Union for the Conservation of Nature (IUCN) classifies three as Critically Endangered, seven as Endangered, six as Vulnerable, and five as Near Threatened.[1] [7]

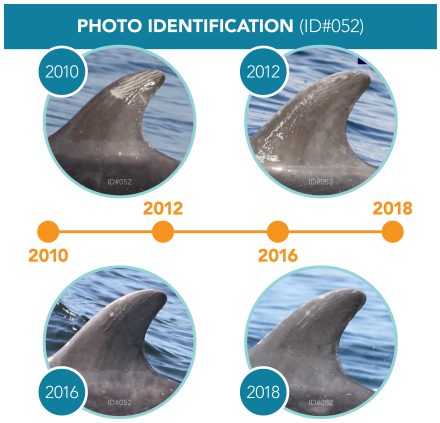


Figure 1. Photo-ID of an individual cetacean throughout the years. [11].

However, photo identification is very time consuming. Researchers need to manually compare photo through photo, and as the era of digital photography exponentially increases the volume of photos submitted to catalogs around the world, this task becomes more unaffordable and impossible to do by manual human eye matching in scales were it can have a bigger impact. This is where artificial intelligence comes into play, as recent advances in deep learning in particular has directed the way to automated image processing by the use of neural networks modeled on the human brain. Through the correct use of this technology, we have the potential to revolutionize the matching speed of these images to identify known individuals. If successful, researchers can reduce image identification times by over 99%.[22]

In the artificial intelligence context this type of problem can be catalogued into 4 main phases. The first one, is to identify the cetacean in the image (e.g. through bounding boxes), the second one, is to differentiate between whales and dolphins, the third one to classify the species of the cetacean, and finally, after knowing its species we can identify the individual. More specifically this type of problem is somewhat similar to those of face recognition [24] since some of the steps normally used in these, are also used in previous applications in the area of Photo-ID, *Applying deep learning to right whale photo identification* [22].

Taking everything previously mentioned into account, we propose to approach this problem firstly through the use of Meta's object detection method named DETIC, specifically through the already implemented DETIC method to the Happywhale dataset by Phalanx [20]. In this way, we can focus directly on the classification problem of the 30 cetaceans species using Convolutional Neural Networks (CNN) such as ResNET and EffNET. For the division of the database, we chose to use the 2-fold cross validation technique since the number of images is considerably reduced with respect to the original database. As the project and the metrics obtained improve, we will tackle the Photo-ID problem, firstly using the state of the art methods, and then, using parallel ideas from the face recognition problems we will be iterating the model, as Photo-ID and face recognition are somewhat similar challenges.

2. Related Work

The automatization of the photo identification task through predictive models such as CNN has also been approached with giraffes [26] (Giraffa camelopardalis) and the Australian skinks [5]. In [5] they use the Interactive Individual Identification System (I3S Pattern) for the identification of individual skinks, their database consists of 1153 images of 30 different adult lizards. They evaluate their dataset with their I3S Pattern computer algorithm [14], but also with 12 observers that are professional field biologist who specialize in plant or animal surveys, although none have specific experience with the study species, and the other 12 with no experience whatsoever. Experienced observers identified a higher proportion of photos correctly (74%) than those with no experience (63%) while the I3S software correctly matched 67% as the first ranked match and 83% of images in the top five ranks [14]. On the other hand, in [26] they first develop an algorithm to extract giraffes from a total of 4000 photos through combining CNN-based object detection, SIFT pattern matching, and image similarity networks obtaining a total of 178 individuals and 5,019 images with a single giraffe each. Then they quantify the performance of deep metric learning and retrieve the identity of known and unknown individuals. With this process the performance of their CNNs reach a top 5 accuracy of about 90% [26].

Author	Method	Accuracy
Andrej Marichenko	Happywhale [0.679]	0.679
Andrij	Happywhale - Effnet B7 fork with Detic Training	0.699
Rabbit	0.720_&EFF_B5_640_Rotate	0.720
Andrij	happywhale_arcface_baseline_eff7_tpu_768_inference	0.729
Isamu	simple ensemble of public best kernels	0.750

Table 1. State of the art in HappyWhales dataset.

Parallel to the Right Whale challenge, is the HappyWhale challenge [11] were submissions mainly use Detic [28] or YOLOV5 [25] method to solve the detection problem by cropping the cetacean from the images, some use a combination of both Yolov5 and Detic. Specifically, ANDRIJ in his *"happywhale_arcface_baseline_eff7_tpu_768_inference"* [3] uses Detic and EfficientNetB6 [19] as the base model were he makes predictions with one model only. He updated the method and now has a model ensemble that predicts based upon 5 trained models, each trained separately as it is faster, using this technique the accuracy increases 5% from the base model obtaining 0.729 (Table 1) in the Photo-ID problem. ISAMU [12] obtains currently best results in HappyWhale competition (Table 1), he uses a simple ensemble method based on ANDRIJ and his previously mentioned method [3] and also an his other forked code *"Happywhale Effnet - B7 fork with Detic Training"* [2], on RABBIT and his *"0.720_EfF_B5_640_Rotate"* method [21], and finally it also uses Andrej Marichenko's *"Happywhale [0.679]"* [18]. All of these methods use as evaluation metric accuracy and their codes are open source.

One of the detection methods that correspond to the state of the art is Detic, which is a process created to improve the most significant contemporary shortcomings in this branch of computer vision, the main problem being the limited vocabulary of object detectors. In this method the detectors used as classifiers for a classification database are trained to significantly expand the vocabulary to thousands of concepts and using these trained models for different challenges. This is possible because of the structure used for the annotations of the predictions, which are not directly on

the boxes, allowing them to be more easily used and accommodated in methods such as neural networks [28]. On the other hand, the method that is at the same level on the detection process is YOLOV5, which is characterized by assigning labels having a fairly high degree of prediction given by its process of dividing the images into cells and in turn, sets of cells in grids where each cell has its own object detection independent of the others which uses the tensors of pytorch currently showing on the basis of COCO the fastest and most accurate results [25].

In addition, EfficientNet [19] is a fairly new CNN which has been used by most of the participants of the competition since it corresponds to a method with an architecture that presents in ImageNet a performance of 84.3% of precision being 8 times smaller than the rest of the networks of the state of the art, as well as the fastest. As can be seen in figure 2, which shows the performance of the methods in the reference database for the state of the art, having at the moment the best result. According to the authors of the network, its architecture is focused on creating a balanced system between the depth, width and resolution of the images using an effective composite coefficient which corresponds to the next step in the vanguard to the structure of ResNet and MobileNets.[19] Additionally, it is also worth expanding the definition of Ensemble methods, these methods combine different models in order to obtain one optimal predictive model. Instead of making a single model, one can make several models, use ensemble methods to take into account all of these models, average them and produce one final optimal method [17].

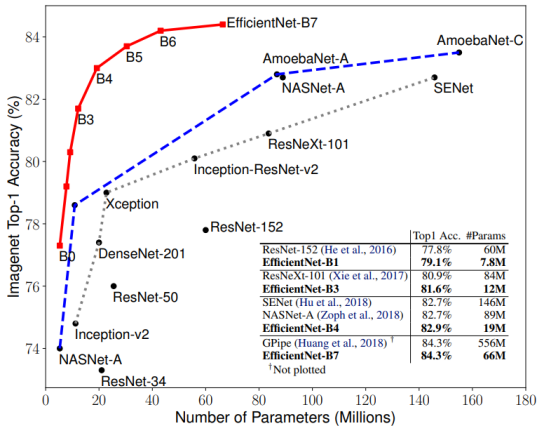


Figure 2. Graph on the performance of the EficcientNet architecture compared to the rest of the neural networks with the highest accuracy results in ImageNet taking into account its number of parameters. [19]

Regarding studies on the same problem, the HappyWhales foundation has a carried similar challenge in the past, however, with a dataset that only focused on more that 3000 individuals of Humpback whales [10]. This database is smaller in size and variety since it only consists of 33,300 images and only one specie of cetacean. On the other hand, this has been attempted with the Right Whale photo identification dataset from *"Applying deep learning to right whale photo identification"* [22] (it is also only one specie of cetacean). Taking these past applications of photo identification to deep learning, we can affirm that this competition is the state of the art, not only because of the greater number of images, but mainly because it consists of 26 species of cetaceans surpassing previous datasets as it aims for a greater generalization of the problem that is being studied. Due to the lack of studies of photo identification in animals, specifically cetaceans, currently there is a lot of room for improvement in the area and in this new framework of HappyWhales.

3. Approach

First of all, before performing the classification and photo identification tasks it is necessary to adjust the images to a fixed size taking into account that the methods proposed in this article are all neural network architectures. Therefore, an open source fin detection code of Meta's[15] object detection named DETIC is used, specifically through the already implemented DETIC method to the Happywhale dataset by Phalanx[20]. The new database of 512x512 images is used, where only the recognized fins are found in the image.

3.1. Classification Baseline

We use ResNet50 as a baseline [15], which comes from the well known family of CNN architectures and vastly used in relevant datasets such as ImageNet. The main feature of the ResNet's architecture family is the implementation of a deep residual learning framework to deal with the degradation problem of deep networks, as they get deeper and deeper, accuracy gets saturated and degrades rapidly[27]. That is why with residual learning, instead of expecting layers to fit a desired underlying mapping, ResNet can explicitly fit the desired layers with a residual mapping. Denoting the desired mapping as $H(x)$, x as the input of these layers and $F(x)$ the output. One can explicitly let these layers approximate a residual function $F(x) := H(x) - x$, this way the original function thus becomes $F(x) + x$ as represented in Figure 3 [15]. It is also worth mentioning that the database has errors in the classification categories, which had to be adjusted by us to minimize the bias in the experimentation, since there are certain species that have more than one category, as a result of a bad wording. An example of these are the categories "bottlenose_dolphin" and "bottlenose_dolpin" which belong to the same species, but in the database are taken as two different categories due to a typo of letter. The finding of the error was duly commented in Kaggle as well as other competitors, however, at the date

of this article the authors of the database have not made the correction.

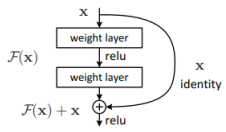


Figure 3. Residual learning example. [15]

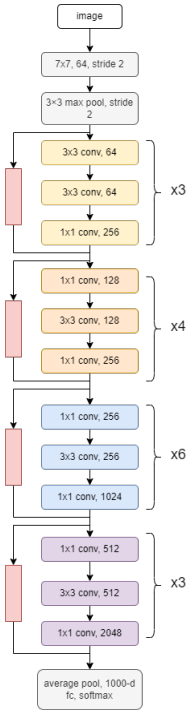


Figure 4. Architecture of ResNet50 used as baseline. In each rectangle the first argument indicates the size of the kernel and the other the number of convolutions. The red rectangles represent the shortcut connections used to implement residual learning while the multiplication numbers indicate how many times each block/layer is repeated (including the shortcut connection). Where need for the shortcut connections, dimension are adjusted.

The results presented in Tables 2 cannot be compared with the state of the art since the database used corresponds to an individual ID classification competition of each individual, therefore, the results found in the literature are focused only on the competition. Considering that we modified the annotations to perform in the first instance a classification using 2-fold cross validation technique on the training folder because we do not have the annotations of the test images, it makes this task a method which is not described in the literature and does not have a test server. Respecting the hyperparameter used, the models were trained in 10 epochs, with a learning rate of 0.0001, a weight decay of 0.0002 with an Adam optimizer and Cross-Entropy loss. Additional to the normalization of the dataset images, a random horizontal flip with probability of 0.5 was applied to the dataset.

Model	Precision	Recall	F1-score
ResNet50_Fold1	0.71	0.7	0.7
ResNet50_Fold2	0.77	0.7	0.71

Table 2. Macro average evaluation of ResNet50 in both folds.

3.1.1 Photo-ID

For the identification of each individual, a head of ArcFace [13] with Efficientnet as backbone was used as the baseline, which corresponds to the state-of-the-art method for face recognition. In the implementation of deep convolutional neural networks for face recognition, a multi-class classification task is performed, so this task is adjusted to the database used in this article, being the fins the tool for the individual identification of the different cetaceans. Arcface is a Deep Face Recognition method which fits any deep convolutional neural network since its main component is not the model architecture but the loss function[16]. The authors of this method propose the Additive Angular Margin loss function (Figure 6), with the objective of obtaining more significant discriminative features for face recognition. Within the literature it has been shown that Arcface helps to stabilize the training process as well as to increase the discriminative power between classes, since the commonly used loss functions such as softMax, where the features that are learned and optimized are separable for closed sets, contrary to the individual indentifying task where it is not discriminative enough.[13]

The Additive Angular Margin function calculates the angle between the current feature and the target weight and an additive angular margin is added to the target angle in order to retrieve the logit which is scaled by a fixed feature norm. After this, the softMax methodology is performed. On the other hand, EfficientNet is a CNN with 813 layers and fewer parameters, this CNN gradually increases the models width, deep and resolution (Figure 5) achieving better results than other models that rapidly get saturated and have a lot of parameters that make them inefficient.

In that order of ideas, for the current problem, the base line is proposed taking into account the best network that fits Arcface according to the literature which is Efficientnet, adjusting our database directly from the Sking-Cheng code [cita], doing the validation in the same way as in classification, using two fold cross validation. The optimizer used is Adam being used by the authors, as well as 20 epochs, with a learning rate of 0.0001 and weight decay of 0.0002.

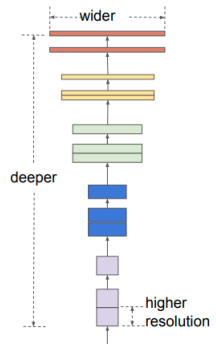


Figure 5. Efficient compound scaling in EfficientNet. [19]

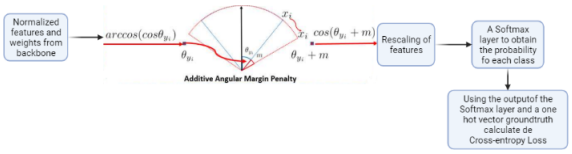
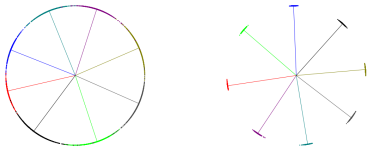


Figure 6. ArcFace head Structure. Here, the cosine of the features and weights from the backbone get feeded into Arcface, then calculate the $arccosine(\theta yi)$, get the angle between the features xi and the GroundTruth weight Wyi. The penalty m of the angular margin is added on the target angle θyi and the cosine of $yi + m$ is calculated. Finally, the result is multiplied by the feature scale s and passed through a Softmax which is used together with the Groundtruth to calculate the Cross-Entropy Loss [13] .



(a) Softmax (b) ArcFace

Figure 7. Distribution form of the distances between 8 classes for softmax where it is evident that there is no distance between classes, unlike ArcFace which allows to create a distance around each of the classes. [13]

Models	Accuracy	Top-5 Accuracy
Arcface_fold1	0.14	0.19
Arcface_fold2	0.15	0.20

Table 3. Results of Photo-ID baseline.

3.2. Proposed Method

The proposed method for cetacean identification is divided into the two main problems mentioned above.For the classification task, an experimentation is carried out using the deep convolutional neural networks corresponding to the state of the art,taking the baseline as a reference point, in order to find the combination of parameters with the best performance. In this order of ideas, changes are made both in the structure and in the architecture of the models, making a progressive evaluation of the variation of the parameters to make a combination among those that show better results. Therefore, we seek to vary the networks, if they have pre-trained weights, the data augmentation tools, the optimizer and the learning rate.

On the other hand, for the photo ID task, a similar experimentation methodology is proposed in the proposed method, where the tools of the state of the art are implemented taking into account the validation system presented for the partitioning of two fold cross validation data and a comparison is made with a division of the database with training folders, validation and test, since more than half of the individuals in the database have only one image, so the database is redistributed having in the test folder only individuals with more than one image and distributing that at least one of their photos is in the training folder. The new partitioning is done by percentages of 60% for train, 19% for validation and 19% for test. The hyperparameters of the model are adjusted taking into account the best results obtained in the validation.[13]

4. Experiments

4.1. Dataset

The database used for the development of this task is authored by the Happywhale foundation whose mission is to collect images of whales and dolphins around the world to monitor the different species of these animals for their preservation. Consistent with its ideal, the Happywhale foundation creates an open competition on the Kaggle website with a database of 78,989 images with more than 15,000 unique individuals and 26 different species of whales and dolphins (as can be seen in the distribution of Figure 8) with the objective of finding an artificial intelligence system capable of recognizing the different individuals in different images to create an automated monitoring system [11].

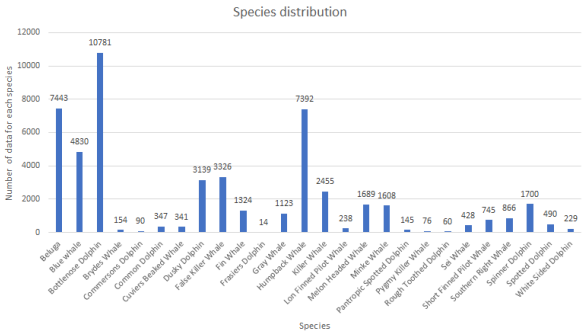


Figure 8. Species distribution graph in the database with annotations.

The database has annotations for 51033 images in a system that first recognizes whether it is a whale or a dolphin,

then the species to which it belongs and as a last step an ID created for individual recognition. The other 27956 images do not have annotations since they are reserved by the company for the evaluation system of the competition. For this article, only the images with annotations will be taken into account, since the aim is to carry out an individual evaluation in each of the classification categories. Likewise, it is pertinent to mention that the dimensions of the images do not have a single value.

Figure 9. visualizations of random images on the database.[11]

image	species	individual_id
51033 unique values	bottlenose_dolphin 19% beluga 15% Other (33926) 66%	15587 unique values
00021adfb725ed.jpg	melon_headed_whale	cadddb1636b9
000562241d384d.jpg	humpback_whale	1a71fbb72250
0007c33415ce37.jpg	false_killer_whale	60008f293a2b
0007d9bca26a99.jpg	bottlenose_dolphin	4b00fe572063
00087baf5cef7a.jpg	humpback_whale	8e5253662392
000a8f2d5c316a.jpg	bottlenose_dolphin	b9907151f66e

Figure 10. Annotations for 51033 images with the species to which it belongs and the ID created for individual recognition. [11].

By taking only the images with annotations, it is necessary to divide the data into training and evaluation folders. In the present case we chose to use the 2-fold cross validation technique since the number of images is considerably reduced with respect to the original database. The use of this data partitioning technique takes advantage of the total number of images since two models are created where one of the folders is used for training and the other for testing in a repeated manner by exchanging the folders. In this way, it is possible to give a result with a statistical analysis with a level of significance according to the means of the metrics and their standard deviations.[29]

To evaluate the classification problem, it was decided to use a multiclass confusion matrix. The confusion matrix corresponds to a tabulation system in which it is possible to visualize the performance in a prediction process. In its simplest form, it is used as a matrix of a binary classification process, however, it is possible to extrapolate it to multiclass data sets. From the confusion matrix we will obtain our specific evaluation metrics of precision, recall and f-score.

$$Precision = \frac{Tp}{Tp + Fp} \tag{1}$$

$$Recall = \frac{Tp}{Tp + Fn} \tag{2}$$

$$Fscore = 2\frac{Precision * Recall}{Precision + Recall} \tag{3}$$

On the other hand, for the Photo-ID problem, as the challenge of HappyWhales suggests, the metric to use is accuracy and Mean Average Precision (or Top-5 Accuracy).

$$Accuracy = \frac{Tp + TN}{Tp + Fn + Fp + Tn} \tag{4}$$

$$MAP@5 = \frac{1}{u}\sum_{u=1}^{U} P(k) * rel(k) \tag{5}$$

In eq 5 (Mean Average Precision), U is the number of images, (k) is the precision at cutoff *k, n* is the number predictions per image, and *rel(k)* is an indicator function equaling 1 if the item at rank k is a relevant (correct) label, zero otherwise.

4.2. Evaluation experiments

4.2.1 Classification 2-Folds Cross Validation

As for the classification results, there are 6 different models, varying between pre-trained weights, optimizer, Lr Decay and architecture used, there are 12 models of weights since a training is performed for each of the folders and the results are compared, in this article it is evident that the difference in metrics between each of the folders is not significant and, therefore, there is a validation of the results that the results are not composed of a random process. The best result corresponds to the EfficientNet model pre-trained on ImageNet with Adam's optimizer. The pre-trained models have the best results overall, however, their results are not only given by the ImageNet weights; this can be seen in the ResNet 50 model not pre-trained with Adam's optimizer which is only 15 points away from its pre-trained version. All models were trained to the same number of epochs and the batch size varied due to the availability of the computational tools.

As a final evaluation of the results obtained, the best model is taken into account and the metrics for each of the 26 classes that the database has in the classification task are taken into account. The results per class vary in their performance independently of how large or small each class is,

Models	Pretained in ImageNet	Lr Decay	Optimizer	Fold1			Fold2		
				Precision	Recall	F1-score	Precision	Recall	F1-score
ResNet50	No	None	Adam	0.71	0.7	0.7	0.77	0.70	0.71
ResNet50	No	None	SGD	0.05	0.05	0.05	0.05	0.05	0.05
ResNet50	No	None	ASGD	0.13	0.14	0.12	0.13	0.09	0.07
ResNet50	Yes	None	Adam	0.86	0.79	0.81	0.85	0.84	0.84
ResNet50	Yes	0.1	Adam	0.64	0.58	0.58	0.76	0.69	0.70
EfficientNet	**Yes**	**None**	**Adam**	**0.92**	**0.88**	**0.89**	**0.93**	**0.84**	**0.86**

Table 4. Results of Clasification models.

so it cannot be inferred that, being an unbalanced database, there are bad metrics for classes with a very small number of images.

Classes	Precision	Recall	F1-score	# Images
Beluga	0.99	1.00	0.99	3721
Blue Whale	0.99	0.98	0.99	2414
Bottlenose Dolphin	0.98	0.98	0.98	5389
Brydes Whale	0.94	0.86	0.90	76
Commersons Dolphin	1.00	0.91	0.95	44
Common Dolphin	0.90	0.94	0.92	173
Cuviers Beaked Whale	0.96	0.88	0.92	170
Dusky Dolphin	0.98	1.00	0.99	1569
False Killer Whale	0.95	0.98	0.96	1662
Fin Whale	0.96	0.95	0.96	661
Frasiers Dolphin	0.50	0.17	0.25	6
Short Finned pilot Whale	0.98	0.89	0.93	370
Gray Whale	0.97	0.97	0.97	561
Humpback Whale	0.98	0.98	0.98	3695
Killer Whale	0.99	0.97	0.98	1226
Long Finned Pilot Whale	0.89	0.92	0.90	118
Melon Headed Pilot Whale	0.92	0.97	0.94	844
Minke Whale	0.98	0.96	0.97	803
Pantropic Spotted Dolphin	0.80	0.46	0.58	72
Pygmy Killer Whale	0.89	0.68	0.77	37
Rough Toothed Dolphin	0.81	0.59	0.68	29
Sei Whale	0.97	0.98	0.97	213
Southern Right Whale	0.97	0.99	0.98	432
Spinner Dolphin	0.96	0.95	0.95	849
Spotted Dolphin	0.85	0.94	0.89	244
White Sided Dolphin	0.90	1.00	0.95	114

Table 5. Results per class of EfficientNet trained in fold1 and evaluated in fold2. Best Classification model.

4.2.2 Photo-ID 2-Folds Cross Validation

Regarding the results obtained in the individual identification task, 3 different models were made, 2 of them correspond to a validation process of 2 fold cross validation, varying on the base line the Lr Decay to Lr values of 3e-4 and a min decay of 1e-6, likewise, another transform was added in the data augmentation in addition to the filp, which is Color Jitter.

On the other hand, the last model corresponds to an experimentation where the distribution of the database is changed with a partition of 60% for training, 19% for validation and 19% for test, with the same hyperparameters of the previous experimentation. The results are corresponding to the test folder. This last experiment is performed to observe the behavior of the model taking into account a more balanced distribution considering that there are 9258 individuals with only one image, 3091 with two images, 773 with 3 images and 2355 with more than three images. This last experimentation shows to have the best results between the other experiments and the base line, which allows to define that this new distribution of the database allows to obtain better metrics by balancing the data.

Classes	Precision	Recall	F1-score	# Images
Beluga	0.98	1.00	0.99	3722
Blue Whale	0.95	0.99	0.97	2416
Bottlenose Dolphin	0.94	0.99	0.96	5392
Brydes Whale	0.74	0.91	0.82	78
Commersons Dolphin	0.93	0.91	0.92	46
Common Dolphin	0.98	0.82	0.89	174
Cuviers Beaked Whale	0.97	0.84	0.90	171
Dusky Dolphin	1.00	0.99	1.00	1570
False Killer Whale	0.98	0.90	0.94	1664
Fin Whale	0.96	0.92	0.94	663
Frasiers Dolphin	1.00	0.12	0.22	8
Short Finned Pilot Whale	0.89	0.90	0.89	375
Gray Whale	0.94	0.96	0.95	562
Humpback Whale	0.99	0.95	0.97	3697
Killer Whale	0.98	0.94	0.96	1229
Long Finned Pilot Whale	0.96	0.88	0.92	120
Melon Headed Pilot Whale	0.93	0.95	0.94	845
Minke Whale	0.96	0.93	0.94	804
Pantropic Spotted Dolphin	0.69	0.67	0.68	73
Pygmy Killer Whale	0.81	0.33	0.47	39
Rough Toothed Dolphin	0.77	0.32	0.45	31
Sei Whale	0.99	0.98	0.99	215
Southern Right Whale	0.99	0.98	0.99	434
Spinner Dolphin	0.95	0.94	0.95	851
Spotted Dolphin	0.86	0.87	0.86	246
White Sided Dolphin	0.97	0.95	0.96	115

Table 6. Results per class of EfficientNet trained in fold2 and evaluated in fold1. Best Classification model.

For the results of the 2 fold cross validation it can be observed that there is no significant difference between the metrics and therefore it can be inferred that it does not correspond to a random process.

Models	Accuracy	Top-5 Accuracy
Arcface_ColorJitter_lrdecay_fold1	0.14	0.19
Arcface_ColorJitter_lrdecay_fold2	0.13	0.19
Arcface_ColorJitter_New dataset	**0.18**	**0.23**

Table 7. Results of Photo-ID changing the data augmentation and the distribution of the dataset.

5. Discussion

In conclusion, in the classification problem, for our best method we obtain F1-scores of 0.89 and 0.86 in fold1 and fold2 respectively using EfficienteNet_b7, which outperforms our baseline of ResNet50 (Not Pretrained) by about 19%, and Resnet50 using the same hyperparameters and optimizer by 8% in fold1 and 2% in fold2. There are several factors that contribute to the better performance of EfficientNet, first and most notably is the implementation of transfer learning by using pretrained weights from ImageNet, this difference can be clearly noticed when comparing the experiments of ResNet50 where only the Pretrain-

ing parameter differs. There, one can observe that transfer learning improves performance of Resnet50 by about 11% to 13% taking into account both folds. The other factor that clearly affects performance is the optimizer, as we test Adam, Stochastic Gradient Descent (SGD) and Average Stochastic Gradient Descent (ASGD). Here, we observe that SGD and ASGD clearly don't work well with the established problem as their results are terrible (figure 4). The reason to this terrible results may reside in the number of training epochs, as all Classification models were trained for 10 epochs, and one of the main problems of SGD and ASGD is that the path taken to the minima is usually very noisy, so it takes longer to converge, however, ASGD somewhat softens the noise which can be observed in the superior result to SGD. So, probably, if left for more training epochs the models that use SGD and ASGD could increase their performance. Adam doesn't suffer from this as it implements two advantages, specifically that it maintains a per-parameter learning rate that improves performance on problems with sparse gradients and those per-parameter learning rates are adapted based on the average of recent magnitudes of the gradients for the weight, meaning it does well in noisy problems [4]. Adam proves to be more efficient than other optimizers and has done so for many others [6], and in [23] is recommended as the best overall choice. Finally, the other reason of the better performance is the architecture of EfficientNet, which uses a Compound scaling that gradually increases the models width, deep and resolution with fewer parameters, achieving better results without getting saturated and using better the resources.

On the other hand, and important observation must be made in that some classes from the 2 folds differ a lot in their performance, this may be due to more difficult examples in one fold than another. This problem can be explained in part by the fact that the implemented detection method, Detic, wasn't perfect (when it couldn't detect the cetacean it did nothing) and some cetaceans aren't centered in the image and may be too small when compared to the surrounding water in the image.

Additionally, The results obtained for the final model in the identification task are not entirely favorable, the metrics reflect a possible problem on the part of the model to adjust the facial recognition process to the flippers. During the training and validation process the values of the hyperparameters were adjusted which show to be quite sensitive compared to the literature[16], however, a favorable optimization process was observed in the decrease of the loss and the validation process for each of the epochs. Therefore, it is inferred that one of the main problems is in the unbalanced form of the data, taking into account that the distribution of the data was adjusted with the objective of minimizing this quantity (an improvement is seen but it is not very significant), one of the objective tasks of the database is lost, which is to be able to identify new individuals, it would be pertinent in the future to use the model proposed with the complete database and test its performance in the test of the competition page. The parameters such as the optimizer and the data augmentation method of rotating the image were not varied since its good performance in this task is well known. As for the ArcFace parameters, there is no experimentation on these parameters, leaving those established by the authors as an angular margin of 0.5. As there are more than 15000 individuals, it is not possible to detect the closeness between individuals by species to validate the fact that the Additive Angular Margin loss function is correctly separating the classes at the species level.[13]

Therefore, it is concluded that the task may be somewhat altered by the distribution of individuals in the database taking into account the results from the challenge (Table 1) and that we don't have access to the test annotations. 9258 individuals only have one photo and 3091 have 2 in the training dataset. It is also pertinent to perform a more exhaustive training process since each model was run for a number of 25 epochs. Likewise, as a future work, it is planned to perform a more exhaustive experimentation in this section by varying more paramters and implementing an inference function which validates the different levels of multiclass in the form of a cladogram, as well as using One Shot Learning and Few Shot Learning for the individuals with one photo or very few, as only 110 individuals have more than 50 photos.

6. Ethical Considerations

This article takes into account the ethical consideration of obtaining a method capable of decreasing the recognition time of cetaceans and obtaining a unique information network focused on monitoring the marine ecosystem. The organization that created the competition, HappyWhales, does not seek to generate any monetization on this project and are aware that it is not a project intended to be industrialized, so they encourage the scientific community to work through the prize of the competition which is sponsored by the organizing which has achieved great advances in the preservation of marine life. In there last project through the data obtained they were capable to regulated speed of cruise ships in regions rich in whales.

On the other hand, one of the ethical problems in the publication of this project is that it can be used to locate areas of high marine life sightings by criminal groups for the trafficking or commercialization of the different species described, which is an illicit practice that the authorities have sought to regulate over time. This consideration goes unnoticed since the control agencies have access to the same information and can increase the monitoring of these risk zones.

References

[1] World Cetacean Alliance. Why cetaceans. 1

[2] ANDRIJ. Happywhale - effnet b7 fork with detic training. 2022. 2

[3] ANDRIJ. happywhale_arcface_baseline_eff7_tpu_768_inference. 2022. 2

[4] Jason Brownlee. Gentle introduction to the adam optimization algorithm for deep learning. 2017. 8

[5] Mark N. Hutchinson C. Michael Bull Claire E. Treilibs, Chris R. Pavey. Photographic identification of individuals of a free-ranging, small terrestrial vertebrate. 2020. 2

[6] Jimmy Lei Ba Diederik P. Kingma. Adam: A method for stochastic optimization. 2015. 8

[7] Thea Bechshoft Gaby Bellazi ECM Parsons, Sarah Baulch. Key research questions of global importance for cetacean conservation. *Endangered Species Research*, 27(2):113–118, 2015. 1

[8] Glacialis Expeditions. Photo identification what for and what is important? 2020. 1

[9] IUCN – SSC Cetacean Specialist Group. Status of the world's cetaceans, 2021. 1

[10] Happywhale. Humpback whale identification. 2019. 3

[11] Happywhale. Happywhale - whale and dolphin identification. 2022. 1, 2, 5, 6

[12] ISAMU. simple ensemble of public best kernels. 2022. 2

[13] Niannan Xue Jiankang Deng, Jia Guo. Arcface: Additive angular margin loss for deep face recognition. 2019. 4, 5, 8

[14] Renate Reijns Jurgen den Hartog. Interactive individual identification system, 2016. 2

[15] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition. *arXivy*, 2015. 3, 4

[16] Petiushko A Komkov S. Real-world adversarial attack on arcface face id system. *International Conference on Pattern Recognition*, 2021. 4, 8

[17] Evan Lutis. Ensemble methods in machine learning: What are they and why use them? *Towards Data Science*, 2017. 3

[18] Andrej Marichenko. Happywhale [0.679]. 2022. 2

[19] Quoc V. Le Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. 2020. 2, 3, 5

[20] Phalanx. whale2-cropped-dataset. 2022. 2, 3

[21] RABBIT. 0.720_eff_b5_640_rotate. 2022. 2

[22] Christin Brangwynne Khan Maciej Klimek Jan Kanty Milczek Marcin Mucha Robert Bogucki, Marek Cygan. Applying deep learning to right whale photo identification. *Conservation Biology*, 33(3), 2018. 1, 2, 3

[23] Sebastian Ruder. An overview of gradient descent optimization algorithms. 2016. 8

[24] Richard Szeliski. Computer vision: Algorithms and applications 2nd edition. pages 368–369. 2

[25] Ultralytics. You only look once v5. 2020. 2, 3

[26] Bruno Spataro Simon Chamaille-Jammes Dominique Allaine Christophe Bonenfant Vincent Miele, Gaspard Dussert. Revisiting giraffe photo-identification using deep learning and network analysis. 2020. 2

[27] Gao L Wen L. A transfer convolutional neural network for fault diagnosis based on resnet-50. *Neural Computing and Applications*, 2021. 3

[28] Armand Joulin Xingyi Zhou, Rohit Girdhar. Detecting twenty-thousand classes using image-level supervision. 2022. 2, 3

[29] Shukla S. Yadav, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *6th International conference on advanced computing*, 2016. 6

7. Credits

Mauricio Salim Gómez Chicre - Paper, coding of classification baseline and adaptation of Photo-ID baseline
Sebastian Ojeda Alonso - Paper, coding, experimentation of Photo-ID implementations, adaptation of classification baseline.