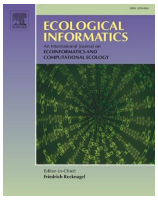
Ecological Informatics 91 (2025) 103396

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Advances in deep learning-driven photo identification and meta analysis of cetaceans in large data repositories

Alexander Barnhill [a,*], Jared R. Towers [b,c], Tasli J.H. Shaw [b,d], Magdalena Arias [e], Adrián Bécares [a], Thomas Doniol-Valcroze [c], Lorenzo von Fersen [f,g], Rodrigo Genoves [h,m], Tim Rörup [a], Gary J. Sutton [b,i], Sheila Thornton [j], Michael Weiss [k], Andreas Maier [a], Elmar Nöth [a], Christian Bergler [l]

[a] Friedrich-Alexander-Universität Erlangen-Nürnberg, Pattern Recognition Lab, Erlangen, Germany
[b] Bay Cetology, Alert Bay, British Columbia, Canada
[c] Pacific Biological Station, Fisheries and Oceans Canada, Nanaimo, British Columbia, Canada
[d] Humpback Whales of the Salish Sea, Duncan, British Columbia, Canada
[e] Center for Applied Research and Technology Transfer in Marine Resources "Almirante Storni" (CIMAS, CONICET), Faculty of Marine Sciences, National University of Comahue, Argentina
[f] Nuremberg Zoo, Nuremberg, Germany
[g] YAQU PACHA e.V., Nuremberg, Germany
[h] Kaosa, Rio Grande, Brazil
[i] Ocean Wise, Vancouver, British Columbia, Canada
[j] Marine Mammal Conservation Physiology Program, West Vancouver, British Columbia, Canada
[k] Center for Whale Research, Friday Harbour, Washington, USA
[l] Technical University of Applied Sciences Amberg-Weiden, Department of Electrical Engineering, Media and Computer Science, Amberg, Germany
[m] Oceanographic Museum 'Prof. Eliézer de C. Rios', Federal University of Rio Grande - FURG, Rio Grande, Brazil

ARTICLE INFO

Keywords:
Deep learning
Photo identification
Marine conservation
Data curation
Resource efficient machine learning

ABSTRACT

Photo-identification of cetaceans remains a labor-intensive task, requiring expert annotation of long-tailed image datasets in which most individuals are rarely encountered. We present a scalable, end-to-end framework that automates this process using lightweight deep learning models optimized for resource-constrained environments. Our modular pipeline integrates state-of-the-art detection (YOLOv8-small), individual identification via metric learning (EfficientNet-B0 with a contrastive head), and auxiliary modules for image quality scoring, side classification, and identifiability prediction. Unlike previous approaches limited to single-species applications or high-resource settings, our framework generalizes across five cetacean populations with diverse visual characteristics. We achieve top-1 identification accuracies of 0.92 for Bigg's killer whales (*Orcinus orca rectipinnus*), 0.96 for Southern resident killer whales (*Orcinus orca ater*), 0.96 for Lahille's bottlenose dolphins (*Tursiops truncatus gephyreus*), 0.82 for common minke whales (*Balaenoptera acutorostrata scammoni*), and 0.85 for humpback whales (*Megaptera novaeangliae*), yielding a cross-species accuracy of 0.90. To support image triage in large datasets, we include a quality scoring module that predicts image utility using learned embedding features. This module achieves an R^2 of 0.799, enabling intelligent prioritization of data. Runtime evaluations show processing speeds of 1.6–3.2 images/s on CPU and 9.6–23.3 FPS with GPU acceleration, making it suitable for archival and real-time applications. We also evaluate the impact of demographic metadata (age, sex) on identification performance and provide practical recommendations for future dataset design. The system is available via a web interface designed to support real-world conservation workflows with minimal computational overhead.

* Corresponding author.
E-mail addresses: alexander.barnhill@fau.de (A. Barnhill), c.bergler@oth-aw.de (C. Bergler).

https://doi.org/10.1016/j.ecoinf.2025.103396
Received 31 May 2025; Received in revised form 12 August 2025; Accepted 12 August 2025
Available online 22 August 2025

1. Introduction

Automated identification of individuals using machine learning has been extensively studied in humans, typically focusing on features like facial structure and iris patterns (Manna et al., 2020; Mohammed et al., 2020; Zanlorensi et al., 2018). Building on these foundations, similar techniques are increasingly applied to non-human animals to estimate population abundance and track individuals over time (Bogucki et al., 2019; Pollicelli et al., 2017; Schofield et al., 2019). Automation is critical for managing large photo datasets, as it enables up-to-date, accurate information by ensuring image quality without relying solely on labor-intensive manual annotation.

Photo-identification (photo-ID) has become a cornerstone method in marine mammal science and has been used for investigation of social dynamics and evolution (Dungan et al., 2016; Nielsen et al., 2021; Nielsen et al., 2023; Zanardo et al., 2018), behavioural analysis (Genoves et al., 2020; Gero et al., 2014; McMillan et al., 2019; Tixier et al., 2016; Towers et al., 2013; van Weelden et al., 2025), and population abundance (Hammond et al., 1990; Jordaan et al., 2020; Jourdain et al., 2021; Matkin et al., 2012) for several species. This process relies on manual image analysis which involves experts matching physical features visible in photographs of individuals to existing reference images. Despite advances in technology, most cetacean photo-ID datasets worldwide still depend on such manual workflows.

Efforts to automate identification, such as the FIN-PRINT pipeline for killer whales (*Orcinus orca*) (Bergler et al., 2021), have improved efficiency by handling the most commonly photographed individuals. Nevertheless, these approaches often focus on a subset of the population and do not fully address temporal changes in individual appearance, such as dorsal fin growth over years, or other important factors like eye patch detection, photo quality assessment, or determining the side of the animal photographed (left or right). Many automated systems also concentrate solely on species or individual classification, overlooking these supporting tasks crucial for maintaining dataset continuity and quality (Hou et al., 2020; Nguyen et al., 2017; Patton et al., 2023).

Traditional supervised learning approaches require large, balanced datasets with many samples per individual, which is often impractical due to the elusive nature of wildlife and the long-tailed distribution of data (Fu et al., 2022). In contrast, contrastive (metric) learning offers an alternative by learning feature embeddings that cluster images of the same individual closer together while pushing different individuals apart. This method has shown success in re-identification tasks for humans (Ren et al., 2019; Yang et al., 2017) and various cetaceans (Bouma et al., 2018; Patton et al., 2023) as well as other mammal species (eg. Miele et al., 2021). Its flexibility makes metric learning well suited for datasets with few images per class and high intra-class variability (Li et al., 2022; Shao and Peng, 2024).

Recent advances in marine monitoring include real-time underwater detection models (Pan et al., 2025; Zhou et al., 2023) and image enhancement techniques to overcome optical distortions (Yao et al., 2025). While these developments improve data acquisition and preprocessing, our work focuses on a complementary need: an end-to-end pipeline for large-scale identification of individual cetaceans from photographic datasets. Related fields, such as medical image analysis (Das et al., 2023) and person re-identification in surveillance (Saini et al., 2024), illustrate the broad applicability and impact of deep learning for identification under challenging visual conditions.

This paper presents two main contributions. First, we introduce a modular, deployable pipeline combining state-of-the-art deep learning models to perform dorsal fin (and in applicable species) eye patch detection, relevance and quality filtering, side determination, and individual identification across five populations in four diverse cetacean species. All components balance accuracy with computational efficiency to enable practical deployment in resource-constrained environments. Second, we analyze the effects of long-term dataset evolution on model performance, exploring biologically informed data partitioning strategies such as by age and sex to improve generalization and robustness. Together, these contributions provide both practical tools and insights to support ongoing studies of cetaceans using photo-ID.

2. Materials

The methods developed in this study are primarily based on a large data repository of the Bigg's killer whale (*Orcinus orca rectipinnus*) (BKW) population. This dataset was chosen due to its extensive collection of images spanning many years, along with rich metadata about individual whales, including sex and year of birth. These factors enable nuanced investigations into key considerations for training models, such as determining appropriate data splits and accounting for demographic variability. Moreover, the BKW dataset exhibits a long-tailed data distribution, a characteristic frequently encountered in ecological studies, allowing the exploration and evaluation of techniques for managing such imbalances.

All models developed and implemented in this work, except those explicitly designed for identifying individual whales, are trained exclusively on the BKW data and subsequently evaluated on additional datasets described in more detail in Supplementary Material section S1. These include a variety of cetacean species including Southern Resident Killer Whales (SRKW) (*Orcinus orca ater*), Lahille's bottlenose dolphins (LBD) (*Tursiops truncatus gephyreus*), Common Minke whales (CMW) (*Balaenoptera acutorostrata scammoni*), and Humpback whales (HW) (*Megaptera novaeangliae*). Utilization of these additional populations enables an assessment of the models' ability to generalize across diverse cetacean populations, each exhibiting distinct identifying characteristics. The statistics for each population utilized in this study, including number of individuals in each population, as well as the number of images per individual, are provided in Table 1, similarly, examples of the individuals extracted from their source material can be seen in Fig. 1.

The datasets used in this study comprise images collected by dozens of photographers using diverse equipment and under varying environmental conditions, including different lighting, angles, and sea states. Images span a wide geographic range and include contributions across multiple years. As such, both the training and evaluation splits are inherently heterogeneous, supporting generalization across realistic variations in capture conditions.

3. Methods

The goal of this pipeline is to extract a structured hierarchy of information from individual images using a series of deep learning models. The process begins with the detection of regions of interest that contain identifying features. For the species studied here, this involves detecting the dorsal fin. The aim of the model is to identify the smallest bounding box that circumscribes the fin (and saddle patch) to minimize background noise. This cropped region is then used in subsequent steps to assess image quality, determine the visible side of the individual (left or right), and ultimately generate an identification prediction.This detection framework can be extended to other identifying features, such as the eye patch which is frequently visible in killer whale photographs and can provide high utility for identification.

The second stage focuses on evaluating the quality of the bounding box contents. Although dorsal fin detection is generally robust, false positives are not uncommon. These often stem from model errors, such as misclassifying boats, humans, or landscape features. Furthermore, some regions may include parts of the animal, such as a fluke or pectoral fin, that are not useful for identification. Quality assessment is therefore divided into two stages: (1) determining whether the content is within the domain of interest (for example, a dorsal fin), and (2) determining whether it is suitable for identification. This is conceptually similar to the Valid-versus-Invalid (VVI) filtering approach in FIN-PRINT (Bergler et al., 2021).

An additional component is side classification, which involves

Table 1
An overview of the data used for training individual identification models for each of the groups / species presented here.

Dataset	Total Images	Total Individuals	Median Img/ID	Mean Img/ID	Std Dev Img/ID	Min Img/ID	Max Img/ID
BKW	90,501	405	124	≈ 223	≈ 236	3	1302
SRKW	18,250	83	186	≈ 219	≈ 112	18	505
LBD	4440	23	200	≈ 193	≈ 51	83	291
CMW	624	8	87.5	≈ 78	≈ 44	16	147
HW	3713	81	42	≈ 46	≈ 22	10	145

For each experiment the data was split using 70 % / 15 % / 15 % for training, validation, and testing, respectively.
Abbreviations
BKW: Bigg's Killer Whale
SRKW: Southern Resident Killer Whale.
LBD: Lahille's Bottlenose Dolphin.
CMW: Common Minke Whale.
HW: Humpback Whale.

Fig. 1. A selection of images displaying not only the variety in the species and groups presented here, but also the variety in the image quality. Photos courtesy of Jared R. Towers (Bigg's Killer Whale, Common Minke Whale), Center for Whale Research (Southern Resident Killer Whale, NMFS/NOAA Permit 27,038), Tasli J.H. Shaw (Humpback Whale), Rodrigo Genoves (Lahille's Bottlenose Dolphin).

determining whether the left or right side of the individual is shown. As individuals may differ in coloration or scarring between sides, documenting both perspectives is critical for accurate re-identification and cataloging efforts.

The final step involves identifying the individual depicted. Although the broader cetacean population can be considered an open-world

recognition problem (Bendale and Boult, 2015), the populations analyzed here are well-monitored, with most individuals appearing in the dataset multiple times. Consequently, the open-world challenge is not addressed in this work. A graphical representation of the pipeline is shown in Fig. 2. In this depiction, it is important to note that side classification occurs after the identification step. While this ordering can be adjusted based on the practitioner's preference, the experiments showed minimal difference between performing orientation prediction before or after identification. This is largely because the images had already been filtered by the domain and identifiability modules. As a result, there was little need for an "other" class to indicate low-quality or ambiguous images, since such images had already been excluded from the identification process.

Model architecture, parameter counts, loss functions, and training

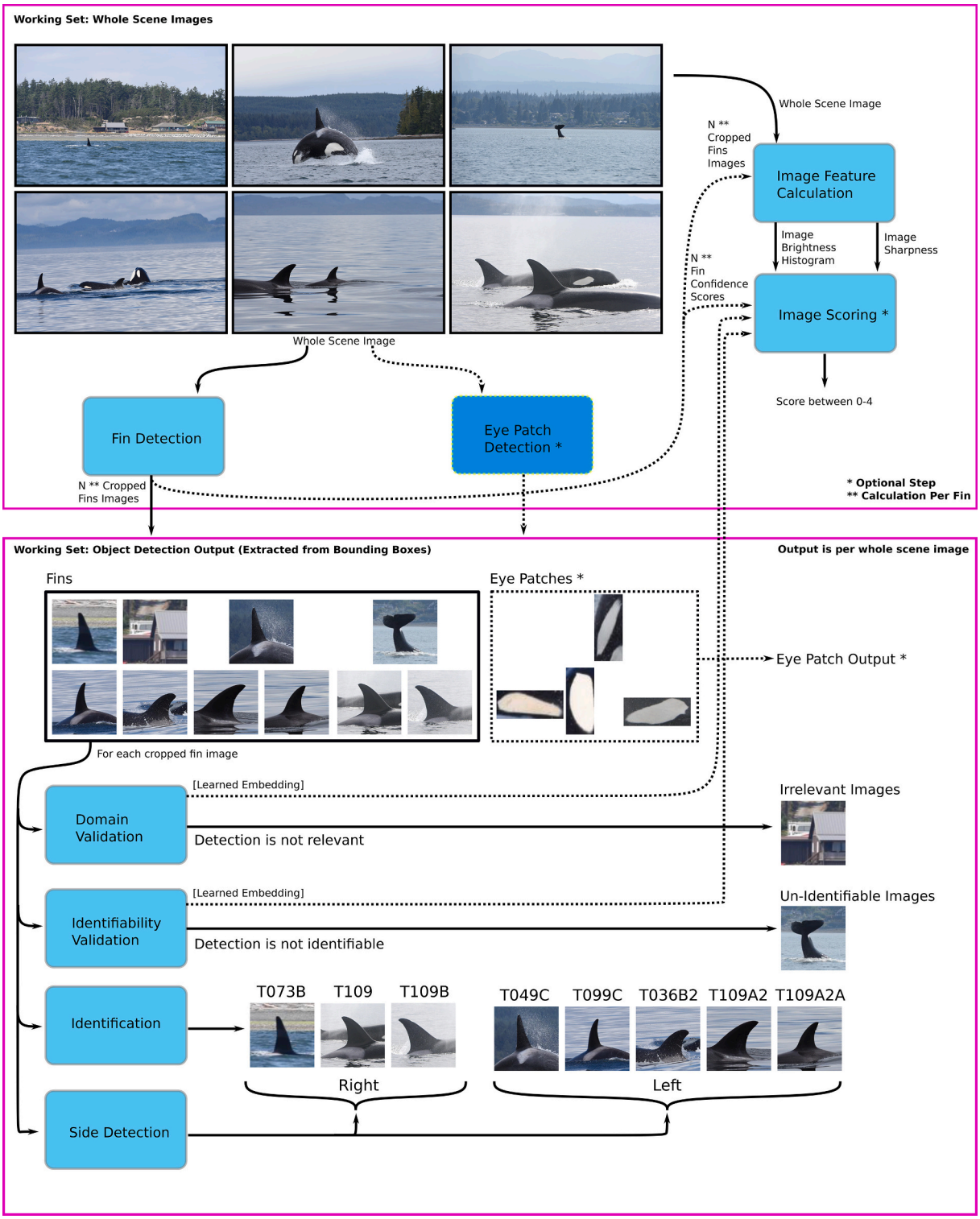


Fig. 2. Overview of the identification pipeline. The workflow processes raw images through several stages: filtering out irrelevant or unidentifiable content, detecting regions of interest, predicting orientation, and identifying individuals. Optional steps are indicated (*). The final output is a categorized set of images based on detected content.

configurations are provided in Table 2. The metrics reported for each step in the pipeline are the average across five runs. The data for each module was split into 70 % / 15 % / 15 % portions for training, validation, and testing. Each model was trained on a single Tesla V100 GPU (16GB). To prevent data leakage while enabling robust identity learning, train and test splits were constructed such that images of the same individual could appear in both sets, but only from distinct encounters; this ensures that no photos from a single sighting event are shared across splits, preserving the integrity of evaluation.

3.1. Detection of regions of interest

Detecting the region of interest (ROI) containing the dorsal fin is a critical first step in the identification pipeline for dorsal-finned marine mammals. For killer whales, the goal is to tightly capture both the dorsal fin and the saddle patch while minimizing background. This task is performed using a YOLOv8-small object detection model (Jocher et al., 2023), which balances accuracy with computational efficiency, making it well-suited for real-time or edge-device deployment. Example detections, both usable and unusable, are shown in Fig. 3.

YOLOv8 introduces several architectural refinements over earlier versions, including an anchor-free detection head, decoupled classification and regression branches, and a lightweight backbone with C2f modules for improved gradient flow and reduced parameter count. The model was initialized with pretrained weights and fine-tuned using domain-specific fin imagery.

Species-specific features like the eye patch, though not always visible, can be helpful for cataloging and ecotype classification. A secondary YOLOv8-based model was trained on a dedicated eye patch dataset and integrated into the pipeline as an optional module. This approach generalizes well to other features, such as flukes or pectoral fins, even with relatively small datasets (see Fig. 4).

To prepare detections for downstream processing, each cropped region is standardized. Fin images are centered and resized to 512 × 512 px, preserving the aspect ratio. While this may include some background, it retains the necessary visual detail. For eye patches, the longest side is scaled to 512 px to support manual interpretation rather than further model input.

Training data for dorsal fin detection are based on the Extended Annotation Dataset (EADD) from FIN-PRINT (Bergler et al., 2021), here referred to as the Fin Detect Dataset (FDD). To assess the model's robustness, 1500 negative samples were added (≈20 % of the original dataset), forming the Fin Detect Dataset with Backgrounds (FDD + B). Summary statistics for both datasets are provided in Table 3.

Eye patch detection training data were manually annotated to create the Eye Patch Detection Dataset (EDD), comprising 2641 images. These images are selected from the BKW dataset which had been manually annotated as having eye patches. As with dorsal fin detection, an additional 528 background-only images were added (≈20 % of the original dataset), resulting in the Eye Patch Detection Dataset with Backgrounds (EDD + B). Details are summarized in Table 4.

Each object detection model was evaluated using standard metrics: precision, recall, and mean average precision (mAP). A predicted bounding box is considered correct if its Intersection over Union (IoU) with a ground truth box exceeds a specified threshold. Precision measures the proportion of correct detections among all predicted boxes, while recall measures the proportion of ground truth objects that were correctly detected. Average Precision (AP) summarizes the precision-recall tradeoff at various confidence thresholds. The mean Average Precision (mAP) is the mean of the AP values, typically reported as mAP@50 (IoU threshold = 0.5) and mAP@50–95 (mean AP over IoU thresholds from 0.50 to 0.95 in steps of 0.05).

3.2. Extracting quality assessments and identification-relevant information

For each bounding box predicted by the fin detection network, we apply a series of deep learning models to assess image characteristics relevant to identification. These models evaluate whether an image is likely to contain a cetacean of interest, whether the individual is visually identifiable, and which lateral side of the individual is shown. These assessments enable automatic filtering of low-quality or irrelevant images and help prioritize high-quality images for archival or field-based re-identification. Images that are not identifiable but are still within the domain (e.g., containing a cetacean) remain accessible for manual review.

All models in this stage, as well as those used for individual identification, follow a shared architectural design consisting of a feature extraction backbone, a neck module for representation transformation, and a task-specific classification head. These heads support binary tasks (e.g., identifiability) and multi-class prediction (e.g., orientation).

For feature extraction, we employ a compact convolutional architecture optimized for both accuracy and computational efficiency. Specifically, we use the B0 variant from a family of models that scale depth, width, and resolution in a balanced manner. This architecture was selected for its favorable trade-off between performance and resource usage, making it well-suited for deployment on a wide range of hardware, including edge devices.

A total of 5000 images were randomly sampled from the BKW dataset (2011−2021) and processed through the fin detector, yielding 7007 cropped images due to multiple detections per input. Manual review identified 378 images (5 %) as out-of-domain (e.g., birds, boats, people). The remaining 6629 in-domain images were further evaluated for identifiability; 681 (10 %) were marked as unidentifiable due to occlusion, poor angle, or absence of a visible fin. To mitigate class imbalance, identifiable and non-identifiable subsets were resampled to create balanced training sets. Representative examples from each class are shown in Fig. 5.

A third model predicts the orientation of the individual (left, right, or other), as both sides are required for comprehensive re-identification.

Table 2
An overview of the model architectures, their loss functions, number of parameters, optimizer, initial learning rates, and maximum training epochs for each module in the pipeline presented here.

Task	Architecture	Loss Function(s)	# Parameters	Optimizer	Learning Rate	Max Epochs
Fin Detection	YOLOv8 (Small)	CIoU Loss, BCE / Focal Loss, BCE Loss	11.2 M	SGD	1e-2	50
Eye Patch Detection	YOLOv8 (Small)	CIoU Loss, BCE / Focal Loss, BCE Loss	11.2 M	SGD	1e-2	50
Domain Relevance	EfficientNet B0	Cross Entropy	5.3 M	Adam	1e-3	50
Identifiability Determination	EfficientNet B0	Cross Entropy	5.3 M	Adam	1e-3	50
Side Detection	EfficientNet B0	Cross Entropy	5.3 M	Adam	1e-3	50
Identification	EfficientNet B0	Sub-center ArcFace w/ Cosine Similarity	5.3 M	Adam	1e-3	200

Abbreviations
YOLO: You Only Look Once
CIoU: Complete Intersection Over Union.
BCE: Binary Cross Entropy.
SGD: Stochastic Gradient Descent.

Fig. 3. Examples of images which contain properly positioned images useful for killer whale photo identification (top row), and images which are either not useful for photo identification or which are of irrelevant targets which may yet still appear in large photo repositories (bottom row). This highlights the large differences contained within these volumes and demonstrates the necessity to only focus on relevant images. All images courtesy of Jared R. Towers.

Fig. 4. Examples of the variety of images available which contain eye patches. These may be visible during a number of events such as surfacing, spy hops and breaches. These examples demonstrate the wide range of orientations in which identifying features may appear within large image archives. All images courtesy of Jared R. Towers.

Table 3
An overview of the samples contained within the train, validation, and test splits for the Fin Detect Dataset (FDD) and the Fin Detect Dataset with Backgrounds (FDD + B), which contains additional samples that do not contain any dorsal fins.

Dataset	∑	Training				Validation				Test			
		Valid	Invalid	∑	%	Valid	Invalid	∑	%	Valid	Invalid	∑	%
FDD	7510	5257	0	5527	70.0	1127	0	1127	15.0	1126	0	1126	15.0
FDD + B	9010	5257	1050	6307	70,0	1127	225	1352	15.0	1126	225	1351	15.0

Table 4
The samples for the training, validation, and test sets for the eye patch detection models.

Dataset	∑	Training				Validation				Test			
		Valid	Invalid	∑	%	Valid	Invalid	∑	%	Valid	Invalid	∑	%
EDD	2641	1849	0	1849	70.0	396	0	396	15.0	396	0	396	15.0
EDD + B	3169	1849	370	2219	70,0	396	79	475	15.0	396	79	475	15.0

The EDD differs from the EDD + B only in that the EDD + B contains images which do not contain eye patches. Those images without eye patches are therefore marked as *invalid*.

Two datasets were prepared: one with balanced left/right images (1635 each), and another incorporating an "other" class (126 non-identifiable images and 300 additional left/right images for balance). Dataset details are provided in Table 5. Examples of images with clearly identifiable orientation (left or right) and images with more ambiguous orientation are shown in Fig. 6.

To evaluate the performance of these models used in filtering and classifying dorsal fin images, standard evaluation metrics such as accuracy, precision, and recall are employed. Accuracy measures the proportion of correct predictions, both positive and negative, out of all predictions made. Precision ensures that the model is not falsely identifying irrelevant images as relevant by calculating the proportion of true positives out of all positive predictions. Recall, on the other hand, measures the model's ability to correctly identify all relevant images, calculated as the proportion of true positives out of all actual positive instances. These metrics together offer a comprehensive view of the model's performance in terms of identifying usable images for further review and re-identification.

Image quality plays a critical role in the identification of individuals within complex biological systems. In the context of marine mammal conservation, particularly for photo-ID tasks, human experts intuitively assess various visual cues, including focal clarity, contrast, subject size and prominence, and species-specific visual traits, to determine whether an image is suitable for analytical use. To approximate this human intuition at scale, a dedicated quality evaluation module was implemented which is capable of assigning a numeric score to each image

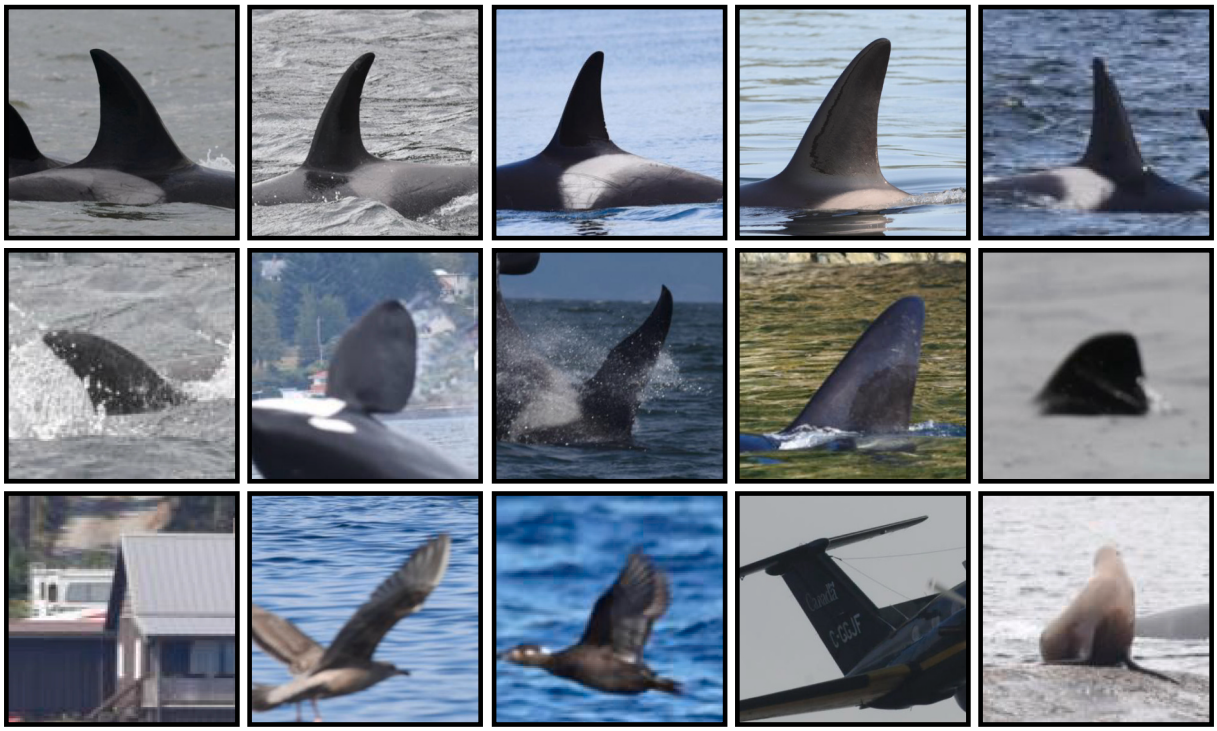


Fig. 5. Examples of the quality of images after the initial extraction step. Those suitable for downstream tasks such as identification (in-domain, identifiable) are shown in the top row. Those images which are relevant but not identifiable (in-domain, unidentifiable) are in the middle row. Images which are not relevant are shown in the bottom row (out-of-domain). All source images courtesy of Jared R. Towers.

Table 5
The samples for each of the tasks in the determination of quality-related characteristics.

Dataset	Total Images	Classes	# Img per Class
		In Domain	500
Domain Relevance	878	Out Of Domain	378
		Identifiable	909
Identifiability	1590	Not Identifiable	681
		Left	1635
Side	3720	Right	1635
		Left	300
		Right	300
Side w/ Other Class	726	Other	126

based on a set of automatically derived visual features.

The module uses as inputs both image-level and region-level characteristics, primarily derived from the fin detection output. For each image, up to *n* detected fin regions are examined. For each detected fin, we extract the following features:

- Detection confidence score from the YOLOv8 model
- Domain relevance representation
- Identifiability representation
- Brightness and sharpness histograms

For the representations for domain relevance and identifiability, there are two options: either the output logits could be used, or the learned embedding. Both are experimented with here. In addition, we compute the following global features for the full image:

- Total number of detected fins
- Relative area of the largest detected fin (normalized by image size)
- Full-image sharpness and brightness histograms

All features are concatenated into a single feature vector and standardized to have zero mean and unit variance before being passed to a multi-layer perceptron (MLP). The MLP outputs a continuous quality score ranging from 0 to 4.

To train the quality module, 5000 full-scene images were randomly sampled from the BKW dataset. Each image was independently scored by two annotators, one of whom is an expert in killer whale identification, using the following rubric:

- 0: No identifiable individual present
- 1: Whales present but not identifiable
- 2: Whales present and identifiable
- 3: High-quality image with few individuals
- 4: High-quality image with a clearly visible, identifiable individual

After data selection and review of the data distribution across the five scores, subsampling was applied to the dataset to reduce data imbalance. This resulted in 2, 244 images across the five scores. Due to the subjective nature of image quality, label smoothing with a factor of 0.5 was applied during training to reduce overfitting and improve robustness.

The metrics used to evaluate this module are the coefficient of determination (R^2), which measures how much of the variance in the data is explained by the model, with a value of one indicating perfect prediction, and the mean squared error (MSE). While the quality score is inherently subjective, it offers a scalable and consistent proxy for expert judgment, enabling large datasets to be filtered and curated more efficiently for downstream identification tasks.

3.3. Identification of individuals

The final step, individual identification, relies on the availability of high-quality, in-domain images produced by the preceding stages. For many datasets, including the BKW data, traditional fully supervised classification is challenged by a long-tailed distribution, in which most individuals are sparsely represented. While techniques such as over- and

Fig. 6. Examples of the variety of images which, once extracted from their source scene photograph, display different positions of individuals. Some of these images (top row) are usable for identification purposes while others (bottom row) are less suited for the task. All source images courtesy of Jared R. Towers.

under-sampling can be used to address this imbalance, contrastive learning has proven particularly effective in such settings (Li et al., 2022; Shao and Peng, 2024; Zhu et al., 2022) and is therefore employed here to train robust identification models.

To enhance the discriminative power of the learned embeddings, we employ the Sub-Center ArcFace loss (Deng et al., 2020). This loss function extends ArcFace (Deng et al., 2019) by replacing each class's single prototype with multiple subcenters, allowing for better modeling of intra-class variability. During training, cosine similarities are computed between each embedding and all subcenters of the ground-truth class, and the highest similarity is used in the margin-based softmax formulation. We use an angular margin of 28.6 degrees (converted to radians), a scale factor of 64, and three subcenters per class, consistent with prior work.

We further strengthen embedding quality through the use of batch hard mining (Hermans et al., 2017), which selects the hardest positive and hardest negative examples within each batch to drive training. This encourages the model to focus on challenging distinctions between individuals. In addition, we replace standard global pooling with Generalized Mean (GeM) pooling (Radenovic et al., 2019), a parameterized pooling operation that generalizes both average and max pooling by raising feature values to a learnable exponent before aggregation. GeM pooling has been shown to produce more informative and spatially aware embeddings, particularly beneficial for fine-grained visual tasks such as individual recognition.

A model was trained separately for each dataset using the contrastive learning approach described above. Training and evaluation involved generating lower-dimensional embeddings for each image and assessing how well these representations grouped semantically similar individuals. Because the training objective encourages samples from the same individual to cluster together in embedding space, evaluation was conducted using precision at 1 (P@1), nearest centroid classification (NCC) accuracy, and adjusted mutual information (AMI).

Precision at 1 (P@1) quantifies the proportion of times the nearest sample in embedding space shares the same identity as the query image. Nearest centroid classification evaluates how well the embedding of a test image corresponds to the centroid of its class, where each centroid is computed as the mean embedding of that class in the training set. This method reduces memory and computation requirements, as only one centroid per class must be stored to enable fast and scalable classification. NCC accuracy reflects the proportion of correctly assigned test samples based on proximity to these centroids.

Finally, adjusted mutual information (AMI) measures the agreement between the predicted clustering structure and the ground-truth labels. It corrects for chance using the expected mutual information under a random model. AMI scores range from −1 to 1, with 1 indicating perfect agreement, 0 reflecting chance-level clustering, and negative values denoting anticorrelated cluster assignments. Scores for both P@1 and NCC range from 0 to 1, with 1 indicating perfect performance.

4. Experiments

The experimental setup has two primary aims. The first is to show the effectiveness and generalizability of the model pipeline and to provide concrete and actionable methods for the practitioner which focus on efficiency as well as efficacy. The second is to show how known characteristics of a population, such a sexual dimorphism in killer whales, can impact the data viability in large image corpora.

4.1. Pipeline evaluation

The training and evaluation of the pipeline focuses on providing actionable data for the practitioner. This includes determining the optimal input image size for object detection, evaluating strategies for combining datasets, assessing the generalizability of image quality classification, and analyzing the impact of incorporating an ‘other’ orientation class. Each module except for the final identification module was trained on the BKW dataset. Evaluation of the modules for domain-relevance and identifiability were also performed using the additional cetacean groups, with each image in those datasets being assumed to be both domain-relevant as well as identifiable, as they were pre-selected for these qualities.

For object detection, three different input image sizes are used for both dorsal fin and eye patch processing, namely 416x416px, 640x640px, and 1280x1280px. This is done to provide empirical evidence for detection quality as it relates to varying sizes of target features. It is important to note the difference in relative sizes of regions of interest between eye patches and dorsal fins, the latter of which tend to be more prominent and larger in a given image. The difference in input image size plays a significant role not only in final performance but also required hardware overhead, with the use of larger images requiring more significant hardware resources not only for training but also for end-goal usage. The impact of including ‘negative’ samples, images without targets of interest, was also investigated. Furthermore, in order to investigate how the detection models perform when both classes of interest are together in the same training data (dorsal fin and eye patch),

a combined model was trained and evaluated. By doing so it is shown what this tradeoff is, if any, and may allow for further reduction in computation by utilizing one fewer model in the pipeline.

The best-performing fin detection model was then used to extract the dorsal fins from the additional cetacean groups, which were then used for evaluation of quality, as mentioned above, as well as for training their own identification models.

4.2. Data partitioning

To explore how identification performance evolves with long-lived, socially dynamic individuals, experiments were conducted by organizing the data into biologically meaningful cohort splits from the BKW dataset. Specifically, the impact of age and sex—given known patterns of sexual dimorphism and visual development in killer whales—was examined. The dataset was divided into four cohorts:

- SD-10-A/J: Individuals aged above and below 10 years, assessing identification performance based solely on age.
- SD-A/J: Adults and juveniles defined by both age and sex, where adulthood was defined as 10 years for females and 20 years for males (Towers et al., 2019).

A baseline model trained on the full dataset without cohort separation was also included. These experiments provide insight into whether expert-driven stratification can improve model performance and guide decisions on whether and how to exclude or filter data based on age- or sex-related characteristics. An overview of these datasets is provided in Table 6.

Modeling age as a continuous variable combined with sex in a learnable feature embedding was also explored; however, no statistically significant improvement in predictive performance was observed. Therefore, the threshold-based approach was retained due to its interpretability and ease of comparison across groups.

5. Results

5.1. Detection of regions of interest containing identifying features

- The results of the dorsal fin detection experiments, summarized in Table 7, demonstrate that the model achieves its highest performance using the smallest tested input size of 416 × 416 pixels. Notably, including background (i.e., empty) images in the dataset degrades performance across both mAP@50 and recall metrics for the 416px and 640px configurations, though precision sees a slight improvement. In contrast, the 1280px configuration shows marginal improvements across all metrics when background images are included. However, these gains still fall short of the average performance achieved with the 416px setting, and come with increased inference time due to the larger image resolution.

Similarly, the eye patch detection results in Table 8 reveal that the 640px configuration—without background images—consistently outperforms other settings across nearly all metrics. While the 1280px version trained with background images slightly surpasses the 640px configuration in mAP@50–95, it underperforms in precision and recall, both of which are more critical for accurate downstream use.

The disparity in optimal image size becomes even more pronounced when detecting both dorsal fins and eye patches in a single model. As shown in Table 9, performance declines across all metrics for every image size. This drop may result from the substantial size difference between the features, with eye patches being considerably smaller and more difficult to detect than dorsal fins. This scale imbalance is a well-documented issue in object detection literature and motivated the exploration of task-specific models (Hua and Chen, 2025). Additionally, the imbalance in the training data, where dorsal fins appear nearly twice

Table 7
Detection performance for fin-only images (FIN) and a combined dataset including fin and background images (FIN+B), evaluated across three input image sizes.

Data	Size [px]	mAP@50	mAP@50–95	Precision	Recall
FIN	416	0.9716	0.6355	0.9593	0.9363
FIN+B	416	0.9655	0.6258	0.9647	0.9233
FIN	640	0.9697	0.6301	0.9601	0.9311
FIN+B	640	0.9649	0.6271	0.9623	0.9258
FIN	1280	0.9538	0.5954	0.9526	0.9091
FIN+B	1280	0.9607	0.6053	0.9594	0.9121

Including background images slightly decreases performance, particularly at higher resolutions.

Table 8
Detection performance for eye patch-only images (EYE) and a combined dataset including eye patch and background images (EYE+B), across varying input image sizes.

Data	Size [px]	mAP@50	mAP@50–95	Precision	Recall
EYE	416	0.9701	0.6226	0.9626	0.9274
EYE+B	416	0.9589	0.6424	0.9559	0.9147
EYE	640	0.9932	0.6633	0.9871	0.9758
EYE+B	640	0.9786	0.6736	0.9697	0.9613
EYE	1280	0.9638	0.659	0.9798	0.9465
EYE+B	1280	0.9779	0.6678	0.9682	0.9616

Overall performance improves with larger input size, but including background images slightly reduces precision.

Table 9
Detection performance when combining both fin and eye patch detections in a single model (FIN + EYE), across different input sizes.

Data	Size [px]	mAP@50	mAP@50–95	Precision	Recall
FIN + EYE	416	0.9135	0.5983	0.9077	0.8394
FIN + EYE	640	0.9272	0.6328	0.8829	0.8739
FIN + EYE	1280	0.8291	0.5524	0.8656	0.7844

Performance is generally lower than the individual models, reflecting increased task complexity.

Table 6
An overview of the datasets used for sexual dimorphism cohort experiments.

Dataset	Total Images	Total Individuals	Median Img/ID	Mean Img/ID	Std Dev Img/ID	Min Img/ID	Max Img/ID
ALL - Baseline	81,463	269	258	302	252	5	1302
Sexual Dimorphism - Adults (SD-A)	48,354	199	156	242	216	8	1010
Sexual Dimorphism - Juveniles (SD-J)	33,109	144	260	290	232	4	1174
Above 10 Y.O. (SD-10-A)	63,406	231	178	274	258	8	1302
Below 10 Y.O. (SD-10-J)	18,057	92	177	196	135	4	583

The baseline shows a total number of 269 individuals overall. The number of individuals and number of photos in each dataset is determined by the age of each individual from this baseline dataset at the time of the creation of the datasets. Therefore, as this dataset represents over a decade, it is the case that some individuals span both adult and juvenile datasets for each of the cohorts. However, each dataset is a subset of the overarching baseline dataset. For each experiment the data was split using 70 % / 15 % / 15 % for training, validation, and testing, respectively.

as often as eye patches, may cause the model to prioritize fin detection, further degrading its ability to accurately locate eye patches.

5.2. Quality evaluation and filtering analysis

Table 10 reports the statistics on the test set for the models trained on the BKW data for determining if a cropped image of what is supposed to be a dorsal fin is actually relevant to the domain as well as if it is identifiable. While the models are all trained on the BKW data, they are evaluated on a number of different species all identifiable by their dorsal fins. This is evidenced by the fact that the performance degrades as the dorsal fin in question becomes less like the training material. For instance, in terms of both identifiability and domain relevance, the SRKW results show top performance across all reported metrics. This is likely due not only to the similarity of the two populations but also because the images used for evaluation within the SRKW data are all high-quality photos which are already prepared to be used for identification purposes. Similarly, the results for domain relevance are high for MW data, as their dorsal fins resemble those of killer whales. However, the scores for identifiability of the MW data suffer somewhat, possibly due to the relative size of the dorsal fin and the lack of saddle patch. This trend continues for the LBD, which, at least in terms of shape, displays many similarities to the killer whale and therefore performs well for domain relevance, but not so for identifiability. Again this is likely due to the same reasons as for the MW data. Finally, displaying the largest contrast to the killer whale, the HW performance degrades significantly in both domain relevance and identifiability.

The precision-recall (PR) and receiver operating characteristic (ROC) curves for the training data (BKW) are presented in Fig. 7, while the recall values for varying thresholds for the evaluation datasets are presented in Fig. 8.

The final module evaluates whether the image shows the left or right side of the individual, with results shown in Table 11. When only clear left or right-side images are included, performance is nearly perfect. However, introducing ambiguous images, such as those captured from the front or back, causes significant drops in accuracy, precision, and recall.

This degradation highlights the importance of module sequencing. Ideally, such ambiguous images should be filtered out by earlier modules before reaching this stage of the pipeline.

5.3. Individual identification

Table 12 presents the results of the final identification experiments across for each group or species included in this study. Despite the long-tailed distribution of the BKW dataset, the metric learning approach proves effective at identifying the majority of individuals, even when as few as three images per individual are available. This is supported by the performance on the CMW dataset, which, while not long-tailed, contains between 11 and 103 images per individual for training.

Table 10
Summary of the results of the experiments on determination of quality-related features domain relevance and identifiability.

Dataset	Task	Accuracy	Precision	Recall
BKW	Domain	0.9818	0.9818	0.9818
	Identifiable	0.9238	0.9238	0.9238
SRKW	Domain	0.9997	0.9997	0.9997
	Identifiable	0.9936	0.9936	0.9936
LBD	Domain	0.9791	0.9791	0.9791
	Identifiable	0.9081	0.9081	0.9081
CMW	Domain	0.9917	0.9917	0.9917
	Identifiable	0.8147	0.8147	0.8147
HW	Domain	0.6310	0.6310	0.6310
	Identifiable	0.7480	0.7480	0.7480

Each model was trained on the BKW dataset and evaluated on the others.

Strong performance is also observed for both Lahille's bottlenose dolphins and southern resident killer whales across all metrics. This is likely due to more careful dataset curation and the relatively small number of individuals in these groups compared to the BKW population. The metric learning approach is also effective for re-identifying humpback whales using only dorsal fin imagery. This offers a viable alternative to more traditional fluke-based identification methods (Blount et al., 2022; Cheeseman et al., 2022, 2024; Ranguelova et al., 2004).

A comparison to FIN-PRINT (Bergler et al., 2021) further highlights the effectiveness of the metric learning approach. A model was trained on the same most commonly seen individuals using the same dataset and data split referred to as KWID11–17 in that study. Evaluation on the same 2018 test data yielded an average accuracy of 0.943 with the Nearest Centroid Classifier and a precision@1 of 0.937, compared to the FIN-PRINT pipeline's accuracy of 0.828.

Using the expanded KWIDE11–17 dataset and the same 2018 evaluation split, this approach achieved an NCC accuracy of 0.886 and a P@1 of 0.864, compared to an accuracy of 0.845 in the original approach. While the NCC accuracy and traditional accuracy metrics are not directly comparable, they offer insight into the relative effectiveness of the two identification strategies. These results are summarized in Table 13.

5.3.1. Embedding visualizations and cluster analysis

Further analysis was conducted using Uniform Manifold Approximation and Projection (UMAP) to visualize the embedding space of 20 randomly selected individuals from the BKW dataset, with up to 1000 samples per individual. Intra- and inter-class distances were computed to assess embedding separation, and their distributions are illustrated in Fig. 9. The intra-class distance distribution had a mean of 0.259 and a standard deviation of 0.297, while the inter-class distance distribution had a mean of 1.00 and a standard deviation of 0.066. These results indicate that the learned embeddings are generally well separated across individuals, although some overlap remains in challenging cases where individual differentiation is more difficult. This degree of separation, as well as the associated difficulties, are further shown in Fig. 10.

5.4. Cohort analysis

Table 14 summarizes the results of the cohort analysis conducted on the BKW population. As expected, identification performance is generally lower for younger individuals compared to adults. This likely reflects the ongoing morphological changes in developing individuals, particularly in males. However, incorporating sexual dimorphism into the cohort definitions improves identification performance for juvenile males and females compared to using an age-based threshold alone.

Adult performance remains similar across both cohort definitions, indicating that accounting for sexual dimorphism does not negatively affect identification performance in this group. This suggests that distributing individuals based on biological factors rather than only on age can lead to more robust identification outcomes.

5.5. Continuous quality scoring for dataset evaluation and triage

The results in Table 15 show the performance of the quality evaluation module. Across all input image sizes, using the intermediate embedding representation instead of the final logit output improved both the coefficient of determination (R^2) and mean squared error (MSE), indicating better predictive performance. Higher input resolutions also led to improved model accuracy, with the best performance ($R^2 = 0.799$, MSE $= 0.362$) achieved using 640 $\times$ 640 px inputs and embedding features.

5.6. Runtime considerations

A major concern surrounding the use of state-of-the-art deep learning

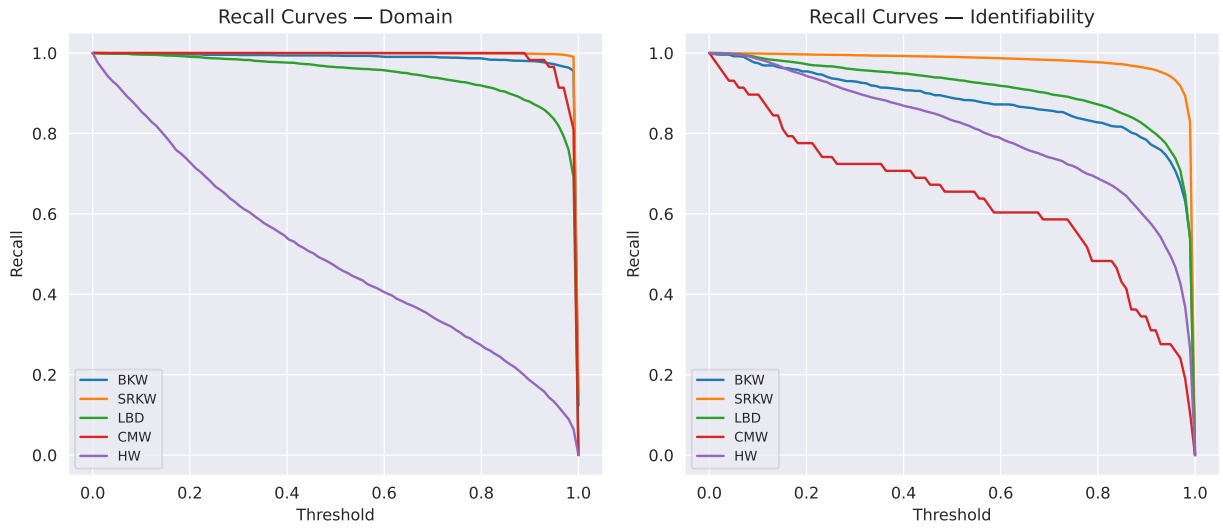


Fig. 7. Precision-Recall (left) and Receiver Operating Characteristic (ROC) curve (right) for the Domain Relevance and Identifiability Classification tasks. The classifier was trained and evaluated using the Bigg's killer whale (BKW) dataset. The PR curve highlights the model's ability to identify relevant images under varying classification thresholds, while the ROC curve illustrates the trade-off between the true positive rate and false positive rate.

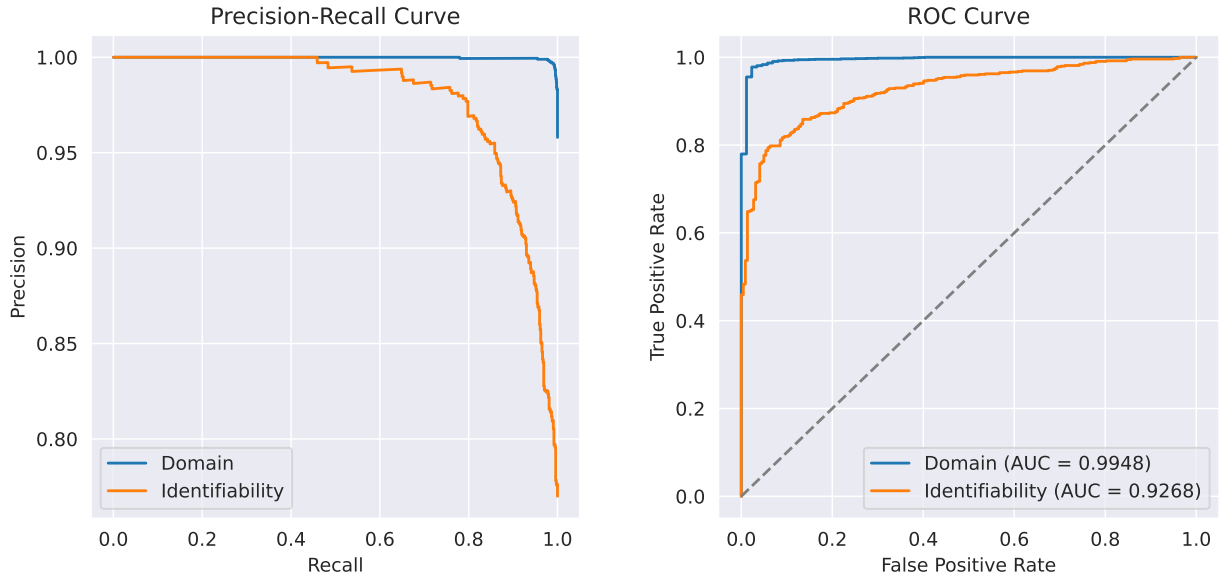


Fig. 8. Recall curves for the Domain Relevance (left) and Identifiability (right) classification tasks, plotted across all five species-specific datasets: Bigg's killer whales (BKW), Southern Resident killer whales (SRKW), Lahille's bottlenose dolphins (ELP), Common minke whales (Minke), and Humpback whales (HW). Each curve illustrates how recall varies with classification threshold, where the positive class is defined as domain-relevant or identifiable, respectively. Consistently high recall across a wide threshold range indicates strong sensitivity of the classifiers across species.

Table 11
The results of the experiments for determining which side of the individual is displayed.

Dataset	w/ Other	Accuracy	Precision	Recall
BKW	No	0.9918	0.9918	0.9918
	Yes	0.9097	0.9097	0.9097

Either there exist only left and right-side images within the data (w/Other: No), or additional images are also present which cannot readily be described as left or right, such as from the front or back of the animal (w/ Other: Yes).

models is their computational demand and the speed of inference. To address this, we conducted a series of experiments to evaluate the performance of our pipeline under different configurations and hardware setups. In each case, the same set of 10,000 images was used for

Table 12
A summary of the results for the final identification step of the pipeline across all tested groups / species.

Dataset	NCC Accuracy	P@1	NMI
BKW	0.9169	0.8689	0.8459
SRKW	0.9575	0.9534	0.9085
LBD	0.9646	0.9605	0.907
CMW	0.8234	0.7434	0.5889
HW	0.8542	0.7934	0.6694

Each model was trained only on data from that group. The nearest centroid classifier accuracy (NCC Accuracy), as well as the nearest neighbor precision (P@1), and the normalized mutual information (NMI) of each group of experiments are reported. The results are an average across five model training runs for each group.

Table 13
Comparison of the FIN-PRINT approach (Bergler et al., 2021) with the method presented in this study, evaluated on the same 2018 test set described in the FIN-PRINT manuscript.

Data	Accuracy (FIN-PRINT)	NCC Accuracy	P@1
KWID11–17	0.828	0.943	0.937
KWIDE11–17	0.845	0.886	0.864

Although the evaluation protocols differ, FIN-PRINT uses probability-based classification while the approach presented here uses nearest centroid classification (NCC) and precision at rank 1 (P@1), the results provide a general indication of the relative performance of the two systems under identical training and testing conditions.

evaluation. These images span a wide temporal range, reflecting variability in photographer style, location, and technological advancement, particularly the increase in image resolution over time due to improvements in sensor quality and storage capacity.

The pipeline was tested on two systems: a standard laptop equipped with an Intel i7 CPU and a high-performance computing (HPC) cluster utilizing a single Tesla V100 GPU (16GB). On the CPU-only system, the pipeline achieved between 3.2 and 1.6 frames per second (FPS), depending on the configuration. The simplest configuration includes only the essential steps for identification: fin detection, embedding generation, and nearest-centroid classification. The most complex configuration (and therefore includes all optional components, such as full-scene image scoring and eye patch detection.

In contrast, GPU acceleration significantly improved performance, with the pipeline achieving between 23.3 and 9.6 FPS on the same dataset. These results demonstrate that the pipeline is not only viable on modest hardware but also performant enough to process large datasets in reasonable time frames. Moreover, GPU acceleration enables near real-time processing, making the system suitable for both large-scale archival tasks and real-time applications. These results are summarized in Table 16.

5.7. Ablation study

To contextualize the role of individual pipeline components, a simplified ablation study was conducted. A subset of possible configurations was omitted to preserve clarity, with the selected experiments considered sufficient to illustrate the contributions and interactions of the key modules. This series of experiments examined several factors hypothesized to influence final identification performance, including: (1) whether "empty" images were included during fin detection training (FIN vs. FIN+B), (2) the input resolution used during detection (416 px, 640 px, or 1280 px), (3) the presence or absence of domain relevance filtering, and (4) the presence or absence of identifiability filtering. The identification module remained unchanged across all conditions; only the data source and pre-identification filtering stages were varied. The effects of these design choices are summarized in Table 17, which highlights the impact of each processing stage on both the quantity of available data and downstream identification accuracy. Again, these results are the average across 5 independent training runs.

Given that this ablation focuses on pre-identification filtering, the primary metrics reported are: (1) the proportion of images removed at each filtering stage, and (2) the performance of the final identification step, measured by nearest centroid classification (NCC) accuracy and precision@1. To ensure comparability across configurations, the classification classes and hardware resources were held constant throughout all runs.

6. Discussion

This work presents a fully automated pipeline to process and curate large image volumes for cetacean photo-ID. It includes detecting identifying features such as dorsal fins and eye patches, extracting regions of interest, evaluating image quality and identifiability, and generating hypotheses about individual identity and body position. The modular format and focus on minimal architecture enables these components to operate independently or in sequence, adapting to different data

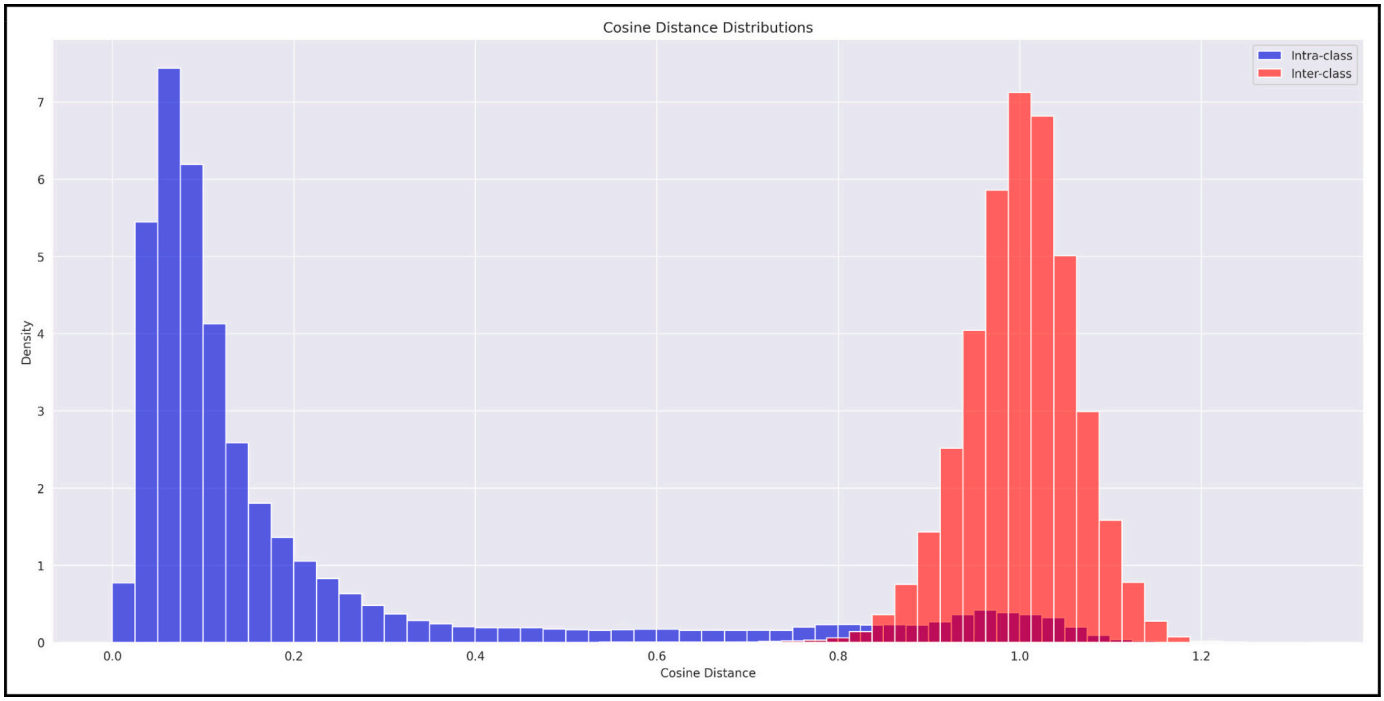


Fig. 9. Histogram of intra-class (blue) and inter-class (orange) pairwise distances for learned embeddings of 20 randomly selected individuals from the BKW dataset (up to 1000 samples per individual). The intra-class distribution (mean = 0.259, SD = 0.297) shows tighter clustering compared to the inter-class distribution (mean = 1.00, SD = 0.066), indicating good separation between individuals with some overlapping cases. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

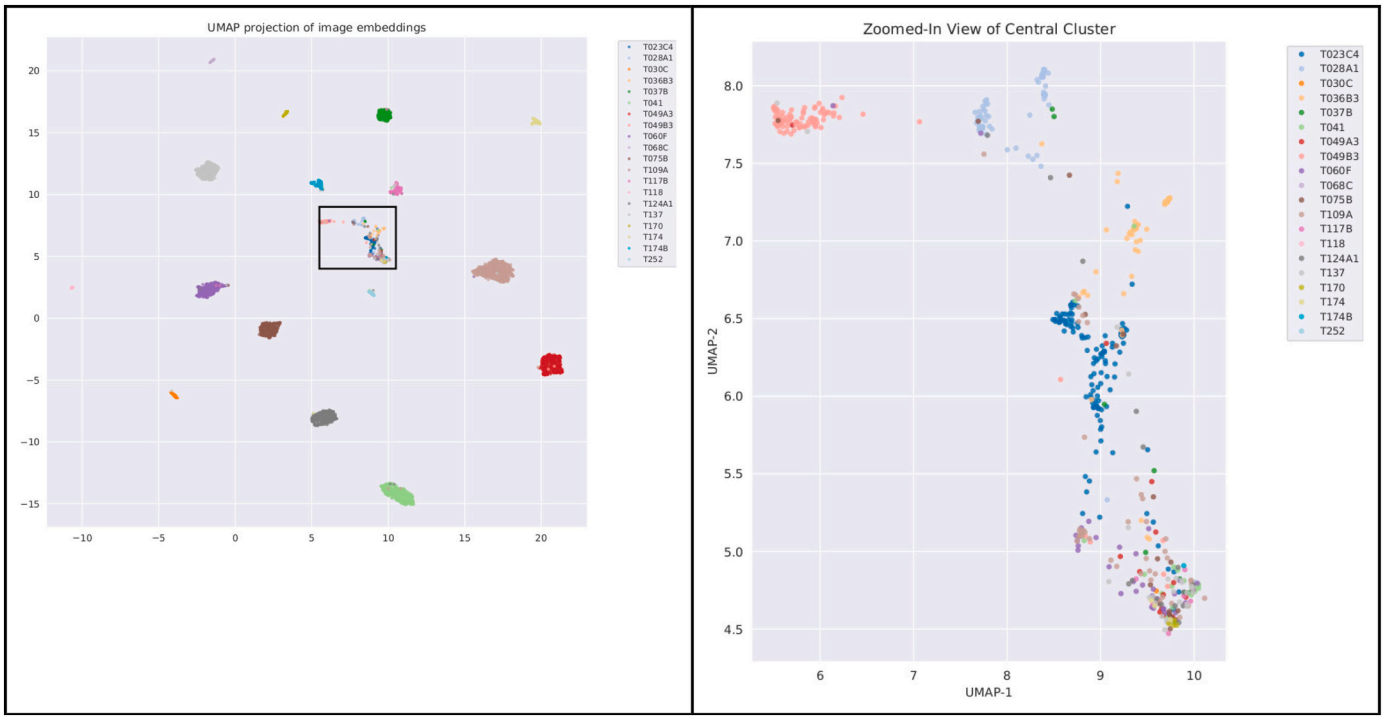


Fig. 10. UMAP projection of learned embeddings for 20 randomly selected individuals from the BKW dataset (up to 1000 samples per individual). The left panel shows the full embedding space, where most individuals form distinct clusters. The right panel provides a zoomed-in view of a central region, highlighting areas of overlap and increased inter-individual similarity, which may represent more challenging identification scenarios.

Table 14
The results from the cohort analysis for the BKW data.

Cohort	NCC Accuracy	P@1	NMI
All (Baseline)	*0.9272*	*0.92*	*0.8645*
Above 10 y.o.	0.9343	0.9327	0.8795
Below 10 y.o.	0.8773	0.8552	0.8064
Adult	0.9334	0.9361	0.8815
Juvenile / Calf	0.8922	0.8658	0.8204

The baseline reflects a model trained on the results for which the birth year and sex are known. For the individual cohorts, only those images are included during training which align with the cohort limits. The individuals are the same across all cohort boundaries to avoid attempting to identify an individual which has not been seen before. The cohorts involving juveniles, either Below 10 y.o or Juvenile / Calf are defined as being strictly below the threshold, meaning 0–9 years old for the first cohort and 0–9 years old for females and 0–19 years old for males in the second cohort. The nearest centroid classifier accuracy (NCC Accuracy), as well as the nearest neighbor precision (P@1), and the normalized mutual information (NMI) of each group of experiments are reported. The results are an average across five model training runs for each group.

Table 15
Performance of the image quality evaluation module across input sizes and feature types.

Input image Size	Quality Representation	R^2	MSE
416	Logit	0.650	0.632
416	Embedding	0.689	0.598
640	Logit	0.744	0.468
640	Embedding	0.799	0.362
1280	Logit	0.717	0.488
1280	Embedding	0.763	0.411

Reported metrics include the coefficient of determination (R^2) and mean squared error (MSE). Results show that embedding-based representations consistently outperform raw logits, and that moderate increases in input size (up to 640px) improve predictive accuracy.

Table 16
Runtime results from each step of the pipeline over an average of 10,000 full-scene images.

	CPU (Intel Core i7-1255U)		GPU-Accelerated (Tesla V100-SXM2-16GB)	
Operation	Mean [ms]	Median [ms]	Mean [ms]	Median [ms]
Fin Detections	280	342	20	15
Fin Extractions	1	2	0	0
Feature Detections	269	318	27	26
Feature Extractions	0	0	0	0
Domain Checks	20	6	17	4
Identifiability Checks	20	7	16	2
Embedding Generation	21	6	16	4
Centroid Classification	12	9	7	5
Quality Check	1	0	1	0

The mean and median time in milliseconds (ms) are reported for each main module in the pipeline.

volumes and hardware constraints. The pipeline performs well across several cetacean species, highlighting its generalizability. Importantly, we show that contrastive learning can mitigate the challenges posed by long-tailed data distributions common in ecological datasets, without the need for manual rebalancing. Furthermore, we provide a method of data triage by way of our continuous quality evaluation scoring method, which does not rely on data manually labelled with identifying information. The results of the ablation study also show the impact of the fin quality evaluation modules on downstream identification tasks, further highlighting the importance of these modules and the structure of our pipeline. Here it is to be noted that each quality evaluation method after the initial object detection increases downstream identification performance, regardless of original resolution used by the detection mechanism. Also notable is that the object detection mechanisms consistently overestimated the number of detected fins, thereby necessitating some

Table 17
The results from an ablation study showing the effect of various pipeline components on performance.

Trial	Input Size	Detection Data	Domain Check	ID-able Check	Data Change	NCC Acc.	P@1
1	416	FIN+B	−	−	+ 0.121	0.800	0.785
2	416	FIN	−	−	+ 0.155	0.789	0.776
3	416	FIN+B	+	−	+ 0.094	0.815	0.792
4	416	FIN	+	−	+ 0.115	0.812	0.784
5	416	FIN	−	+	− 0.096	0.905	0.864
6	416	FIN+B	−	+	− 0.103	0.905	0.856
7	416	FIN+B	+	+	− 0.106	0.906	0.863
8	640	FIN+B	−	−	+ 0.108	0.825	0.808
9	640	FIN+B	−	+	− 0.104	0.909	0.868
10	640	FIN+B	+	+	− 0.106	0.917	0.869
11	640	FIN	+	+	− 0.103	0.911	0.867
12	1280	FIN+B	−	−	+ 0.110	0.811	0.788
13	1280	FIN+B	+	−	+ 0.089	0.835	0.811
14	1280	FIN+B	−	+	− 0.081	0.901	0.845
15	1280	FIN+B	+	+	− 0.084	0.911	0.859
16	1280	FIN	+	+	− 0.028	0.905	0.849

Each row corresponds to a trial run with different configurations of the detection data, image input size, domain relevance and identifiability checks. Detection Data indicates whether the object detector was trained with images that include no detectable individuals (FIN + B) or whether every image in the object detection dataset contained target(s) of interest (FIN). Data Change refers to the relative change in number of images retained after detection and filtering. NCC Acc. refers to nearest-centroid classification accuracy, and P@1 denotes precision at rank 1. Note: The results in trial 10 correspond to the BKW results in Table 12.

kind of filtering mechanism to remove what must be erroneous detections. While we did not evaluate on a held-out geographic region or camera type, the natural diversity of our dataset, in that it spans many photographers, locations, and conditions, serves as a strong proxy for cross-distribution generalization.

Our approach builds on prior work in deep learning for wildlife identification (e.g., Bergler et al., 2021; Bogucki et al., 2019; Gore et al., 2016; Hou et al., 2020; Patton et al., 2023), extending existing systems in several ways. While most systems focused on single-species applications, we demonstrate transferability across species with diverse identifying features. For example, Miele et al. (2021) applied deep metric learning to giraffe identification using coat patterns, whereas our system handles a broader range of visual cues, such as fin shape and eye patch morphology, common to multiple cetacean species. Similarly, FINPRINT (Bergler et al., 2021) developed a robust pipeline for killer whale identification. Our work builds on this by incorporating body position prediction and image relevance evaluation, as well as investigating the effect of introducing age and sex cohorts to training data, enabling improved image triaging and prioritization in large ecological datasets.

In shark identification, studies such as Le et al. (2022) employed segmentation and embedding strategies tailored for individual recognition. While effective within their domain, such approaches often require computationally intensive models (e.g., VGG-based networks) and are less adaptable to edge deployment. In contrast, our use of compact architectures like YOLOv8-Small and EfficientNet-B0 balances performance and resource efficiency, enabling field deployment in resource-limited settings.Contour-based identification methods, such as those used by Hughes and Burghardt (2017) for white sharks, have also proven effective, particularly when dorsal fin trailing-edge marks are distinct and persistent. However, these approaches are species-specific and may not generalize well to taxa where flank patterns or saddle patches carry more identifying information. Our pipeline is designed to accommodate such diversity, offering a unified framework that can be adapted to a variety of marine mammal morphologies.

While this approach shows strong performance in identifying known individuals, and while the pipeline has functions outside of straightforward identification (e.g. large dataset triage), several limitations remain. First, as mentioned earlier, the system assumes a closed-world setting and does not explicitly detect or handle unknown individuals. Second, juveniles, whose morphological features can more readily change over time, can pose challenges for consistent identification. Third, while orientation detection is used to disambiguate left and right fin images, misclassification or ambiguous viewpoints can still introduce noise into the learned embedding space. These issues and ambiguities should be addressed in future work.

The pipeline's modular design, centered around a single detection step (typically dorsal fin detection), allows practitioners to customize processing based on available computational resources and specific project needs. For instance, lightweight deployments could run only the detection and identification modules, while cloud-based systems could include full image quality assessment and quality-characteristic analysis.

Looking forward, we plan to further compress and optimize the pipeline, including through weight sharing and transfer learning, to minimize hardware requirements while preserving accuracy. These enhancements will support prioritization in monitoring workflows and improve training dataset curation. Finally, this pipeline can be easily extended to other taxonomic groups, including terrestrial and avian species, thereby broadening its utility for conservation efforts across diverse ecosystems.

7. Conclusion

This study presents a comprehensive pipeline for automated species identification, demonstrated through experiments on marine mammals, especially killer whales. We developed models for object detection (fin and eye patch), quality assessment (image usefulness, identifiability, and side classification), and individual identification using contrastive learning combined with nearest centroid classification. These components address the critical need to scale ecological data curation with accuracy and efficiency.

Our motivation was to reduce manual effort in species identification workflows and democratize access to these tools. To that end, the pipeline was fully integrated into finwave (www.finwave.io), an open-access platform enabling researchers and citizen scientists to contribute data and apply identification tools without requiring deep machine learning expertise.

Importantly, our runtime analysis shows that the pipeline operates near real-time on GPU hardware, while still maintaining functionality on simpler, more accessible computing devices. This balance of speed and deployability ensures broad usability in diverse field and laboratory settings.

By providing configurable code and deployment scripts for URI-based model access, we support easy integration into existing workflows across multiple species and research contexts.

Overall, this pipeline and its deployment on finwave offer a scalable, efficient, and accessible solution that lowers technical barriers,

accelerates conservation research, and supports global population monitoring efforts. By combining scalable deep learning methods with ecological insight, this study contributes a practical, extensible tool for individual recognition and image management in photo-ID research. It supports long-term population monitoring, enhances the quality of ecological data pipelines, and advances the role of machine learning in biological sciences.

CRediT authorship contribution statement

Alexander Barnhill: Writing – original draft, Writing – review & editing, Visualization, Methodology, Project administration, Formal analysis, Investigation. **Jared R. Towers:** Writing – review & editing, Supervision, Data curation. **Tasli J.H. Shaw:** Writing – review & editing, Data curation. **Magdalena Arias:** Writing – review & editing, Data curation. **Adrián Bécares:** Methodology, Software, Investigation. **Thomas Doniol-Valcroze:** Writing – review & editing, Data curation. **Lorenzo von Fersen:** Writing – review & editing, Data curation. **Rodrigo Genoves:** Writing – review & editing, Data curation. **Tim Rörup:** Methodology, Software, Investigation. **Gary J. Sutton:** Writing – review & editing, Data curation. **Sheila Thornton:** Writing – review & editing, Data curation. **Michael Weiss:** Writing – review & editing, Data curation. **Andreas Maier:** Writing – review & editing, Supervision, Funding acquisition, Resources. **Elmar Nöth:** Writing – review & editing, Supervision, Funding acquisition. **Christian Bergler:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This analysis was supported by funding from Friedrich-Alexander-Universität Erlangen-Nürnberg. Data collection and curation for Bigg's killer whales was funded primarily by the Species at Risk Program at Fisheries and Oceans Canada with support from Bay Cetology. Data collection and curation for southern resident killer whales was funded by the Species at Risk Program at Fisheries and Oceans Canada and supported by the US National Marine Fisheries Service, Northwest Fisheries Science Center. We thank the many identification photo collectors for the data contributions which have helped make this analysis possible and two anonymous reviews for their suggestions to improve the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103396.

Data availability

The data used in this study were obtained from a wide range of contributors, including government agencies, research institutions, and independent organizations. Many of these contributors have shared data under specific agreements that restrict redistribution or public release. As such, the full dataset cannot be made publicly available due to legal and ethical considerations, including data ownership and privacy obligations to non-author contributors. However a representative dataset featuring 20 individuals from the BKW dataset, encompassing 500 photos, has been made available at zenodo.org/records/16422215. This dataset includes the features necessary to illustrate the modules detailed here and therefore includes the raw image data as well as information regarding individual ID as well as the sex of that individual and their age at the time the photograph was taken. The code necessary to reproduce the experiments listed here will be available at github.com/alexanderbarnhill/finprintv2.

References

Bendale, A., Boult, T., 2015. Towards open world recognition. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, pp. 1893–1902. Available at: https://doi.org/10.1109/CVPR.2015.7298799.

Bergler, C., et al., 2021. FIN-PRINT a fully-automated multi-stage deep-learning-based framework for the individual recognition of killer whales. Sci. Rep. 11 (1), 23480. Available at: https://doi.org/10.1038/s41598-021-02506-6.

Blount, D., et al., 2022. Flukebook: an open-source AI platform for cetacean photo identification. Mamm. Biol. 102 (3), 1005–1023. Available at: https://doi.org/10.1007/s42991-021-00221-3.

Bogucki, R., et al., 2019. Applying deep learning to right whale photo identification. Conserv. Biol. 33 (3), 676–684. Available at: https://doi.org/10.1111/cobi.13226.

Bouma, S., et al., 2018. Individual common dolphin identification via metric embedding learning. In: 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–6. Available at https://doi.org/10.1109/IVCNZ.2018.8634778.

Cheeseman, T., et al., 2022. Advanced image recognition: a fully automated, high-accuracy photo-identification matching system for humpback whales. Mamm. Biol. 102 (3), 915–929. Available at: https://doi.org/10.1007/s42991-021-00180-9.

Cheeseman, T., et al., 2024. Bellwethers of change: population modelling of North Pacific humpback whales from 2002 through 2021 reveals shift from recovery to climate response. R. Soc. Open Sci. 11 (2), 231462. https://doi.org/10.1098/rsos.231462.

Das, S.K., Roy, P., Singh, P., Diwakar, M., Singh, V., Maurya, A., Kumar, S., Kadry, S., Kim, J., 2023. Diabetic foot ulcer identification: a review. Diagnostics 13 (12), 1998. https://doi.org/10.3390/diagnostics13121998.

Deng, J., et al., 2019. ArcFace: additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4685–4694. Available at https://doi.org/10.1109/CVPR.2019.00482.

Deng, J., et al., 2020. Sub-center arcface: boosting face recognition by large-scale noisy web faces. In: computer vision–ECCV 2020: 16th European conference, Glasgow, UK, august 23–28, 2020, proceedings, part XI 16. Springer, pp. 741–757.

Dungan, S.Z., et al., 2016. Social structure in a critically endangered Indo-Pacific humpback dolphin (*Sousa chinensis*) population. Aquat. Conserv. Mar. Freshwat. Ecosyst. 26 (3), 517–529. https://doi.org/10.1002/aqc.2562.

Fu, Y., et al., 2022. Long-tailed visual recognition with deep models: a methodological survey and evaluation. Neurocomputing 509, 290–309. Available at: https://doi.org/10.1016/j.neucom.2022.08.031.

Genoves, R.C., et al., 2020. Fine-scale genetic structure in Lahille's bottlenose dolphins (*Tursiops truncatus gephyreus*) is associated with social structure and feeding ecology. Mar. Biol. 167 (3), 34. https://doi.org/10.1007/s00227-019-3638-6.

Gero, S., et al., 2014. Behavior and social structure of the sperm whales of Dominica, West Indies. Mar. Mamm. Sci. 30 (3), 905–922. Available at: https://doi.org/10.1111/mms.12086.

Gore, M.A., Frey, P.H., Ormond, R.F., Allan, H., Gilkes, G., 2016. Use of photo-identification and mark-recapture methodology to assess basking shark (Cetorhinus maximus) populations. PLoS One 11 (3), e0150160. https://doi.org/10.1371/journal.pone.0150160.

Hammond, P.S., Mizroch, S.A., Donovan, G.P., 1990. Individual Recognition of Cetaceans: Use of Photo-Identification and Other Techniques to Estimate Population Parameters. Reports of the International Whaling Commission (Special Issue, 12).

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. ArXiv, abs/1703.07737.

Hou, J., et al., 2020. Identification of animal individuals using deep learning: a case study of giant panda. Biol. Conserv. 242, 108414.

Hua, W., Chen, Q., 2025. A survey of small object detection based on deep learning in aerial images. Artificial Intelligence Review 58 (6). https://doi.org/10.1007/s10462-025-11150-9.

Hughes, B., Burghardt, T., 2017. Automated visual fin identification of individual great white sharks. Int. J. Comput. Vis. 122 (3), 542–557. Available at: https://doi.org/10.1007/s11263-016-0961-y.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics.

Jordaan, R.K., et al., 2020. Abundance, survival and population growth of killer whales *Orcinus orca* at subantarctic Marion Island. Wildl. Biol. 2020 (4), wlb.00732. https://doi.org/10.2981/wlb.00732.

Jourdain, E., et al., 2021. Killer whale (Orcinus orca) population dynamics in response to a period of rapid ecosystem change in the eastern North Atlantic. Ecol. Evol. 11 (23), 17289–17306. Available at: https://doi.org/10.1002/ece3.8364.

Le, N.A., et al., 2022. An automated framework based on deep learning for shark recognition. J. Mar. Sci. Eng. 10 (7), 942. https://doi.org/10.3390/jmse10070942.

Li, T., et al., 2022. Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6918–6928.

Manna, S., Ghildiyal, S., Bhimani, K., 2020. Face recognition from video using deep learning. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 1101–1106. https://doi.org/10.1109/ICCES48766.2020.9137927.

Matkin, C.O., et al., 2012. Contrasting Abundance and Residency Patterns of Two Sympatric Populations of Transient Killer Whales (Orcinus orca) in the Northern Gulf of Alaska. | EBSCOhost. Available at: https://openurl.ebsco.com/contentitem/gcd:76441623?sid=ebsco:plink:crawler&id=ebsco:gcd:76441623. (Accessed 13 March 2025).

McMillan, C.J., Towers, J.R., Hildering, J., 2019. The innovation and diffusion of "trap-feeding," a novel humpback whale foraging strategy. Mar. Mamm. Sci. 35 (3), 779–796. https://doi.org/10.1111/mms.12557.

Miele, V., et al., 2021. Revisiting animal photo-identification using deep metric learning and network analysis. Methods Ecol. Evol. 12 (5), 863–873. Available at: https://doi.org/10.1111/2041-210X.13577.

Mohammed, N.F., Ali, S.A., Jawad, M.J., 2020. Iris recognition system based on lifting wavelet. In: Mallick, P.K., et al. (Eds.), Cognitive Informatics and Soft Computing. Springer, Singapore, pp. 245–254. https://doi.org/10.1007/978-981-15-1451-7_27.

Nguyen, H., et al., 2017. Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring. In: 2017 IEEE international conference on data science and advanced Analytics (DSAA). IEEE, pp. 40–49.

Nielsen, M.L.K., et al., 2021. A long postreproductive life span is a shared trait among genetically distinct killer whale populations. Ecol. Evol. 11 (13), 9123–9136. Available at: https://doi.org/10.1002/ece3.7756.

Nielsen, M.L.K., et al., 2023. Temporal dynamics of mother–offspring relationships in Bigg's killer whales: opportunities for kin-directed help by post-reproductive females. Proc. R. Soc. B Biol. Sci. 290 (2000), 20230139. Available at: https://doi.org/10.1098/rspb.2023.0139.

Pan, W., Chen, J., Lv, B., Peng, L., 2025. Lightweight marine biodetection model based on improved YOLOv10. Alex. Eng. J. 119, 379–390. https://doi.org/10.1016/j.aej.2025.01.077.

Patton, P.T., et al., 2023. A deep learning approach to photo–identification demonstrates high performance on two dozen cetacean species. Methods Ecol. Evol. 14 (10), 2611–2625. Available at: https://doi.org/10.1111/2041-210X.14167.

Pollicelli, D., Coscarella, M., Delrieux, C., 2017. Wild cetacea identification using image metadata. J. Comput. Sci. Technol. 17 (01), 79–84.

Radenovic, F., Tolias, G., Chum, O., 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (7), 1655–1668. https://doi.org/10.1109/tpami.2018.2846566.

Ranguelova, E., Huiskes, M., Pauwels, E.J., 2004. Towards computer-assisted photo-identification of humpback whales. In: 2004 International Conference on Image Processing, 2004. ICIP '04. 2004 International Conference on Image Processing, 2004, 3. *ICIP '04*, pp. 1727–1730. https://doi.org/10.1109/ICIP.2004.1421406. Available at.

Ren, C.-X., Xu, X.-L., Lei, Z., 2019. A deep and structured metric learning method for robust person re-identification. Pattern Recogn. 96, 106995.

Saini, T., Gandhi, R., Roy, S., Diwakar, M., Singh, P., Mishra, A.K., 2024. Person re-identification using deep learning. In: Proceedings of the 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE). IEEE, pp. 509–512. https://doi.org/10.1109/AECE62803.2024.10911558.

Schofield, D., et al., 2019. Chimpanzee face recognition from videos in the wild using deep learning. Sci. Adv. 5 (9), eaaw0736. https://doi.org/10.1126/sciadv.aaw0736.

Shao, M., Peng, Z., 2024. Distance metric-based learning for long-tail object detection. Image Vis. Comput. 142, 104888.

Tixier, P., et al., 2016. Depredation of Patagonian toothfish (Dissostichus eleginoides) by two sympatrically occurring killer whale (Orcinus orca) ecotypes: insights on the behavior of the rarely observed type D killer whales. Mar. Mamm. Sci. 32 (3), 983–1003. Available at: https://doi.org/10.1111/mms.12307.

Towers, J.R., et al., 2013. Seasonal movements and ecological markers as evidence for migration of common minke whales photo-identified in the eastern North Pacific. J. Cetacean Res. Manag. 13 (3), 221–229. https://doi.org/10.47536/jcrm.v13i3.539.

Towers, J.R., et al., 2019. Movements and dive behaviour of a toothfish-depredating killer and sperm whale. ICES J. Mar. Sci. 76 (1), 298–311. Available at: https://doi.org/10.1093/icesjms/fsy118.

van Weelden, C., et al., 2025. Divergent killer whale populations exhibit similar acquisition but different healing rates of conspecific scars. Behav. Ecol. Sociobiol. 79 (3), 39. https://doi.org/10.1007/s00265-025-03576-6.

Yang, X., Wang, M., Tao, D., 2017. Person re-identification with metric learning using privileged information. IEEE Trans. Image Process. 27 (2), 791–805.

Yao, F., Zhang, H., Gong, Y., Zhang, Q., Xiao, P., 2025. A study of enhanced visual perception of marine biology images based on diffusion-GAN. Complex Intell. Syst. 11 (5). https://doi.org/10.1007/s40747-025-01832-w.

Zanardo, N., et al., 2018. Social cohesion and intra-population community structure in southern Australian bottlenose dolphins (*Tursiops* sp.). Behav. Ecol. Sociobiol. 72 (9), 156. https://doi.org/10.1007/s00265-018-2557-8.

Zanlorensi, L.A., et al., 2018. The impact of preprocessing on deep representations for Iris recognition on unconstrained environments. In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 289–296. Available at https://doi.org/10.1109/SIBGRAPI.2018.00044.

Zhou, L., Cai, J., Ding, S., 2023. The identification of ice floes and calculation of sea ice concentration based on a deep learning method. Remote Sens. 15 (10), 2663. https://doi.org/10.3390/rs15102663.

Zhu, J., et al., 2022. Balanced contrastive learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6908–6917.