A Whale's Tail - Finding the Right Whale in an Uncertain World

Diego Marcos[1](✉), Jana Kierdorf[2], Ted Cheeseman[3], Devis Tuia[4], and Ribana Roscher[2,5]

[1] Wageningen University, Wageningen, The Netherlands
diego.marcos@wur.nl

[2] Institute of Geodesy and Geoinformation, University of Bonn, Bonn, Germany
{jkierdorf,ribana.roscher}@uni-bonn.de

[3] Happywhale, St Albans, UK
ted@happywhale.com

[4] Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
devis.tuia@epfl.ch

[5] Data Science in Earth Observation, Technical University of Munich,
Ottobrunn, Germany

Abstract. Explainable machine learning and uncertainty quantification have emerged as promising approaches to check the suitability and understand the decision process of a data-driven model, to learn new insights from data, but also to get more information about the quality of a specific observation. In particular, heatmapping techniques that indicate the sensitivity of image regions are routinely used in image analysis and interpretation. In this paper, we consider a landmark-based approach to generate heatmaps that help derive sensitivity and uncertainty information for an application in marine science to support the monitoring of whales. Single whale identification is important to monitor the migration of whales, to avoid double counting of individuals and to reach more accurate population estimates. Here, we specifically explore the use of fluke landmarks learned as attention maps for local feature extraction and without other supervision than the whale IDs. These individual fluke landmarks are then used jointly to predict the whale ID. With this model, we use several techniques to estimate the sensitivity and uncertainty as a function of the consensus level and stability of localisation among the landmarks. For our experiments, we use images of humpback whale flukes provided by the Kaggle Challenge "Humpback Whale Identification" and compare our results to those of a whale expert.

Keywords: Attention maps · Sensitivity · Uncertainty · Whale identification

The work is partly funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant number: 01DD20001).

A. Holzinger et al. (Eds.): xxAI 2020, LNAI 13200, pp. 297–313, 2022.
https://doi.org/10.1007/978-3-031-04083-2_15

1 Introduction

For many scientific disciplines, reliability and trust in a machine learning result are of great importance, in addition to the prediction itself. Two key values that can contribute significantly to this are the interpretability and the estimation of uncertainty:

– An interpretation aims at the presentation of properties of a machine learning model (e.g., a decision process of a neural network) in a way that it is understandable to a human [21]. One possibility to obtain an interpretation is sensitivity analysis which provides information about how the models' output is affected by small or specifically chosen changes in the input [18].
– Uncertainty is the quantity of all possible changes in the output that result from uncertainties already included in the data (aleatoric/data uncertainty) or a lack of knowledge of the machine learning model (epistemic/model uncertainty) [6].

Both uncertainty quantification and sensitivity analysis have become a broad field of research in recent years, especially for developing methods to check the suitability and to better understand the decision-making process of a data-driven model [6,21,24]. However, so far, the two areas have usually been considered separately, although a joint consideration has clear benefits, since the analysis of sensitivity can often be considered as a part or first step towards uncertainty quantification.

In this chapter, we will consider a use case from marine science to demonstrate the usefulness of a joint use of sensitivity and uncertainty quantification in landmark-based identification. In particular, we look at the identification of whales by means of images of their fluke. Whale populations worldwide are threatened by commercial whaling, global warming, and the struggle for food in competition with the fishing industry [33]. A protection of whales is essentially supported by the reconstruction of the spatio-temporal migration of whales, which in turn is based on the (re)identification of whales. Individual whales can be identified by the shape of their whale flukes and their unique pigmentation [13]. Three features in particular play a crucial role for whale experts in distinguishing between individual whales (see Fig. 1):

– *Pigmentation-based features.* These features correspond to coloured patches on the fluke, forming unique patterns. They are very clearly visible to the human eye. They can change significantly within the first few years of whale life and in extremely cold water (for example, Antarctica, but also Greenland and the North Atlantic). They may be partially obscured by heavy diatom growth, characterized by a yellow-orange appearance of the fluke.
– *Fluke shape.* This feature is reliable and robust. The outer 20% of the tail may become more distorted and change over time, but the inner 80% and V-notch are reliable and stable. Although it is difficult to detect by the human eye, it has proven to be very useful for machine learning-based approaches [14,15,25].

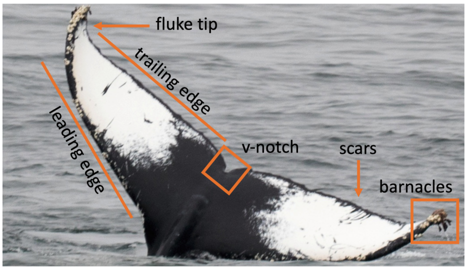

Fig. 1. Important characteristics of a whale fluke.

- *Scars.* The surface of the fluke usually shows contrasting scars. However, the contrast can vary greatly and the scars may change over time. Certain scars grow with the whale, such as killer whale rake marks that form parallel lines or barnacle marks that form circles. In addition, lighting conditions can significantly affect the detectability of scars.

For whale monitoring, whale researchers often use geo-tagged photos with time and location information to reconstruct activities. Since manual analysis is too costly and thus a huge amount of data remained unused, current approaches focus on machine learning [14,15,25].

Despite the accuracy observed in recent competitions [29], limited effort has been devoted to actually quantify sensitivity in the prediction and identify sources of uncertainty. We argue that uncertainty identification remains a central topic requiring attention and propose a methodology based on landmarks and their spatial sensitivity and uncertainty to answer a number of scientific questions useful for experts in animal conservation. Specifically, we tackle the following questions:

- Which parts of the fluke are more consistently useful to identify whales? A whale fluke changes with time and therefore, characteristic features of a fluke may no longer be present and therefore not visualized in the interpretation tool results.
- Can landmarks together with uncertainty and sensitivity indicate the suitability of images for identification? Suitability is influenced, for example, by image quality, position, and size of the object, but also by the presence of relevant features.

These goals are formulated from the perspective of whale research, but are also intended to raise relevant questions from the perspective of machine learning, such as the usefulness of interpretation tools to improve models. In general, the task of re-identifying objects or living beings from images and is a common topic [2,16,26], and the approach and insights presented in this paper can also be applied to similar tasks from other fields.

2 Related Work

Self-explainable Deep Learning Models. Although the vast majority of methods to improve the interpretability and explainability of deep learning models are designed to work *post-hoc* [19,28,32], *i.e.* the important parts of the input are highlighted while the model itself remains unmodified, a few approaches aim at modifying the model so that its inherent interpretability is enhanced, also referred to as self-explainable models [23]. This has the advantage that the interpretation is actually part of the inference process, rather than being computed *a posteriori* by an auxiliary interpretation method, resolving potential trustworthiness issues of *post-hoc* methods [22]. The visual interpretation can be obtained, for example, by incorporating a global average pooling after the last convolutional layer of the model [39] or by levering a spatial attention mechanism [36]. Our self-explainable method is inspired by [36] and [38], and learns a fixed set of landmarks, along with their associated attention maps, in a weakly supervised setting by only using class labels. To gain further insight, the landmarks can be used for sensitivity analysis and uncertainty quantification.

Uncertainty Quantification. The field of uncertainty quantification has gained new popularity in recent years, especially for determining the uncertainty of complex models such as neural networks. In most applications, the predictive uncertainty is of interest, i.e. the uncertainty that affects the estimation from various sources of uncertainty, originating from the data itself (aleatoric uncertainty) and arising from the model (model uncertainty). These sources are often not negligible, especially in real-wold applications, and must be determined for a comprehensive statement about the reliability and accuracy of the result. Several works have been carried out such as [5,30], which explore Monte Carlo dropout or quantify uncertainty analysing the softmax output of neural networks. [7,12,34] give comprehensive overviews of the field, where [6] specifically focuses on the applicability in real-world scenarios.

Sensitivity Analysis. This kind of analysis is usually considered in the context of explainable machine learning. Here, a set input variables, such as pixel values in an image region or a unit in some of the model's intermediate representations [3,31], are perturbed, and the effect of such changes on the result is considered. This approach helps to understand the decision process and causes of uncertainties, and to gain insights into salient features that can be spatial, temporal or spectral. According to [21], sensitivity analysis approaches belong to interpretation tools, as they transform complex aspects such as model behavior into concepts understandable by a human [19,24]. Many approaches use heatmaps that visualize the sensitivity of the output to perturbations of the input, the attention map of the classifier model, or the importance of the features [11]. These tools are extremely helpful and have been used recently to infer new scientific knowledge and discoveries and to improve the model [21,27,31]. Probably the best known principle is study of the effects of masking selected regions of the input, which is systematically applied in occlusion sensitivity maps [20]. For more details, including specific types of interpretation and further implementation, we refer to recent studies [1,8,9].

Sensitivity vs. Uncertainty. There are significant differences between the analysis of uncertainties and sensitivity, and previous applications mostly consider only one of the two. Sensitivity analysis focuses more on the input and the effect of modifications on the predictions, while uncertainty quantification focuses on the propagation of uncertainties in the model. Nevertheless, there are also strong correlations, as shown in [18]. Sensitivity analysis, for example, explores the causes and importance of specific uncertainties in the input data for the decision, while uncertainty analysis describes the whole set of possible outcomes. Both consider variations in the input and their influence on the output to derive statements for decision-making. Our work is based on the preliminary work of [14], in which occlusion sensitivity maps are created by systematically covering individual areas in images of whale flukes in order to identify the characteristic features of flukes for whale identification. Here, we propose to learn a set of compact attention maps such that each specializes in the detection of a fluke landmark. These learned landmarks are use to extend [14] by a combined analysis of the sensitivity of the classification to each landmark and their uncertainty.

3 Humpback Whale Data

3.1 Image Data

In this work, we use a set of humpback whale images from the Kaggle Challenge "Humpback Whale Identification". More specifically, we process their tails, called flukes (see Fig. 1). The data set consists of more than 67.000 images, in which 10.008 different whale individuals, i.e., 10.008 different classes, are represented. We pruned the dataset and used only the 1.646 classes that contained three or more images in the training set of the challenge. For our experiments, we restrict ourselves to use images in the training set because the test set does not provide reference information, as it is generally the case for Kaggle challenges. We split the images into a training set $\mathcal{X}_{\text{train}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ (9.408 images) and a test set $\mathcal{X}_{\text{test}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ (1.646 images, or one per class, i.e. a specific whale individual). The number of images per set is given by N and T, respectively. The set $\mathcal{X}_c = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R\}$ describes a subset that includes R images for one specific class c.

3.2 Expert Annotations

A domain expert participated to the study and provided human annotation of remarkable features helping in the discrimination of the whale individuals. For each annotation the expert was provided with a pair of images and asked to mark a set of features helping in discriminating whether the images were of the same individual or not. Three features are generally used by the expert (personal communication), who therefore provided three features per image analysed. Some examples are shown in Fig. 5a.

4 Methods

4.1 Landmark-Based Identification Framework

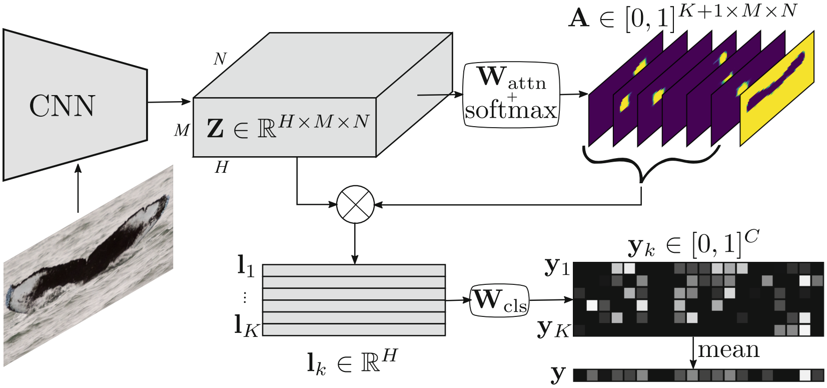
$\mathbf{A} \in [0,1]^{K+1\times M\times N}$
CNN
N
M
$\mathbf{Z} \in \mathbb{R}^{H\times M\times N}$
H
$\mathbf{W}_{\text{attn}}$ + softmax
$\mathbf{l}_1$
$\mathbf{l}_K$
$\mathbf{l}_k \in \mathbb{R}^H$
$\mathbf{W}_{\text{cls}}$
$\mathbf{y}_k \in [0,1]^C$
$\mathbf{y}_1$
$\mathbf{y}_K$
mean
$\mathbf{y}$

Fig. 2. Given the image of a fluke, we extract the feature tensor **Z** using a CNN. A set of compact attention maps **A**, excluding a background map, is then used to extract localized features from **Z**. These features are then averaged and used for classification into C classes, each corresponding to an individual whale.

We propose to learn a set of discriminant landmarks for whale identification such that the model uses evidence from each one separately in order to solve the task. The rationale behind this approach is twofold:

1. Each landmark will gather evidence from a different region of the image, effectively resulting in an ensemble of diverse classifiers, each using a different subset of the data. This independence between the different classifiers provides an improved uncertainty estimation.
2. Since landmarks are trained to attend to a small region of the image, it becomes very easy to visualize where the evidence is coming from with no further computation, thus inherently providing an enhanced level of interpretability.

In order to learn to detect informative landmarks without further supervision than the whale ID, we use an approach inspired by [38]. Likewise, we aim at learning to detect a fixed set of keypoints in the image to establish at which locations landmarks are to be extracted. Unlike [38], we do not use an hourglass-type architecture, but a standard classification CNN with a reduced downsampling rate in order to allow for a better spatial resolution. Another major difference is that we do not use any reconstruction loss and therefore need no decoding elements.

Given an image $\mathbf{X} \in \mathbb{R}^{3 \times MD \times ND}$ and a CNN with a downsampling factor D, the H-channel tensor resulting from applying the CNN to $\mathbf{X}$ is:

$$\mathbf{Z} = \mathrm{CNN}(\mathbf{X}; \theta) \in \mathbb{R}^{H \times M \times N}. \tag{1}$$

We obtain the $K + 1$ attention maps, representing the K keypoints and the background, by applying a linear layer to each location of $\mathbf{Z}$, which is equivalent to a 1×1 convolutional filter parametrized by the weight matrix $\mathbf{W}_{\mathrm{attn}} \in \mathbb{R}^{H \times (K+1)}$, followed by a channel-wise softmax:

$$\mathbf{A} = \mathrm{softmax}(\mathbf{Z} * \mathbf{W}_{\mathrm{attn}}) \in \mathbb{R}^{(K+1) \times M \times N}. \tag{2}$$

Each attention map $\mathbf{A}_k$, except for the $(K+1)^{\mathrm{th}}$, which captures the background, is applied to the tensor $\mathbf{Z}$ in order to obtain the corresponding landmark vector:

$$\mathbf{l}_k = \sum_{u=1}^{M} \sum_{v=1}^{N} \mathbf{A}_k(u,v)\mathbf{Z}(u,v) \in \mathbb{R}^{H}. \tag{3}$$

Each landmark $\mathbf{l}_k$ undergoes a linear operation in order to generate the C classification scores, where C is the total number of classes, associated to it:

$$\mathbf{y}_k = \mathbf{l}_k \mathbf{W}_{\mathrm{class}} \in \mathbb{R}^{C}. \tag{4}$$

We apply different losses to the classification scores $\mathbf{y}$, the landmark feature vectors $\mathbf{l}$ and the attention maps $\mathbf{A}$. For the classification scores, we use a cross-entropy loss, providing the only gradients for learning the weights of the linear operator $\mathbf{W}_{\mathrm{class}} \in \mathbb{R}^{H \times C}$:

$$\mathcal{L}_{\mathrm{class}}(\mathbf{y}, c) = -\log\left(\frac{\exp(y(c))}{\exp(\sum_i y(i))}\right) \tag{5}$$

In addition, we make sure that landmark vectors are similar across images of the same individual. We use a triplet loss for each landmark k, which is computed on the landmark vector $\mathbf{l}_k^a$, used as anchor in the triplet loss, a positive vector from the corresponding landmark stemming from an image of the same class, $\mathbf{l}_k^p$, and a negative one from a different class $\mathbf{l}_k^n$:

$$\mathcal{L}_{\mathrm{triplet}}(\mathbf{l}_k^a, \mathbf{l}_k^p, \mathbf{l}_k^n) = \max(\|\mathbf{l}_k^a - \mathbf{l}_k^p\|_2 - \|\mathbf{l}_k^a - \mathbf{l}_k^n\|_2 + 1, 0) \tag{6}$$

Regarding the losses applied to the landmark attention maps, which have the role of ensuring learning a good set of keypoints for landmark extraction, we apply two losses:

$$\mathcal{L}_{\mathrm{conc}}(\mathbf{A}) = \frac{\sum_{k=1}^{K} \sigma_u^2(\mathbf{A_k}) + \sigma_v^2(\mathbf{A_k})}{K}, \tag{7}$$

which aims at encouraging each attention map to be concentrated around its center of mass by minimizing the variances of each attention map, $\sigma_u^2(\mathbf{A_k})$ and $\sigma_v^2(\mathbf{A_k})$, across both spatial dimensions and

$$\mathcal{L}_{\max}(\mathbf{A}) = \frac{\sum_{k=1}^{K} 1 - \max(\mathbf{A_k})}{K}, \tag{8}$$

which ensures that all landmarks are present in each image.

These four losses are combined as a weighted sum to obtain the final loss:

$$\mathcal{L} = \lambda_{\text{class}}\mathcal{L}_{\text{class}} + \lambda_{\text{triplet}}\mathcal{L}_{\text{triplet}} + \lambda_{\text{conc}}\mathcal{L}_{\text{conc}} + \lambda_{\max}\mathcal{L}_{\max}, \tag{9}$$

where λ_{class}, λ_{triplet}, λ_{conc} are scalar hyperparameters.

4.2 Uncertainty and Sensitivity Analysis

Patch-Based Occlusion Sensitivity Maps. Determining occlusion sensitivity maps is a strategy developed by [37] to evaluate the sensitivity of a trained model to partial occlusions in an input image. The maps visualize which regions contribute positively and which contribute negatively to the result. The approach is to systematically mask different regions for a given input image, choosing a rectangular patch in our case. Two parameters, namely patch size p and step size, are chosen by the user, and the choice affects the result in terms of precision and smoothness. In the area around position $\mathbf{u}$ occluded by the patch, the pixel-wise results of the classifier for each class are compared with the results obtained after part of the image was occluded. For the expected class c, the score $\boldsymbol{s}$ is predicted for the corresponding position u of the patch. The difference $\delta \boldsymbol{s}_{cu}$ is given by.

$$\delta \boldsymbol{s}_{cu} = \boldsymbol{s}_c \text{ - } \tilde{\boldsymbol{s}}_{cu} \tag{10}$$

where the original predicted score for each class is denoted by $\boldsymbol{s}_c$ and the predicted score based on occlusion is given by $\tilde{\boldsymbol{s}}_{cu}$. Performing this for the entire image yields a heat map of occlusion sensitivity.

Landmark-Based Sensitivity Analysis. Similarly to the patch-based occlusion sensitivity maps presented previously, landmark-based sensitivity analysis eliminates individual landmarks, by setting all the elements in the corresponding feature vector $\mathbf{l}_k$ to zero, in order to analyze their effect on the output, allowing to understand the impact that each landmark has on the final score. In addition to this, we also measure the impact that removing a landmark has on the accuracy across the validation set. In both cases, the same landmark k is removed for all images in the test, thus preventing it from contributing to the final score. This allows us to probe the importance of each landmark across the whole test set.

Landmark-Based Uncertainty Analysis. Due to occlusions, unreliable fluke features or wrongly placed landmarks, different groups of landmarks in the same image may provide evidence for conflicting outputs. Similarly, each individual landmark detector may receive conflicting signals from the previous layer about where to place the landmark on the image. This disagreement can be used to In order to measure this disagreement, we perform two experiments applying different types of Monte Carlo dropout (i.e. test time dropout) to the landmarks.

Class Uncertainty Through Whole Landmark Dropout. We randomly choose half of the landmarks and use them to obtain a class prediction y_r. We perform this operation R times to obtain a collection of class predictions $\mathbf{R} = \{y_1, \ldots, y_R\}$. The agreement score a is then computed as the proportion of random draws that output the most frequently predicted class:

$$a = \frac{1}{R} \sum_{r=1}^{R} [y_r = \mathrm{mode}(\mathbf{R})]. \tag{11}$$

Landmark Spatial Uncertainty Through Feature Dropout. In this case we apply standard dropout to the feature tensor $\mathbf{Z}$, thus perturbing the landmark attention maps $\mathbf{A}$. Landmarks that have not been reliably detected will be more sensitive to these perturbations, resulting in higher spatial uncertainty.

5 Experiments and Results

Our experiments address landmark detection focusing on the uncertainty and sensitivity of landmarks, and compare to previous results from patch-based occlusion sensitivity maps from [14] by means of whale identification. Furthermore, the landmarks and occlusion sensitivity maps are compared to the domain knowledge of an expert.

Our method allows to easily reach conclusions at both the dataset level and the image level. For one particular image, due to the spatial compactness of the landmark attention maps, we can visualize the contribution of each landmark to the final classification score. In addition, the fact that each landmark tends to focus on the same fluke features across images allows us to analyze the importance of each landmark at the dataset level.

5.1 Experimental Setup

We use a modified classification CNN, a ResNet-18 [10], with reduced downsampling, by a factor of four, in order to preserve better spatial details. For the final loss we used the same weight for each of the sub-losses $\lambda_{\text{triplet}} = \lambda_{\text{conc}} = \lambda_{\text{max}} = \lambda_{\text{class}} = 1$. We use Adam as an optimizer, with the ResNet-18 model starting with a learning rate of 10^{-4}, while $\mathbf{W}_{\text{attn}}$ and $\mathbf{W}_{\text{class}}$ are optimized starting with a learning rate of 10^{-2}. After every epoch, the learning rates are divided by 2 if the validation accuracy decreases. No image pre-processing is used. The top-1 accuracy reaches 86% on the held-out validation set. For comparison, we trained the same base model without the attention mechanism, obtaining an accuracy of 82%, showing that the landmark-based attention mechanism does not penalize the model's performance.

For comparison, we use our previously computed occlusion sensitivity maps presented in [14], which were based on the data and scores of the classification framework of the second winner solution[1] of the Kaggle Challenge. For preprocessing, the framework applies two steps to the raw image. First, the chosen framework automatically performs image cropping in order to reduce the image content to the fluke of the whale. The cropped images are resized to an uniform size of 256 px × 512 px. In the second step, the framework performs standard-normalization on the input images. The architecture is based on ResNet-101 [10] utilizing triplet loss [35], ArcFace loss [4], and focal loss [17]. With this model, we reach a top-5 accuracy of 94.2%.

5.2 Uncertainty and Sensitivity Analysis of the Landmarks

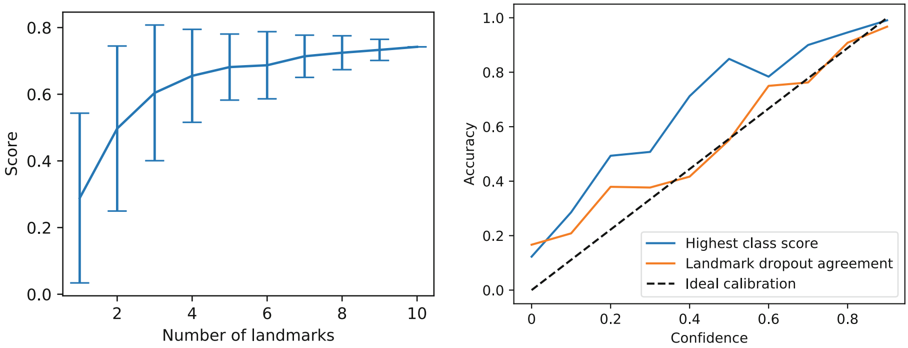

Fig. 3. Left: Average score and standard deviation by randomly selecting an increasing number of landmarks. **Right:** Expected accuracy as a function of two different confidence scores: the highest class score after softmax, and the agreement between 100 landmark dropout runs.

Figure 3 (left) shows the uncertainty of the predicted score, *i.e.* how much the result score varies when a certain number of landmarks is used. It can be seen that the uncertainty becomes smaller the more landmarks are used. The reason for this is that usually several features are used for identification - by the domain expert as well as by the neural network - and with increasing number of landmarks the possibility to cover several features increases. Figure 3 (right) displays the expected accuracy for varying levels of confidence estimates. We compare two estimates: the maximum softmax loss, in blue, and the agreement between 100 runs of MC landmark dropout with a dropout rate of 0.5, in orange. We can see that the latter follows more closely the behaviour of an ideally calibrated estimate (dashed line).

[1] 2nd place: https://github.com/SeuTao/Humpback-Whale-Identification-Challenge-2019_2nd_palce_solution.

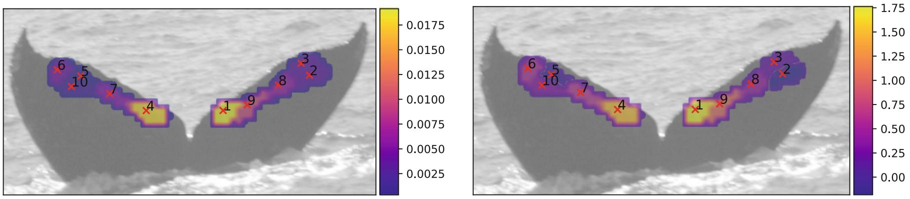

Fig. 4. Top: Average sensitivity heatmap rendered on the landmark locations of one image, representing the average reduction in the score of the correct class after removing each landmark. **Bottom**: Average loss in accuracy, in percent points, after removing each landmark. Photo CC BY-NC 4.0 John Calambokidis.

5.3 Heatmapping Results and Comparison with Whale Expert Knowledge

Figure 4 shows the mean landmark sensitivity (top), as well as the loss of accuracy after removing landmarks (bottom), calculated over the complete data set. When compared to the landmarks near the fluke tips, it can be seen that the landmarks near the notch change the score the most, and flip the classification towards the correct class the most often. This is consistent with the fact that the interior of a fluke changes rather little over time, while the fluke tips can change significantly over time. Also, the pose and activity of the whale when the images are captured might explain this behavior. It is worth noting that all the attention is concentrated along the trailing edge of the fluke. This may be due to the fact that it is the area of the fluke that is most reliably visible in the images, since the leading edge tends to be under water in a number of photos.

In the following, we examine the landmark-based and patch-based tools in terms of the features considered as important by the whale expert on individual images. We show the results on two pairs of images such that each pair belongs to the same individual. Figure 5a highlights the main areas the expert focused on in order to conclude whether they do belong to the same individual or not after inspecting both images side-to-side. Note the tendency of the expert of annotating just a small number of compact regions.

The heatmaps obtained using patch-based occlusion are shown in Fig. 5b. Although the fluke itself is recognised as being important to the classification, no particular area is highlighted, except for one case where the whole trailing edge appears to be important. In addition, some regions outside of the fluke seem to have a negative sensitivity, pointing at the possibility of an artifact in the dataset that is being used by the model. This was observed in previous publications [14], where authors concluded that patch-based occlusion was using the shape of the entire fluke, rather than specific, localised patterns.

The results of the landmark-based approach, in Fig. 5c, show more expert-like heatmaps, with the evidence for and against a match always located on the fluke and generally around the trailing edge and close to the notch. In each case, only a few small regions are responsible for the evidence in favor of assigning each pair to the same individual. However, although both the expert and the

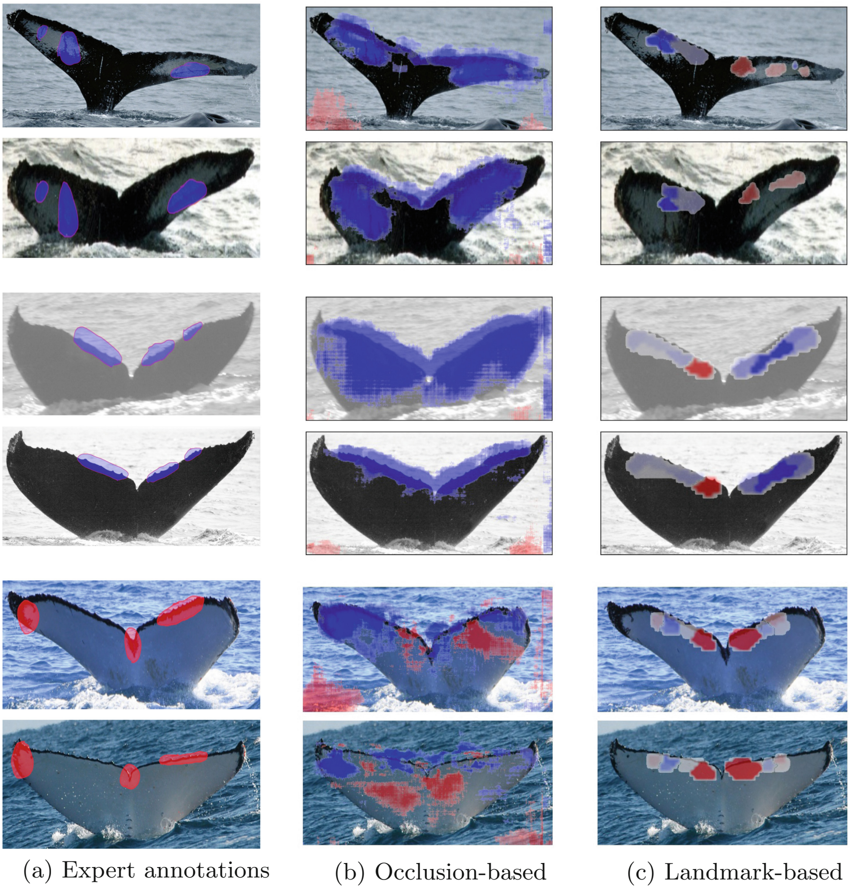

Fig. 5. Heatmaps of attribution. Dark blue/red areas highlight the regions that are estimated to provide evidence for/against the match. The top two pairs are matching pairs (same individual) while the bottom one is not a match. (Color figure online)

landmark-based method have a tendency of pointing at the same general areas around the trailing edge with compact highlights, we do not observe a consistent overlap with the expert annotated images. This may be due to constraints in both the expert and the landmark-based highlights. Unlike the expert, the landmark-based approach tends to focus, by design, in the areas of the fluke that are most reliably visible. The expert, on the other hand, explores all visible fluke features and highlights them in a non-exhaustive manner. On the top image pair, a region that is also annotated by the expert on the left fluke provides most of the positive evidence, but a feature close to the leading edge is ignored. This is probably due to the model learning that the leading edge is less reliable, since it is under water in a large number of photos. On the middle pair, the area to the left of notch is assigned a negative sensitivity while being annotated as important by the

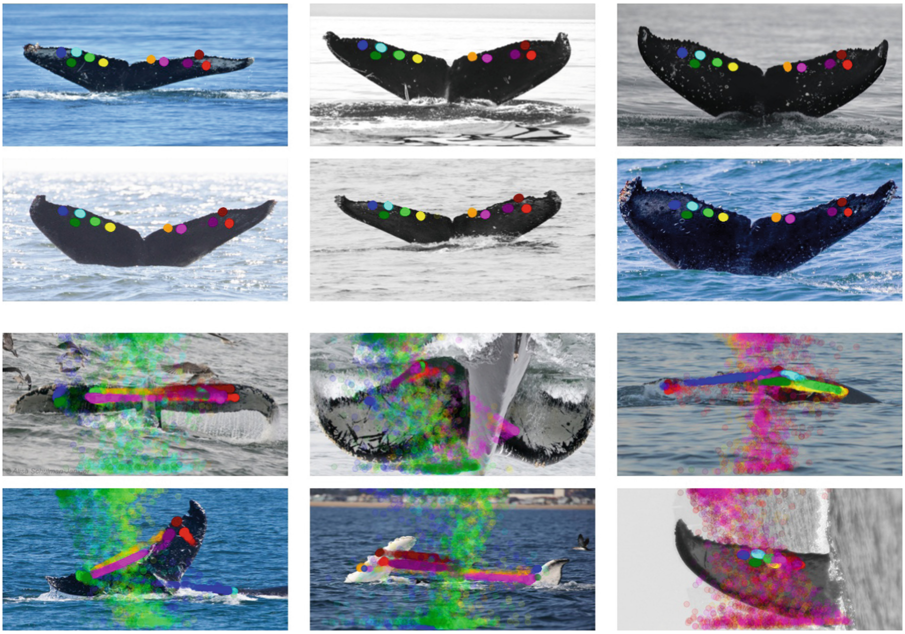

Fig. 6. Spatial uncertainty of each landmark on different whales determined by means of 500 dropout runs on the feature tensor **Z**. Each disk represents the location of a landmark in one run and each of the ten landmarks is colored consistently across images. **Top**: The test images with the lowest uncertainty. **Bottom**: The test images with the highest uncertainty. (Color figure online)

expert. On the bottom pair we see that only the landmarks closest to the notch are used by the model to decide that the images do indeed belong to different individuals, while the expert has also annotated a region close to the fluke tip, which the landmark-based model systematically ignores, likely due to the fact, as with the leading edge, that the tips are less reliably visible in the images.

5.4 Spatial Uncertainty of Individual Landmarks

The visualizations in Fig. 6 display the six images in the test set with the lowest and with the highest uncertainty, each on a different individual. The colored disks represent the positions of each landmark across 500 random application of dropout, with a dropout probability of 0.5, to the feature tensor **Z**. The colors are consistent (e.g. landmark 5, as seen in Fig. 4 is always represented in dark blue). The top rows tend to contain images with clearly visible flukes in a canonical pose. As we can see, the detected keypoints do behave as landmarks, each specializing in a particular part of the fluke, even if no particular element of the loss was designed to explicitly promote this behaviour. The bottom rows contain images with either substantial occlusions or uncommon poses. This shows how the spatial uncertainty uncovered by MC dropout can be used to detect unreliably located landmarks, which in turn can be used to find images with problematic poses and occlusions that are likely to be unsuitable for identification.

6 Conclusion and Outlook

In this work, we explore the use of landmark detection learning using only class labels (i.e. whale identities) and apply it to gain insights into which fluke parts are relevant to the model's decision in the context of cetacean individual identification. Our experiments show that, compared to patch-based occlusion mapping, our approach highlights regions in the images that are systematically located along the central part of the trailing edge of the fluke, which is the part most reliably visible in the images. At the same time, the landmarks highlight compact regions that are much more expert-like than the baseline OSM heatmaps. In addition, we show that the agreement of random subsets of the landmarks is a better estimate of the expected error rate than the softmax score. However, there seems to be little agreement between the specific regions chosen by the expert and the landmark-based highlights.

The use of landmarks makes it easy to match them across images, since each landmark develops a tendency to specialize on a particular region of the fluke. This allowed us to study their average importance for the whole validation set, leading us to conclude that the areas of the trailing edge right next to the notch tend to be the most relied upon. This is probably due to the to the higher temporal stability of the region around the notch, which is less exposed and thus less likely to develop scars, and to the fact that the trailing edge is the part of the fluke most often visible in the photos. Is also worth noting that the proposed method is inherently interpretable, thus not only guaranteeing that the generated heatmaps are relevant to the model's decision, but also doing so at a negligible computational cost, requiring to perform inference once and not using any gradient information. In addition, the accuracy obtained is noticeably higher than a model with the same base architecture but no attention mechanism.

In spite of these advantages, we also observed an inherent limitation of the method when compared to the expert annotations. Our landmark-based model requires to find all landmarks on each image, resulting in a tendency to only focus on the areas of the fluke that are most reliably visible and discarding those that are often occluded, such as the tips and the leading edge. Designing a model that is free to detect a varying number of landmarks is a potential path towards even more expert-like explanations.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Andrew, W., Greatwood, C., Burghardt, T.: Aerial animal biometrics: individual friesian cattle recovery and visual identification via an autonomous UAV with onboard deep inference. In: IROS (2019)

3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: CVPR (2017)
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: CVPR, pp. 4690–4699 (2019)
5. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: ICML, pp. 1050–1059 (2016)
6. Gawlikowski, J., et al.: A survey of uncertainty in deep neural networks. arXiv preprint arXiv:2107.03342 (2021)
7. Ghanem, R., Higdon, D., Owhadi, H. (eds.): Handbook of Uncertainty Quantification. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-12385-1
8. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning, May 2018, arXiv preprints arXiv:1806.00069
9. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
11. Hohman, F.M., Kahng, M., Pienta, R., Chau, D.H.: Visual analytics in deep learning: an interrogative survey for the next frontiers. IEEE Trans. Visual Comput. Graph. **25**(1), 1–20 (2018)
12. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. **110**(3), 457–506 (2021). https://doi.org/10.1007/s10994-021-05946-3
13. Katona, S., Whitehead, H.: Identifying humpback whales using their natural markings. Polar Rec. **20**(128), 439–444 (1981)
14. Kierdorf, J., Garcke, J., Behley, J., Cheeseman, T., Roscher, R.: What identifies a whale by its fluke? on the benefit of interpretable machine learning for whale identification. In: ISPRS Annals, vol. 2, pp. 1005–1012 (2020)
15. Kniest, E., Burns, D., Harrison, P.: Fluke matcher: a computer-aided matching system for humpback whale (Megaptera novaeangliae) flukes. Mar. Mamm. Sci. **3**(26), 744–756 (2010)
16. Li, S., Li, J., Tang, H., Qian, R., Lin, W.: ATRW: a benchmark for Amur tiger re-identification in the wild. In: ACM International Conference on Multimedia, pp. 2590–2598 (2020)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV, pp. 2980–2988 (2017)
18. Loucks, D., Van Beek, E., Stedinger, J., Dijkman, J., Villars, M.: Model sensitivity and uncertainty analysis. Water Resources Systems Planning and Management, pp. 255–290 (2005)
19. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digit. Sig. Process. **73**, 1–15 (2018)
20. Rajaraman, S., et al.: Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images. J. Med. Imaging **5**(3), 034501 (2018)
21. Roscher, R., Bohn, B., Duarte, M.F., Garcke, J.: Explainable machine learning for scientific insights and discoveries. IEEE Access **8**, 42200–42216 (2020)
22. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)

23. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: a review of methods and applications. Proc. IEEE **109**(3), 247–278 (2021)
24. Samek, W., Müller, Klaus-R.: Towards explainable artificial intelligence. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700, pp. 5–22. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_1
25. Schneider, S., Taylor, G.W., Linquist, S., Kremer, S.C.: Past, present and future approaches using computer vision for animal re-identification from camera trap data. Methods Ecol. Evol. **10**(4), 461–470 (2019)
26. Schofield, D., et al.: Chimpanzee face recognition from videos in the wild using deep learning. Sci. Adv. 5(9), **eaaw0736** (2019)
27. Schramowski, P., et al.: Right for the wrong scientific reasons: revising deep networks by interacting with their explanations. arXiv preprint arXiv:2001.05371 (2020)
28. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)
29. Simoes, H., Meidanis, J.: Humpback whale identification challenge: a comparative analysis of the top solutions (2020)
30. Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G.: Evaluation of uncertainty quantification in deep learning. In: Lesot, M.-J., et al. (eds.) IPMU 2020. CCIS, vol. 1237, pp. 556–568. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50146-4_41
31. Stomberg, T., Weber, I., Schmitt, M., Roscher, R.: Jungle-net: using explainable machine learning to gain new insights into the appearance of wilderness in satellite imagery. In: ISPRS Annals, vol. 3, pp. 317–324 (2021)
32. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML, pp. 3319–3328. PMLR (2017)
33. Surma, S., Pitcher, T.J.: Predicting the effects of whale population recovery on northeast pacific food webs and fisheries: an ecosystem modelling approach. Fish. Oceanogr. **24**(3), 291–305 (2015)
34. Wang, H., Yeung, D.Y.: A survey on Bayesian deep learning. ACM Comput. Surv. (CSUR) **53**(5), 1–37 (2020)
35. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**(2), 207–244 (2009)
36. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: ICML, pp. 2048–2057. PMLR (2015)
37. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
38. Zhang, Y., Guo, Y., Jin, Y., Luo, Y., He, Z., Lee, H.: Unsupervised discovery of object landmarks as structural representations. In: CVPR, pp. 2694–2703 (2018)
39. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR, pp. 2921–2929 (2016)