

Robust Re-identification of Manta Rays from Natural Markings by Learning Pose Invariant Embeddings

Olga Moskvayak, Frederic Maire, Asia O. Armstrong, Feras Dayoub and Mahsa Baktashmotlagh

Abstract—Visual identification of individual animals that bear unique natural body markings is an important task in wildlife conservation. The photo databases of animal markings grow larger and each new observation has to be matched against thousands of images. Existing photo-identification solutions have constraints on image quality and appearance of the pattern of interest in the image. These constraints limit the use of photos from citizen scientists. We present a novel system for visual re-identification based on unique natural markings that is robust to occlusions, viewpoint and illumination changes. We adapt methods developed for face re-identification and implement a deep convolutional neural network (CNN) to learn embeddings for images of natural markings. The distance between the learned embedding points provides a dissimilarity measure between the corresponding input images. The network is optimized using the triplet loss function and the online semi-hard triplet mining strategy. The proposed re-identification method is generic and not species specific. We evaluate the proposed system on image databases of manta ray belly patterns and humpback whale flukes. To be of practical value and adopted by marine biologists, a re-identification system needs to have a top-10 accuracy of at least 95%. The proposed system achieves this performance standard.

I. INTRODUCTION

RE-IDENTIFICATION of animal individuals by unique natural markings in photo databases is an effective and non-invasive mark-recapture tool for monitoring populations [1]. Tracking population dynamics of animals such as manta rays is critical owing to their vulnerable conservation status, and economic importance in both ecotourism and fisheries [2]. These species cannot sustain heavy exploitation [3], and the trade of manta ray gill rakers is believed to be responsible for driving population declines upwards of 80% in some locations [4]. Some species such as humpback whales are no longer threatened by commercial whaling. The conservation effort is now focused on the identification of individual humpback whales to better understand their use of breeding and feeding areas [5].

Our research is focused on developing an automated system for visual re-identification of animals that bear unique natural

O. Moskvayak, F. Maire and F. Dayoub are with the School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia.

A. O. Armstrong is with the School of Biomedical Science, The University of Queensland, St Lucia, QLD 4072, Australia.

M. Baktashmotlagh is with the School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia.

Corresponding author O. Moskvayak: olga.moskvayak@hdr.qut.edu.au

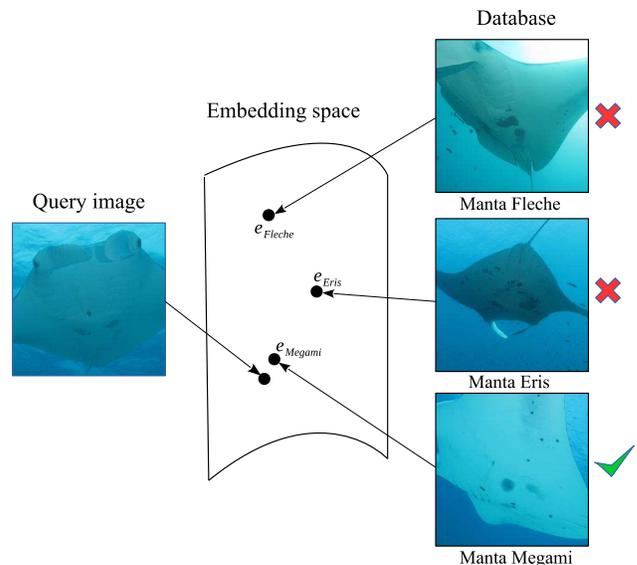


Fig. 1. The proposed system learns embeddings for images from the database. The embeddings of the same individual are brought close together and embeddings of different manta rays are pushed further apart. A new query image is matched to the database by finding the closest points in the embedding space. The system learns embeddings that are invariant to viewing angle and illumination. Photo credit: Chris Garraway.

markings. We demonstrate the suitability of the proposed system on photo databases of manta ray belly patterns and humpback whale flukes. Manta rays have a unique spot pattern on their ventral surface that allows individuals to be distinguished from one another. The spot pattern is conserved throughout the animals life, much like a human fingerprint. Examples of spot patterns are shown in Fig. 1 and Fig. 2. Humpback whales have patterns of black and white pigmentation and scars on the underside of their tails that are unique to each whale.

There are a number of factors that make animal re-identification based on natural markings challenging. Photo databases often rely on input from citizen scientists to fill in data gaps when researchers are not in the field. This means image quality cannot be guaranteed as camera parameters, and the angle of image capture vary. Other factors include poor visibility (especially for underwater images), illumination, and small objects occluding the pattern on the animal.

The current state-of-the-art manta ray recognition system Manta Matcher [6] requires the user to manually align and normalize the 2D orientation of the manta ray within the



Fig. 2. The camera angle can vary dramatically. Here the same manta ray, named Eris, was photographed from two different viewpoints. A homography transformation is required to align the belly patterns of the two images. Photo credit: Chris Garraway.

image, and select a rectangular region of interest containing the spot pattern. The Manta Matcher works best with photos taken perpendicular to the manta’s ventral pattern with no reflective particles in the water and in good lighting conditions. In practice, these constraints limit the use of photos from citizen scientists and some marine biologists still do the identification manually using a handcrafted decision tree. A common idea that has been applied to several species for recognizing individual animals is to search for an affine transformation matching patterns present in two distinct images (lizards [7], arthropods [8], sharks [9], turtles [10]). However, this approach requires annotating body landmarks on each individual image in the same order. This is not suitable for manta rays as we want to accept images of the animals in a wide range of poses with no requirement that all body landmarks are clearly visible.

Convolutional neural networks (CNN) have been applied to the problem of animal identification as a classification problem [11], [12], [13], [14]. It means that the trained model is only able to identify the animals presented during training.

It is highly desirable to have a system that is not only capable of recognizing animals whose images have been used to train the neural network, but also capable of recognizing animals whose images have been added to the database well after the network has been trained without requiring the re-training of the network on these new instances. This paper focuses on this more challenging and less studied problem for animal re-identification.

In this work we focus on eliminating some constraints of previous wildlife matching systems such as requirements for high image quality and a clear view of the animal markings in the image. We propose a solution inspired by advances in deep learning for face re-identification. Our approach uses a CNN to learn embeddings for images of animal markings in such a way that the distance between embeddings of the same individual is smaller than the distance between embeddings of this individual and other animals (see Fig. 1).

The main contribution of this work is a novel visual wildlife re-identification system with the following properties:

- 1) robustness to viewpoint changes, small occlusions and lighting conditions, and therefore ability to match images from citizen scientists;
- 2) re-identification of individuals never seen during training.

The paper is organized as follows: in Section II we discuss related work on re-identification. Our approach to learning embeddings is described in Section III. The experimental setup and results are presented in Section IV.

II. RELATED WORK

The techniques that have been proposed for photo-identification of animal natural markings vary in the core methods used, amount of user involvement and ability to be adapted to different species. We review solutions used in practice for different cases.

Matching natural patterns has been approached by exhaustively generating two-dimensional affine transformations based on user provided key points and comparing each transformation of a candidate example with the examples stored in a repository [7], [8], [9], [10]. The algorithm was implemented in a solution called APHIS (Automated Photo-Identification Suite) and applied for re-identification of lizards [7], arthropods [8], spotted raggedtooth sharks [9] and turtle flippers [10]. However, the method requires a user to select key points and identify the most distinctive spots for each image.

Some methods have been developed for specific species and, while performing well on these, are not easily transferable to other species. High-contrast colour patterns of humpback whale flukes [15] and dolphin dorsal fins [16] are matched by extracting hand-crafted features from corresponding segments obtained by overlaying a grid on a region of interest. This method is not robust to viewpoint changes.

Another approach identifies individual cetaceans from images showing the trailing edge of their fins by generating a representation of integral curvature of the nicks and notches along the trailing edge [17].

Current systems used in practice (Manta Matcher [6], HotSpotter [18]) are based on automated extraction and matching of keypoint features using the Scale-Invariant Feature Transform (SIFT) algorithm [19] with different modifications and enhancements to work on specific cases. While the algorithm works well on images that clearly show the pattern of interest, it is not robust to large changes in camera viewpoint, occlusions and variations in illumination.

The task of animal visual re-identification is related to the face recognition problem that has been extensively studied with deep learning in recent years [20]–[22]. The main idea is learning a function using a CNN that maps from a face image space to a space of a smaller dimension where the distance between the learned embedding vectors corresponds to a face similarity measure [20], [23]. The network is trained on labelled image pairs or triplets to learn a face similarity measure under which the distance between the embeddings of faces from the same person is reduced as much as possible and that of the distance between embeddings of faces of different people is increased. The problem is then reduced to the nearest neighbours search problem in Euclidean space, which can be solved by efficient approximate nearest neighbours search algorithms [24].

The difference between face verification and animal re-identification is that a face image is typically normalized to

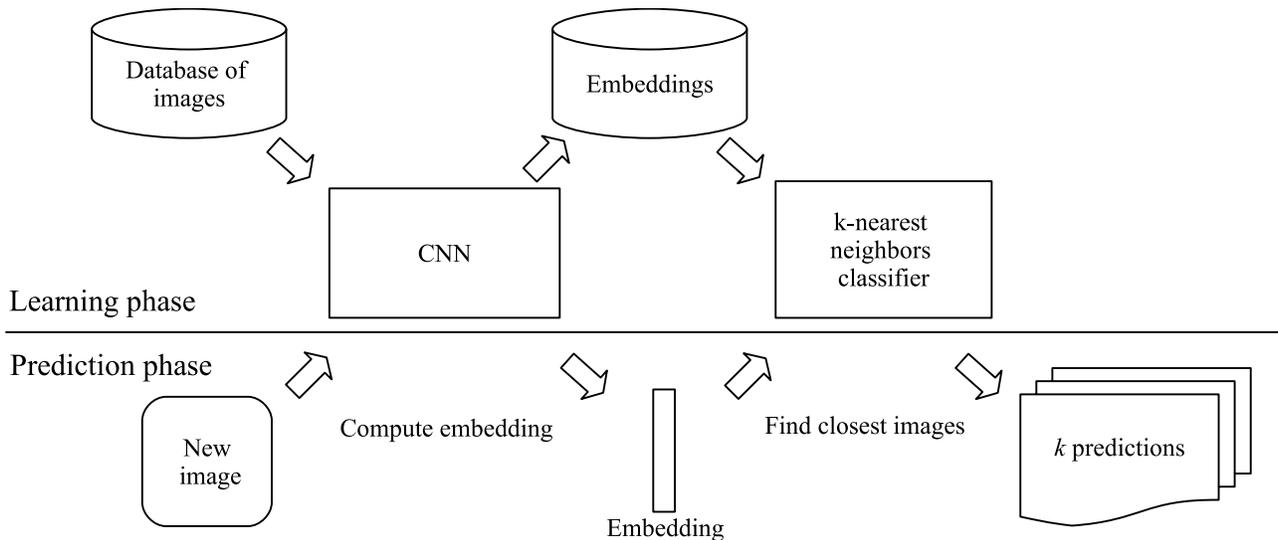


Fig. 3. System architecture. All images from the current database are passed through the trained CNN model to compute embeddings and fit a nearest neighbors classifier in an embedding space. At the prediction step, k predictions are obtained by computing an embedding using trained CNN and finding the closest points from the database using the nearest neighbors classifier.

an upright position whereas a pattern on an animal body is not necessarily in a canonical position and can appear at different angles. See an example of the same manta ray viewed from different vantage points in Fig. 2. A robust identification system should be invariant to the pose of the object of interest and viewing angle. In our previous work [25], we investigated the difficulty of recognizing a set of artificially generated patterns subjected to various projective transformations to simulate the variations in appearance of natural markings from different vantage points. This previous study explored Siamese [23] and Triplet [24] architectures with different loss functions for learning the homographic equivalence between patterns. It was concluded that these architectures with a relatively simple CNN in its core were suitable for pattern re-identification. The results were promising and we have now extended this approach to real images of animal markings in the wild.

III. LEARNING EMBEDDINGS

Throughout the paper, we say that images from the same individual animal *belong to the same class*. Images of different individual animals are said to be from *different classes*. The re-identification task can be formulated as a classification problem where the number of classes is in the order of thousands and not known in advance, and the number of examples for each class is small. The following section gives an overview of the architecture of our re-identification system.

A. System architecture

The system, illustrated in Fig. 3, consists of a CNN that produces embeddings for images and a k-nearest neighbors classifier in the embedding space. During the learning phase, we train a CNN on a database of labelled images. During the prediction phase, a new query image is fed to the network to produce an embedding. The first k animals in the embedding

database that are closest to the embedding of the query image are returned. Two outcomes are possible during the verification of the identity of the animals. Either the marine biologist confirms that the query image corresponds to one of the k returned animals or the query image is considered to be from a never seen before animal. In the first case, the query image is added to the record of the recognized animal. In the second case, a new animal entry is created. Over time, new images are added to the database but the CNN is not systematically retrained on the extended dataset. The network is able to match against images that were in the database during training as well as against images added later.

B. Model

We have adapted a model proposed in FaceNet [21] that learns embeddings for faces by minimizing a triplet loss. Initially, it was claimed that representation learning with the triplet loss is inferior to a combination of classification and verification losses [26], [27]. However, modifications of the triplet loss (angular loss [28], magnet loss [29]) and smart triplet mining strategies (semi-hard [21], batch-hard [30], [31]) has proved that a model can successfully learn an end-to-end mapping between images and an embedding space.

The model consists of convolutional layers to extract features from an input image, a global pooling layer over feature maps and a fully connected layer to produce an embedding vector. We compare different CNN architectures as a base network, see details in Section IV-B3. The convolutional layers output a 3D array (e.g. $8 \times 8 \times 512$) that is then passed to a global pooling layer.

A global pooling layer takes the average of each feature map along the spatial axes (e.g. a tensor $8 \times 8 \times 512$ is transformed into a tensor 1×512). We follow FaceNet [21] and favor a global average pooling layer instead of a fully connected layer after convolutional layers. The global pooling layer makes

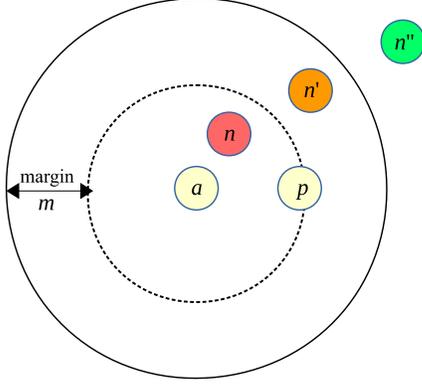


Fig. 4. Examples of a hard triplet, a semi-hard triplet and an easy triplet of embeddings. Anchor embedding a , its closest positive point p (same class as a) and the red negative n (different class from a) form a hard triplet as the negative is closer to the anchor than the positive example. The triplet (a, p, n') is a semi-hard triplet as the orange negative n' lies within the margin from the positive. The triplet contributes a positive value to the loss function. Whereas the triplet (a, p, n'') with the green negative n'' is an easy triplet because it contributes zero to the loss function.

the output of the network invariant to the size of the input images. Moreover, as the layer has no parameters, overfitting is avoided [32]. The layer sums out the spatial information so it is more robust to spatial transformations of its input. Pooled features maps are passed to a fully connected layer to produce an embedding vector.

C. Loss function

Our model is optimized using the triplet loss function [24] which accepts triplets of images. Let us define a triplet (I^a, I^+, I^-) where an image I^a (*anchor*) and an image I^+ (*positive*) are from the same class and an image I^- (*negative*) is from a different class. The function D between two input images I and J is defined as the Euclidean distance between their embeddings $f(I)$ and $f(J)$. That is, $D(I, J) \stackrel{\text{def}}{=} \|f(I) - f(J)\|_2$

The triplet loss function \mathcal{L} encourages the squared distances between positive pairs of embeddings to become smaller than the squared distances between negative pairs of embeddings by a given margin m :

$$\mathcal{L} \stackrel{\text{def}}{=} \sum_{i=1}^N \max(0, m + D(I_i^a, I_i^+) - D(I_i^+, I_i^-)) \quad (1)$$

where N is the number of training triplets.

We also did experiments with the Siamese network architecture [23] and a contrastive loss function [33] over randomly generated pairs, however, the results were not as good as the results obtained with the triplet loss function.

D. Example mining

The strategy for selecting triplets for learning embeddings plays an equal or more important role than the loss [34].

Generating random triplets for training with the triplet loss would result in many triplets that are already in a correct position and contribute zero loss to (1). Several strategies have been proposed to optimize training with the triplet loss function. Batch-hard triplet mining [30] selects the hardest positive (the furthest example from the same class) and the hardest negative (the closest example from a different class) within a batch for each anchor image. Another technique, distance-weighted sampling [34], selects a negative example with a probability function of the distance to the negative example.

We follow the semi-hard triplet mining strategy proposed by [21] as we found experimentally that this approach works better than batch-hard strategy for our application domain. The triplet loss is calculated over triplets that contribute positive value to the loss function. In other words, these negative examples lie within a margin from the positive examples (see Fig. 4). The selected triplets are not necessarily the hardest within a batch but they violate the constraint $D(I_i^a, I_i^+) + m < D(I_i^a, I_i^-)^2$.

The triplet mining strategies listed above require computing embeddings in order to select triplets. This can be achieved by precomputing embeddings every n steps using the most recent network checkpoint. We adopt a more computationally efficient online mining strategy [21] where triplets are generated on the fly after the embeddings have been computed and before the evaluation of loss function and backpropagation phase.

E. Evaluation methodology

We evaluate the performance of the system by computing the following metrics:

- true positive rate on pairs from the test set;
- top- k accuracy on the test set ($k = 1, 5, 10$).

1) *Validation on pairs*: The network performance is evaluated on pairs generated from the test set using a method proposed in [21]. The set of pairs of images from a same class is denoted as P_+ and the set of all pairs from different classes is denoted as P_- . Let us define the set of *true accepts* TA for a threshold d as the set of correctly classified positive pairs with a threshold d :

$$\text{TA}(d) \stackrel{\text{def}}{=} \{(i, j) \in P_+, \text{ with } D(I_i, I_j) \leq d\}.$$

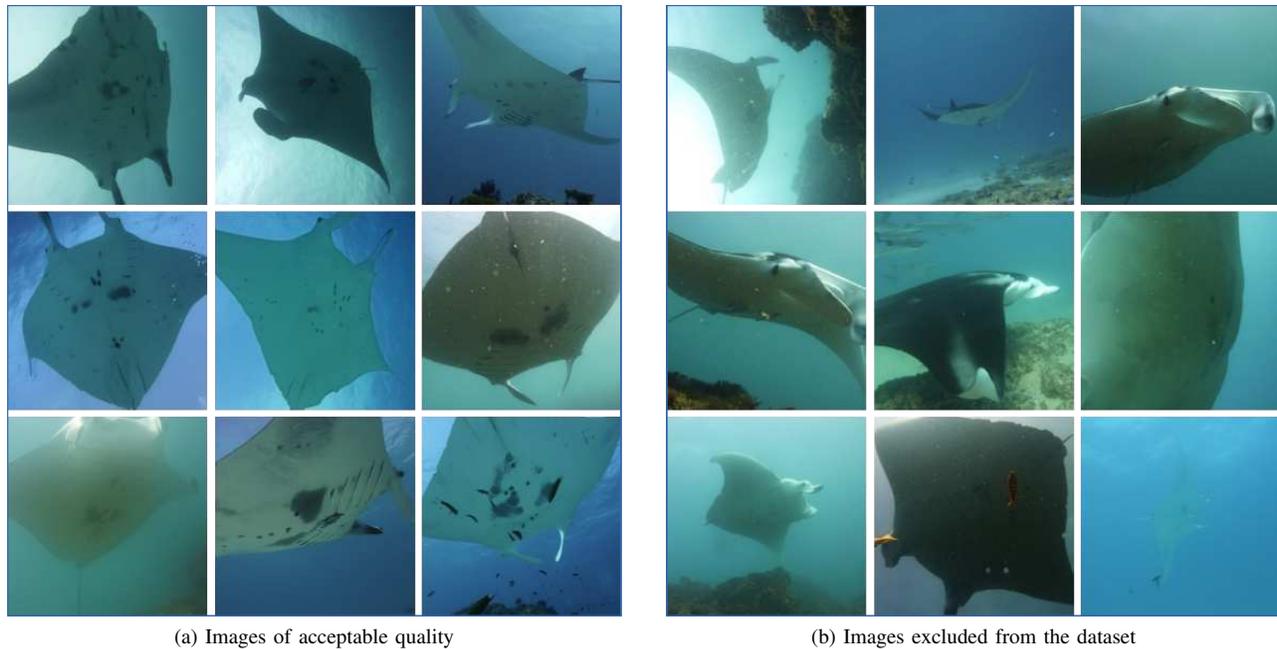
The set of *false accepts* FA is defined as the set of negative pairs that are incorrectly classified as positive with a threshold d :

$$\text{FA}(d) \stackrel{\text{def}}{=} \{(i, j) \in P_-, \text{ with } D(I_i, I_j) \leq d\}.$$

We calculate the *true positive rate* TPR (or *recall*) and the *false acceptance rate* FAR for a given threshold d as:

$$\text{TPR}(d) \stackrel{\text{def}}{=} \frac{|\text{TA}(d)|}{|P_+|}, \quad \text{FAR}(d) \stackrel{\text{def}}{=} \frac{|\text{FA}(d)|}{|P_-|}$$

Thanks to the relatively small size of the test datasets, all possible pairs are generated. The models are evaluated by plotting ROC curves and computing the area under the curve.



(a) Images of acceptable quality

(b) Images excluded from the dataset

Fig. 5. Not all images of manta rays are acceptable for training and testing the system. An image is accepted if the belly pattern is visible (even at oblique angles and in muddy water), see examples in (a). We accept more challenging images than other methods [6]. Examples of excluded images in (b): back of the manta, side view, poor underwater visibility. Photo credit: Lydie Couturier.

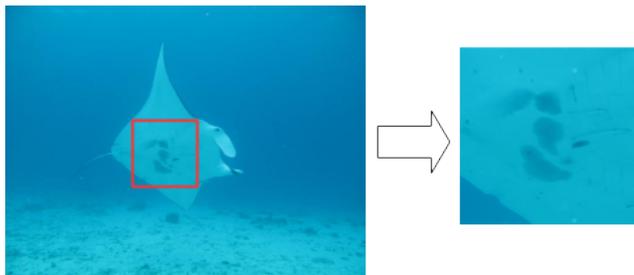


Fig. 6. The user is required to draw a bounding box around the region containing the natural markings. Images are cropped to contain the pattern of interest only. This is the only input required from the user. Photo credit: Chris Garraway.

The models are compared with respect to the true positive rate TPR at the threshold d when the false acceptance rate FAR = 0.01.

2) *Accuracy evaluation on the test set:* From a marine biologist’s point of view, a reliable system should have at least 95% top-10 accuracy. The accuracy of re-identification depends on the number of matching images in the database for each query image. We consider a realistic scenario where each query individual has two matching images in the database (our databases have at least 3 images for each individual). If there are more images per individual in the database, the task of re-identification becomes easier. For training, the dataset is partitioned into a training set and a test set in such a way so each individual animal appears exclusively either in the training set or in the test set. For testing, the database is made of the training set images plus $m = 2$ random images for each test individual. The rest of test images are used as query images. The accuracy is averaged over all test individuals in

multiple runs by moving different images from the test set to the database. We also analyze the effect of varying the number m in Section IV-C6. A similar evaluation procedure has been performed in [6], [18] on different datasets.

IV. EXPERIMENTS

A. Datasets

1) *Manta ray belly patterns:* The experiments have been conducted on a dataset of manta ray images from Project Manta (a multidisciplinary research program based at the University of Queensland, Brisbane, Australia). Images have been manually checked to select the ones that show a pattern on a belly with enough clarity to be recognized by a human. See some examples in Fig. 5 (left). The dataset is challenging as it contains photos of the patterns taken at oblique angles, in a muddy water or with some small occlusions (small fish, water bubbles). Uninformative images such as the view of the back of a manta, partial views or unclear patterns have been removed from the dataset. See examples in Fig. 5 (right). Each image has been manually annotated with a bounding box around the pattern. Then, each image has been cropped to the area inside the bounding box (Fig. 6). Manually highlighting the belly pattern region is the only input required from the user in our application.

The resulting dataset (see details in Table I) is partitioned into the training set (96 individuals) and the test set (24 individuals).

2) *Humpback whale flukes:* The dataset of humpback whale flukes comes from the Happy Whale organization (happywhale.com) [35], [36]. The main challenge with this dataset is the small number of images per whale with two-thirds of the whales having one or two sightings. For training

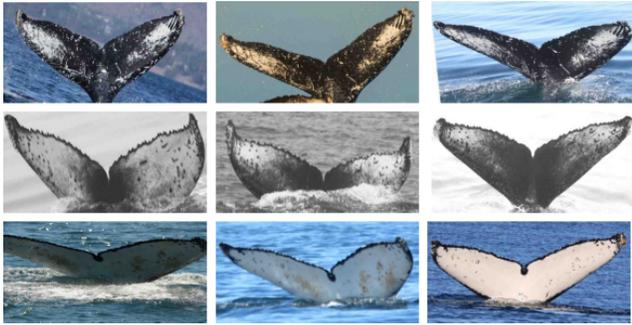


Fig. 7. Images from the whale dataset. Each row shows three images of the same fluke. List of photo credits is provided in acknowledgment.

TABLE I
STATISTICS FOR THE DATASETS

Stats	Manta rays		Whales	
	Dataset	One fold	Dataset	One fold
Number of images	1730	~350	2908	~550
Number of individuals	120	24	633	126
Average # images per ind.	14		5	
Min # images per ind.	6		3	

and testing purposes we select individuals with a minimum of three images per whale resulting in a set of 2908 images for 633 unique whales, see Table I. Most of the images have already been cropped to include only the image of the fluke (see example images in Fig. 7), however there are some noisy examples where the fluke is shown in the distance or text information appears at the bottom. We did not do any cropping, although this may further improve the results.

The challenges encountered with the whale dataset are different from those of the manta ray dataset. Although there is no large variation in pose or viewpoint, there is a limited number of examples per individual, a combination of black-and-white and colour images, a variety of illumination conditions and some noisy images.

B. Implementation details

1) *Batch generation*: Training is performed on batches of $B \stackrel{\text{def}}{=} P \times K$ images, where P is a number of distinct classes in the batch and K is a number of examples per class. During training, the whole batch is fed into the network and embeddings for the batch are computed. Embeddings are then combined into triplets based on pairwise distances according to the semi-hard triplet selection strategy discussed in Section III-D. We use batches of 15 classes with 5 images per class for manta rays and 3 images per class for whales as this is the maximum batch size that fits into the memory of the computer utilized in these experiments.

2) *Data augmentation*: Data augmentation is used extensively during training to increase the variety in the training set. Transformations are applied on the fly so that at every epoch the network receives a new augmentation of the image. For the manta ray dataset the following geometric transformations

TABLE II
PERFORMANCE OF *Inception-Ft* MODEL ON HUMPBACK WHALES AND MANTA RAYS DATASETS SEPARATELY (METRICS ARE AVERAGED OVER 5 SPLITS)

Metrics	Dataset	
	Humpback whales	Manta rays
Top-1	62.78%±1.64	62.05%±3.24
Top-5	88.20%±0.67	93.65%±1.83
Top-10	93.46%±0.63	97.03%±1.11
TPR	73%	71%
AUC	0.980	0.966

were used: rotation up to 90 degrees, horizontal and vertical flips, small shifts up to 10 pixels and zooming in to 10 percent. Most of the whale images have already a normalized view of the fluke upright. Therefore, only small rotation angles are used in data augmentation for whales.

3) *Base networks*: We compare convolutional layers of InceptionV3 [37] and MobileNetV2 [38] as feature extractors to assess the influence of the CNN architecture on the performance of the system. One of the key differences between these two models is the number of parameters and operations. The smaller MobileNetV2 has 3.4 million parameters and 300 million multiply-adds operations [37]. The bigger InceptionV3 has 23 million parameters and 5 billion multiply-adds per inference [37]. The convolutional layers of both networks have been initialized with weights pretrained on Imagenet [39].

The input size of the network depends on the case study: the input images of whale flukes are resized to 224×448 because of the shape of the region of interest; the input images for manta ray pattern are of shape 300×300 for InceptionV3 and 224×224 for MobileNetV2 (pretrained weights for MobileNetV2 are available only for some input sizes). Images are preprocessed the same way as it was done for the model used for fine-tuning (pixel values are scaled from $[0,255]$ to $[-1,1]$).

4) *Training*: The Adam optimizer [40] is used for all experiments with a learning rate 10^{-5} and other hyperparameters with default values ($\beta_1 = 0.9$, $\beta_2 = 0.999$). We used a learning rate 10^{-5} because higher values did not work well with the pretrained weights (the same has been observed in [30] while training the pretrained network with the triplet loss).

In order to produce an accurate evaluation of the performance of the network, we perform k-fold cross-validation for the first experiment (Section IV-C1). All splits are done with respect to individuals so each individual appears only in training or test split. Each dataset is split in five parts and five rounds of training are completed with four folds allocated for training and one fold for testing.

All experiments have been run on a cluster with two Tesla M40 24GB GPUs and 6 CPUs.

C. Performance evaluation

1) *Fine-tuning InceptionV3 based model*: We fine-tune models with InceptionV3 convolutional layers, a global pooling layer and a fully connected layer with 256 outputs on each

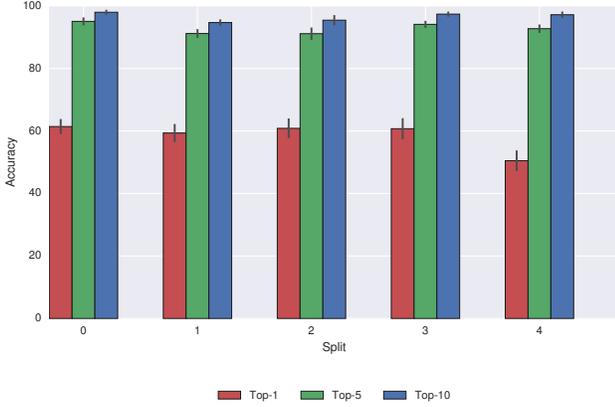


Fig. 8. Top-k accuracy over 5 splits on manta ray patterns with *Inception-Ft* configuration.

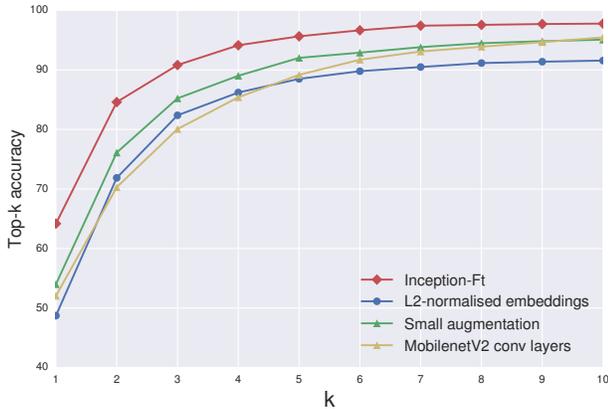


Fig. 9. Top-k accuracy significantly increases at the second prediction for all configurations.

dataset separately. We name this configuration *Inception-Ft* (fine-tuned).

The metrics TPR and AUC are calculated over all possible pairs for the test fold when $FAR = 0.01$ (around 45,000 pairs for manta rays with approximately 2,000 positive pairs depending on the split; around 180,000 pairs with approximately 1,800 positive for the whales dataset). Top-k accuracy is computed for the query set where there are two matching images in the database for each query pattern. We also explore how the accuracy changes depending on the number of matches present in the database in Section IV-C6.

The results of training over five splits on the manta ray dataset show that accuracy does not vary significantly over the splits, see Fig. 8. The top-1 accuracy is 62% for both datasets and the top-10 accuracy is 93% for humpback whales and 97% for manta rays (Table II). Moreover, the graph in Fig. 9 shows that the top-k accuracy increases sharply at the second prediction ($k = 2$) and top-3 accuracy is over 90% for *Inception-Ft* configuration.

We cannot make a meaningful comparison with previous works as the results have been reported on different datasets and the source code is not publicly available. Manta Matcher [6] demonstrates 50.97% top-1 and 67.64% top-10 accuracy

TABLE III
THE LARGER MODEL BASED ON INCEPTIONV3 DEMONSTRATES BETTER PERFORMANCE THAN THE SMALLER MODEL BASED ON MOBILENETV2

Metrics	Base network	
	MobileNetV2	InceptionV3
Top-1	52.06%±4.77	64.18%±4.55
Top-5	89.18%±1.85	95.65%±1.15
Top-10	95.47%±1.40	97.78%±0.62
TPR	60%	73%
AUC	0.970	0.983

TABLE IV
NOT NORMALIZED EMBEDDINGS PERFORMS BETTER THAN l_2 -NORMALIZED

Metrics	Embeddings	
	l_2 -normalized	Not normalized
Top-1	48.72%±4.06	64.18%±4.55
Top-5	88.50%±1.62	95.65%±1.15
Top-10	91.57%±1.75	97.78%±0.62
TPR	61%	73%
AUC	0.959	0.983

on a dataset of 720 images of 265 different manta rays. We think that our dataset is more challenging as it contains images taken from a wider variety of angles and illumination conditions (Fig. 5a). The best results to our knowledge for re-identification of humpback whale flukes have been reported in [17]. The top-1 accuracy of 80% has been achieved on a dataset of a similar size. However, the method is using integral curvature representation of the trailing edge of the flukes and is specifically designed for humpback whales. Our method is generic and not specialized for a particular species.

For the rest of the experiments we change one hyperparameter to evaluate its effect and all other parameters are kept unchanged; the experiments are performed on one split of manta ray dataset.

2) *Influence of the base network*: We evaluate the effect of the model architecture by training two networks with convolutional layers from InceptionV3 and MobileNetV2. The larger model InceptionV3 demonstrates better performance in both validation on pairs (TPR 73% vs 60%) and top-k accuracy

TABLE V
ACCURACY IS NOT SENSITIVE TO THE DIMENSION OF THE EMBEDDING SPACE

Metrics	Embedding length		
	128	256	512
Top-1	64.46%±3.40	64.18%±4.55	65.75%±4.80
Top-5	95.33%±1.08	95.65%±1.15	94.67%±1.61
Top-10	97.76%±0.90	97.78%±0.62	97.47%±0.72
TPR	72%	73%	70%
AUC	0.983	0.983	0.980

TABLE VI
EXTENSIVE AUGMENTATION (ROTATIONS UP TO 360° AND FLIPS) OF
INPUT IMAGES IMPROVE PERFORMANCE COMPARED TO ONLY SMALL
AMOUNT OF AUGMENTATION (ROTATIONS UP TO 10°)

Metrics	Augmentation	
	Small augmentation	Extensive augmentation
Top-1	54.00%±3.32	64.18%±4.55
Top-5	92.03%±1.62	95.65%±1.15
Top-10	95.09%±1.28	97.78%±0.62
TPR	58%	73%
AUC	0.970	0.983

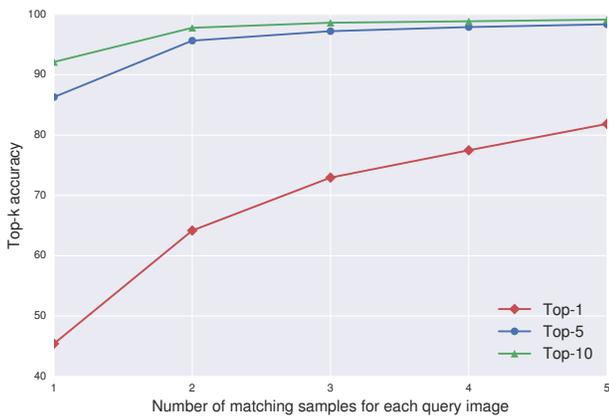


Fig. 10. Accuracy of re-identification depending on the number of matching examples per query individual in the database.

(top-1 accuracy 64% vs 52%), see Table III. However, the difference in performance decreases for higher k and top-10 accuracy is 97% for Inception based and 95% for MobileNet based networks. The advantage of MobileNetV2 is a slightly faster execution but our system does not have to work in real-time. The rest of the experiments are continued with InceptionV3 convolutional layers.

3) *Effect of embedding normalization*: FaceNet architecture [21] uses l_2 -normalization whereas Hermans et al. [30] argue that forcing the norm of the embedding to 1 does not improve performance. Our experiments demonstrate that restricting the embedding space to a hypersphere decreases the accuracy and metrics for verification on pairs. For example, top-1 accuracy drops from 64% to 48% when we apply l_2 -normalization, TPR decreases from 73% to 61% see Table IV. Therefore, the rest of the experiments was done without l_2 -normalization.

4) *Influence of embedding dimension*: We tested three values for the dimension of the embedding space, 128, 256 and 512. Averaged results are reported in Table V. The difference between achieved accuracy is statistically insignificant and we select dimension of 256 for all other experiments. Experiments with smaller embedding spaces (dimensions 32 and 64) showed inferior performance compared to higher dimensional spaces.

5) *Effect of data augmentation*: We have investigated the effect of data augmentation on the manta ray dataset. The pattern on a manta ray belly may appear at different angles so

extensive data augmentation including full rotations and flips has been applied to *Inception-Ft* model. We train the same model with rotations to only 10 degrees and no flips to estimate the influence of data augmentation on the performance.

The experiment shows (Table VI) that the performance results of the tested model are lower when less augmentation is applied during training: top-1 accuracy drops significantly, 54% vs 64%, and TPR has dropped to 58% compared to 73%. This demonstrates that rotations and flips of training examples facilitate learning of pattern invariance to rotations. However, the difference in top-5 and top-10 accuracy is less marked.

6) *Number of matching individuals in the database*: Previous experiments in this paper have been conducted under the condition that there are two matching images for each query individual in the database. This experiment compares accuracy for a different number of matching individuals (from one to five), see Fig. 10. The fewer images in the database for each query individual, the more difficult it is for the network to find the right match. The number of matching examples for an individual in the database is more important for top-1 than for top-5 or top-10 accuracy. Top-1 accuracy is around 45% for only one matching image, it increases to 64% for two matches and reaches 81% when there are five images in the database for each individual. Top-10 accuracy reaches 98% with at least three images per individual in the database which is beneficial for the practical application.

7) *Visualization of predictions*: Fig. 11 shows three query images and top-5 predictions of the system. All predictions share visual similarity with a query image. Three examples of incorrect matches alongside with top-3 predictions and two matching examples from the database are shown in Fig. 12. These examples are challenging as the pattern is only partly visible because of the oblique angle.

We analyze the learned representation with t-SNE [41]. The t-SNE algorithm maps a high dimensional space into a two-dimensional while preserving the similarity between points. The t-SNE plot for the manta ray test set (see Fig. 13) shows examples where embeddings for the same manta ray (manta Telluno, manta Priapus) are clustered together even when the viewpoint is different and small occlusions are present (water bubbles, small fish). Embeddings are less separated for the less distinguishable markings where a pattern consists of a small number of black marks placed sparsely (manta Paw Paw and manta Nova; manta Kimba and manta Cousteau). On the t-SNE plot for the humpback whales test set (see Fig. 14) we observe that individuals are clustered together even when the fluke is visible from different distances (whale 120). The system is invariant to the pose of the fluke (whale 101, whale 21) and viewpoint position (whale 25). The mix between whales occurs for some totally black flukes (whales 136, 12, 72) or for the flukes with a similar colour pattern (whale 61 and whale 21).

V. CONCLUSION

We have presented a novel visual re-identification system for manta rays that is robust to viewpoint changes, variations in lighting and small occlusions. The results have been achieved

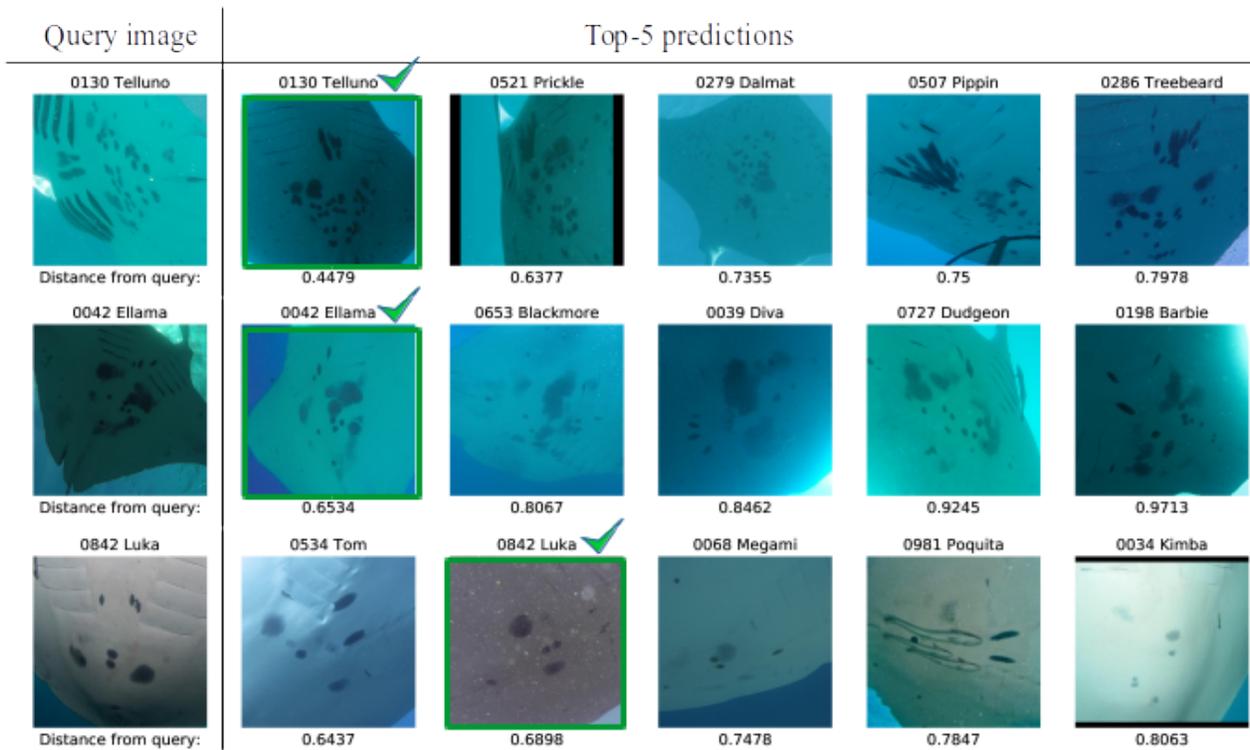


Fig. 11. Three examples of correct predictions. All closest predictions share visual similarity to the query image. The pattern is correctly matched even for examples with a challenging viewpoint and illumination. Photo credits: Fabrice Jaïne, Linda Earthwatch, John Lawson, Chris Garraway, Chris Garraway, Chris Kim, Maggie McNeil, Chris Garraway, Rebecca Fonskov, Kathy Townsend, Chris Dudgeon, Chris Garraway, Sarah Williamson, Ryan Jeffery, Amelia Armstrong, Ryan Jeffery, Josh Gransbury, Chris Garraway.

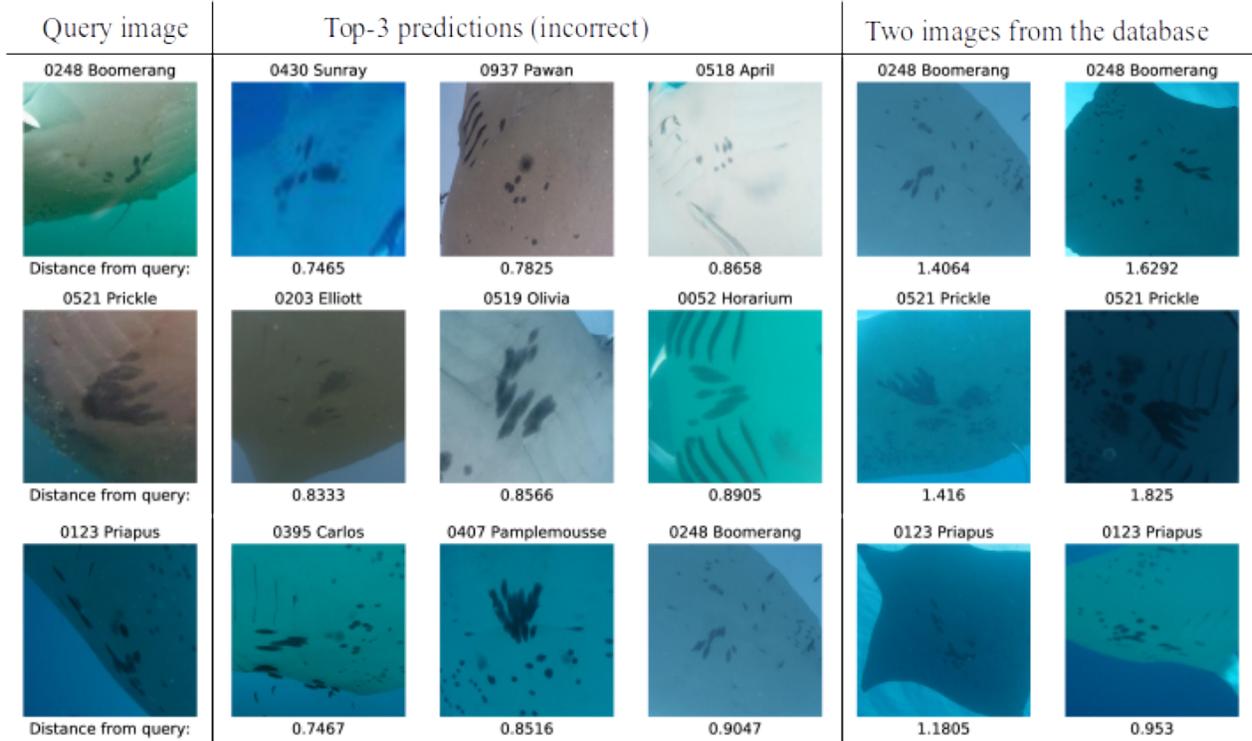


Fig. 12. Three examples of incorrect matching (no match within ten closest predictions). Two images on the right are the only matching examples for each query image in the database. The query images are difficult because of the oblique angle that limits the visibility of the whole pattern. Top-3 predictions share some visual similarity to the query image. Photo credits: John Gransbury, Ian Christie, Amelia Armstrong, Mark Gray, Kathy Townsend, Kathy Townsend, Graeme Haas, Gerard Smith, Amelia Armstrong, Fabrice Jaïne, Michael Rowett, Mounties Earthwatch, Mark Atkinson, Lydie Couturier, Chris Garraway, Kathy Townsend, Chris Garraway, Deg Ed.

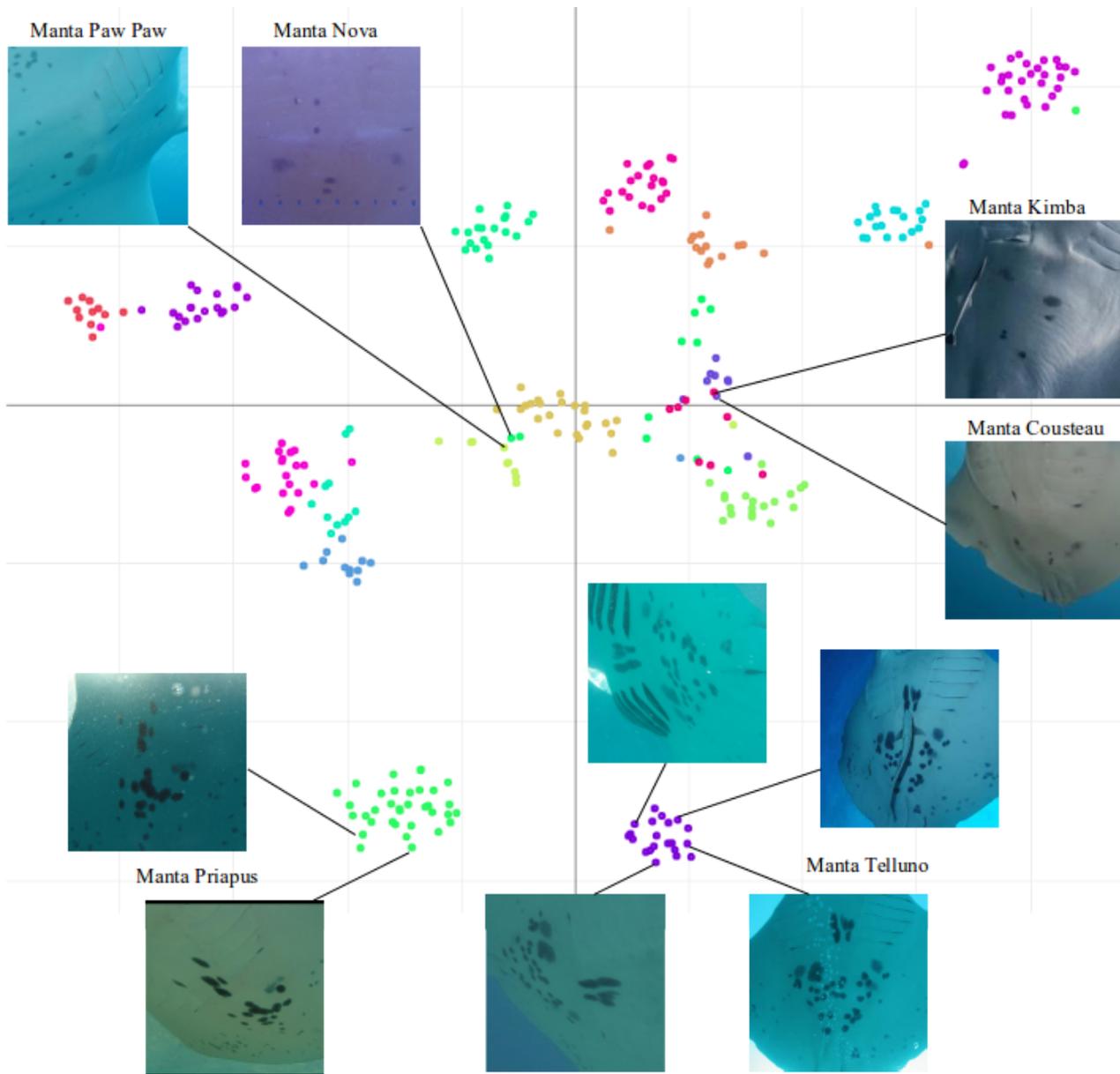


Fig. 13. Visualization of embeddings computed for the manta ray test set (best viewed in colour) using t-SNE [41]. Embeddings for manta Telluno and manta Priapus form tight clusters and show that the learned representation is invariant to rotations, a viewing angle and small occlusions. Mixing between classes happens when the pattern has several sparse dots (manta Paw Paw and manta Nova; manta Kimba and manta Cousteau). Photo credits (in a clockwise order starting from manta Kimba: Steward Barry, Mark Gray, Fabrice Jaïne, Nigel Marsh, Kathy Townsend, Lydie Couturier, Chris Garraway, Chris Garraway, Chris Garraway, Matt Prunty.

by using a combination of InceptionV3 model, the semi-hard triplet mining strategy, the triplet loss function and an extensive geometric augmentation of the input images. The practical value of the system been demonstrated on a manta ray dataset and an humpback whale dataset. The system requires the user to localize the region of interest by drawing a bounding box around it.

In the future, we plan to further improve the system by automating the localization of the patterns of interest. One possible strategy is to train the network on auxiliary tasks like learning to predict the locations of specific body landmarks (tip of the wings and gills of manta rays, fluke tips and notch for whales). This would force the network to learn about the

morphology of the animal. This ability should help induce a better representation of the spatial position of the pattern with respect to the body.

ACKNOWLEDGMENT

The authors would like to thank Project Manta (<https://sites.google.com/site/projectmantisite/home>) and Happywhale organization (happywhale.com) for the datasets of images. Photo credits for published images of humpback whale flukes: Alethea Leddy, Barry Gutradt, Casey Clark, Channel Islands NMS Naturalist Corps, Colin Garland, Dale Frink, Fernando Arcas, JB, John Calambokidis, Kate Cummings,

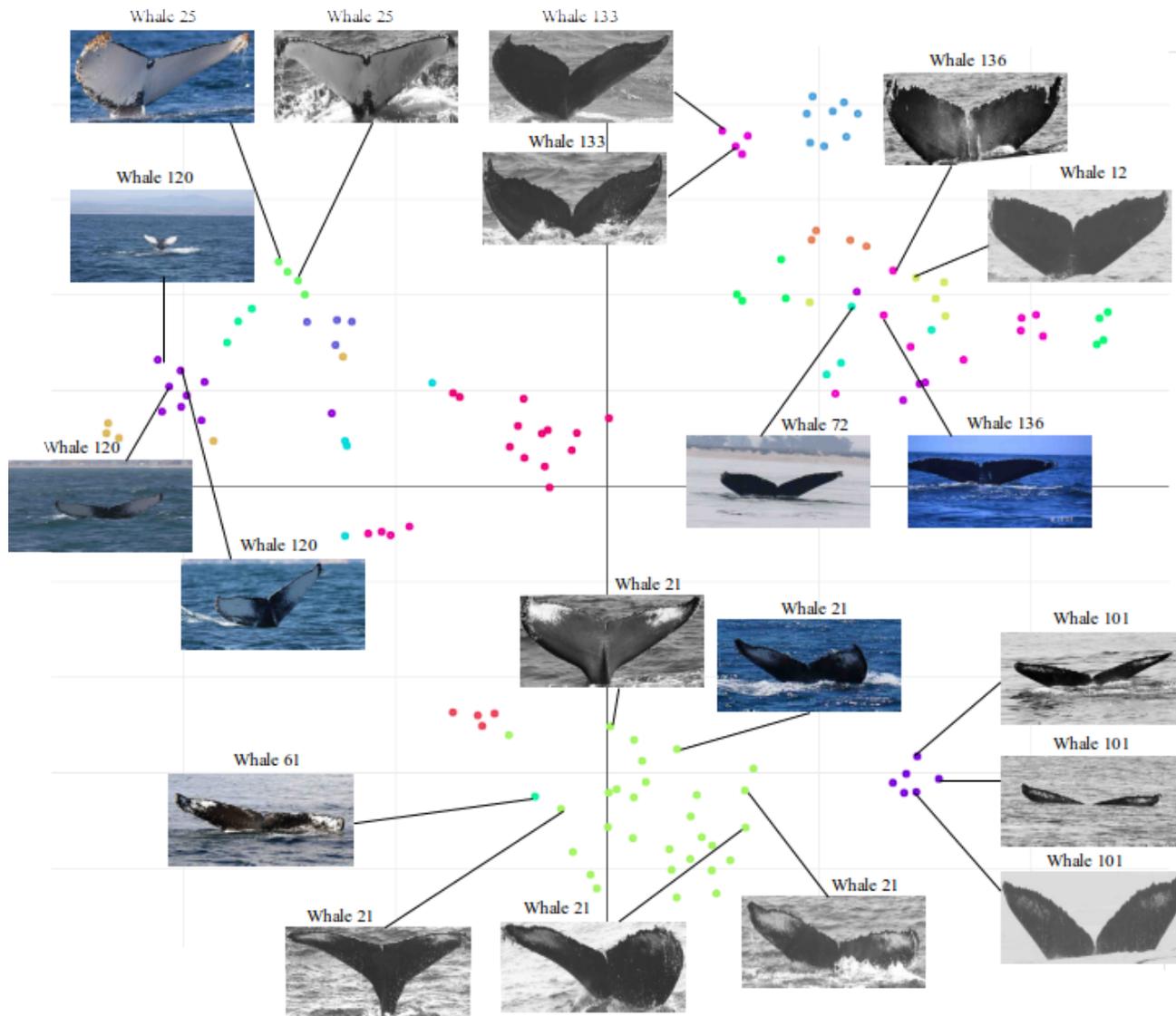


Fig. 14. Visualization of embeddings computed for the test set for humpback whale flukes (best viewed in colour) using t-SNE [41]. Photo credit: Happywhale organization. List of photo credits is provided in acknowledgment.

Kate Spencer, Mark Girardeau, Richard Jackson, Ryan Lawler, Traci Phillips.

Computational resources and services used in this work were provided by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia.

REFERENCES

- [1] A. Marshall and S. Pierce, "The use and abuse of photographic identification in sharks and rays," vol. 80, pp. 1361–79, Apr. 2012.
- [2] J. D. Stewart, F. R. A. Jaime, A. J. Armstrong, A. O. Armstrong, M. B. Bennett, K. B. Burgess, L. I. E. Couturier, D. A. Croll, M. R. Cronin, M. H. Deakos, C. L. Dudgeon, D. Fernando, N. Froman, E. S. Germanov, M. A. Hall, S. Hinojosa-Alvarez, J. E. Hosegood, T. Kashiwagi, B. J. L. Laglbauer, N. Lezama-Ochoa, A. D. Marshall, F. McGregor, G. Notarbartolo di Sciara, M. D. Palacios, L. R. Peel, A. J. Richardson, R. D. Rubin, K. A. Townsend, S. K. Venables, and G. M. W. Stevens, "Research priorities to support effective manta and devil ray conservation," *Frontiers in Marine Science*, vol. 5, no. 314, pp. 1–27, Sept. 2018. [Online]. Available: <https://doi.org/10.3389/fmars.2018.00314>
- [3] D. A. Croll, H. Dewar, N. K. Dulvy, D. Fernando, M. P. Francis, F. Galvan-Magana, M. Hall, S. Heinrichs, A. Marshall, D. Mccauley, K. M. Newton, G. Notarbartolo-Di-Sciara, M. O'Malley, J. O'Sullivan, M. Poortvliet, M. Roman, G. Stevens, B. R. Tershy, and W. T. White, "Vulnerabilities and fisheries impacts: the uncertain future of manta and devil rays," *Aquatic Conservation: Marine and Freshwater Ecosystems*, vol. 26, no. 3, pp. 562–575, Nov. 2015. [Online]. Available: <https://doi.org/10.1002/aqc.2591>
- [4] C. A. Rohner, S. J. Pierce, A. D. Marshall, S. J. Weeks, M. B. Bennett, and A. J. Richardson, "Trends in sightings and environmental influences on a coastal aggregation of manta rays and whale sharks," *Marine Ecology Progress Series*, vol. 482, pp. 153–168, May 2013.
- [5] A. Zerbini, E. J. Ward, P. Kinan, M. H. Engel, and A. Andriolo, "A bayesian assessment of the conservation status of humpback whales (megaptera novaeangliae) in the western atlantic ocean (breeding stock a)," *Journal of Cetacean Research and Management*, vol. Special Issue, no. 3, pp. 131–144, Dec. 2011.
- [6] C. Town, A. Marshall, and N. Sethasathien, "Manta matcher: automated photographic identification of manta rays using keypoint features," *Ecology and evolution*, vol. 3, no. 7, pp. 1902–1914, 2013.
- [7] O. Moya, P.-L. Mansilla, S. Madrazo, J.-M. Igual, A. Rotger, A. Romano, and G. Tavecchia, "Aphis: A new software for photo-matching in ecological studies," *Ecological Informatics*, vol. 27, pp. 64–70, May 2015. [Online]. Available: <https://doi.org/10.1016/j.ecoinf.2015.03.003>

- [8] J. Diaz-Calafat, E. Ribas-Marques, S. Jaume-Ramis, S. Martinez-Nunez, A. Sharapova, and S. Pinya, "Individual unique colour patterns of the pronotum of *rhyrachophorus ferrugineus* (coleoptera: Curculionidae) allow for photographic identification methods (pim)," *Journal of Asia-Pacific Entomology*, vol. 21, no. 2, pp. 519–526, 2018. [Online]. Available: <https://doi.org/10.1016/j.aspen.2018.03.002>
- [9] A. M. Van Tienhoven, J. E. Den Hartog, R. A. Reijns, and V. M. Peddemors, "A computer-aided program for pattern-matching of natural marks on the spotted raggedtooth shark *carcharias taurus*," *Journal of Applied Ecology*, vol. 44, no. 2, pp. 273–280, Feb. 2007. [Online]. Available: <https://doi.org/10.1111/j.1365-2664.2006.01273.x>
- [10] C. R. Gatto, A. Rotger, N. J. Robinson, and P. S. Tomillo, "A novel method for photo-identification of sea turtles using scale patterns on the front flippers," *Journal of Experimental Marine Biology and Ecology*, vol. 506, pp. 18–24, 2018. [Online]. Available: <https://doi.org/10.1016/j.jembe.2018.05.007>
- [11] M. F. Hansen, M. L. Smith, L. N. Smith, M. G. Salter, E. M. Baxter, M. Farish, and B. Grieve, "Towards on-farm pig face recognition using convolutional neural networks," *Computers in Industry*, vol. 98, pp. 145–152, 2018. [Online]. Available: <https://doi.org/10.1016/j.compind.2018.02.016>
- [12] E. Nepovinskyh, T. Eerola, H. Kälviäinen, and G. Radchenko, "Identification of saimaa ringed seal individuals using transfer learning," in *Proc. ACIVS*, Poitiers, France, 2018, pp. 211–222.
- [13] N. Tariq, K. Saleem, M. Mushtaq, and M. A. Nawaz, "Snow leopard recognition using deep convolution neural network," in *Proc. ICISDM*, Lakeland, FL, USA, 2018, pp. 29–33.
- [14] R. Bogucki, M. Cygan, C. B. Khan, M. Klimek, J. K. Milczek, and M. Mucha, "Applying deep learning to right whale photo identification," *Conservation Biology*, Sept. 2018.
- [15] E. Ranguelova, M. Huiskes, and E. J. Pauwels, "Towards computer-assisted photo-identification of humpback whales," in *Proc. ICIP*, Singapore, 2004, pp. 1727–1730.
- [16] A. Gilman, K. Hupman, K. A. Stockin, and M. D. M. Pawley, "Computer-assisted recognition of dolphin individuals using dorsal fin pigmentations," in *Proc. IVCNZ*, Palmerston North, New Zealand, 2016, pp. 1–6.
- [17] H. J. Weideman, Z. M. Jablons, J. Holmberg, K. Flynn, J. Calambokidis, R. B. Tyson, J. B. Allen, R. S. Wells, K. Hupman, K. Urian, and C. V. Stewart, "Integral curvature representation and matching algorithms for identification of dolphins and whales," in *Proc. ICCV*, Venice, Italy, 2017, pp. 2831–2839.
- [18] D. T. Bolger, T. A. Morrison, B. Vance, D. Lee, and H. Farid, "A computer-assisted system for photographic mark-recapture analysis," *Methods in Ecology and Evolution*, vol. 3, no. 5, pp. 813–822, 2012.
- [19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- [20] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition," in *Proc. BMVC*, Swansea, UK, 2015, pp. 41.1–41.12.
- [21] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 815–823.
- [22] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, Montreal, Canada, 2014, pp. 1988–1996.
- [23] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, San Diego, CA, USA, 2005, pp. 539–546.
- [24] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 1386–1393.
- [25] O. Moskvyyak and F. Maire, "Learning geometric equivalence between patterns using embedding neural networks," in *Proc. DICTA*, Sydney, NSW, Australia, 2017, pp. 778–785.
- [26] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Proc. AAAI*, San Francisco, CA, USA, 2017, pp. 3988–3994.
- [27] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," in *Proc. BigMM*, Xi an, China, 2018.
- [28] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proc. ICCV*, Venice, Italy, 2017, pp. 2612–2620.
- [29] O. Rippel, M. Paluri, P. Dollár, and L. D. Bourdev, "Metric learning with adaptive density discrimination," *arXiv preprint arXiv:1511.05939*, 2015. [Online]. Available: <https://arxiv.org/pdf/1511.05939.pdf>
- [30] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017. [Online]. Available: <https://arxiv.org/pdf/1703.07737.pdf>
- [31] C. Wang, X. Zhang, and X. Lan, "How to train triplet networks with 100k identities?" in *Proc. ICCVW*, Venice, Italy, 2017, pp. 1907–1915.
- [32] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013. [Online]. Available: <http://arxiv.org/abs/1312.4400.pdf>
- [33] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. CVPR*, New York City, NY, USA, 2006, pp. 1735–1742.
- [34] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. ICCV*, Venice, Italy, 2017, pp. 2859–2867.
- [35] K. S. Ted Cheeseman, Tory Johnson and N. Muldavin, "Happywhale: Globalizing marine mammal photo identification via a citizen science web platform," Happywhale, Santa Cruz, CA, USA, Rep. SC/67b/PH/02, 2017.
- [36] T. Cheeseman and K. Southerland, "Happywhale progress report 2017-2018," Happywhale, Santa Cruz, CA, USA, Rep. SC/67b/PH/05, 2018.
- [37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, Las Vegas Valley, NV, United States, 2016, pp. 2818–2826.
- [38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, Salt Lake City, UT, USA, 2018.
- [39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015. [Online]. Available: <https://arxiv.org/pdf/1412.6980.pdf>
- [41] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.