Building a Flexible Framework for Automated White Shark Re-Identification

Fabrice Kurmann[1] Connor Pryor[1] Charles Dickens[1] Alexandra Eva DiGiacomo[2]
Samantha Andrzejaczek[2] Eriq Augustine[1] Barbara Block[2] Lise Getoor[1]

[1]Department of Computer Science and Engineering, UC Santa Cruz
[2]Hopkins Marine Station, Stanford University

{fkurmann, cfpryor, cadicken, eaugusti, getoor}@ucsc.edu
{alexandra.digiacomo, sammyaz, bblock}@stanford.edu

Abstract

Animal re-identification, the problem of mapping a new image to an existing curated set of individuals, is crucial for wildlife conservation and population monitoring. Traditionally, domain experts have performed this task manually, but identifying individuals from images is both challenging and labor-intensive. To address this challenge in monitoring white shark populations, we present an automated re-identification approach that assists in matching individuals while supporting human-in-the-loop validation. By training and evaluating on a dataset of dorsal fin images characterized by limited training data and a long-tailed distribution, we develop a system robust to the real-world challenges of image-based wildlife tracking. We leverage a pretrained Vision Transformer (ViT) backbone [8], which we efficiently adapt to produce discriminative shark fin embeddings that are robust to variations in pose, lighting, and image quality. We compare embedding retrieval strategies to optimize retrieval of the most relevant individuals. Our combined approach is situated in a user interface that allows researchers to update and grow their image catalogs, leveraging our re-identification system to deliver high-likelihood match recommendations they can evaluate and, upon approval, accept.

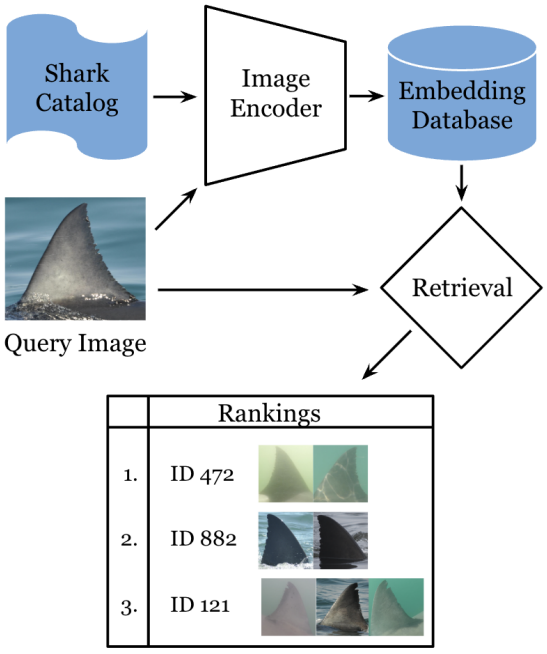

Figure 1. For white shark re-identification an image encoder trained on the contents of a shark catalog generates embeddings for all catalog and new query images. A retrieval algorithm is then used to produce ranked shark suggestions from the embedding database.

1. Introduction

Identifying individual animals from images and videos is essential for understanding population dynamics, tracking migration patterns, and informing conservation efforts. Automated re-identification offers a scalable alternative to labor intensive, highly specialized matching of animal images based on pictured individuals. As a result, the task of animal re-identification using computer vision has received immense attention [5–7, 12, 14–17, 19, 20, 23, 24, 26, 29]. Previous works have developed numerous methods, including applying various neural network architectures and training techniques to recognize and identify individuals based on specific biomarkers for a multitude of species.

In this paper, we specifically examine the problem of white shark identification from dorsal fin images extracted from video [1]. Visual matching in this context presents substantial challenges: due to the vastness of their habitat, limited visibility, and rarity of re-encountering the same individual, white shark sightings are rare and photographic

conditions are often poor due to water quality and lighting [18]. As a result, researchers can typically collect only a small number of quality photographs per individual to identify sharks across sightings. This leads to a sparsely populated database of many individuals, each represented by only a handful of images. As such a dataset grows, it becomes increasingly difficult to match new observations against past records. The matching process quickly becomes time-consuming and error-prone, limiting the scalability of long-term monitoring efforts.

We address these challenges and develop a framework for identifying individual white sharks from dorsal fin photographs leveraging a state-of-the-art vision transformer (ViT) architecture [8]. Figure 1 provides a high-level overview of our matching system, in which a trained image encoder generates embeddings for all catalog and query images and a retrieval algorithm then ranks candidate matches based on proximity to the query of individuals in the embedding database. Our model and retrieval system streamlines re-identification by presenting marine biologists with a ranked list of likely matching individuals for each query image, reducing the labeling workload while maintaining accuracy through targeted human review.

Our key contributions are: (1) training an image model optimized for our task by introducing class-aware triplet sampling and augmented images to improve performance with limited, sparsely distributed data; (2) evaluating two retrieval strategies, nearest-neighbor and prototype-based, for identifying individuals; and (3) demonstrating strong performance on an extensive dataset of dorsal fin images, achieving up to 51% Hits@1 and 91% Hits@50. Additionally, we introduce a shark matching platform which provides an interface to our framework where marine biologists can view and maintain their catalog of labeled shark images and efficiently process and match newly captured photographs. Our results and feedback from users highlight the promise of deep learning coupled with a human-in-the-loop interface to streamline animal re-identification in unique real-world marine monitoring efforts.

2. Related Work

The re-identification of individual animals is a cornerstone of wildlife conservation and ecological efforts, enabling population assessments, tracking movements, and understanding behaviors [1, 3, 10, 27]. Traditional photo-identification relies on manual comparison of distinctive natural markings, such as nicks, notches, scars, and pigmentation patterns [28], a process which is exceptionally time-consuming and prone to human error, especially as image databases grow. Consequently, computer-assisted and automated methods have emerged to accelerate and improve accuracy in individual animal re-identification.

For many large marine species, the dorsal fin or fluke serves as the primary feature for photographic identification. Early computational approaches focused on geometric representations of the fin's trailing edge. Finscan, for example, framed fin comparison as a string-matching problem based on edge curvature [12], while others later introduced an integral curvature representation to improve robustness to viewpoint changes [26]. Hughes and Burghardt introduced the first fully automated contour-based visual identification system for great white sharks from dorsal fin imagery, focusing specifically on matching unconstrained, natural images to animal identity without requiring manual intervention for fin detection and extraction [14]. More recent methods like finFindR have shifted towards deep learning, using Convolutional Neural Networks (CNNs) to extract features from sub-images along the fin [23], making the process more robust and less sensitive to image noise. Similarly, Moskvyak et al. presented a system for robust re-identification of manta rays and humpback whale flukes using pose-invariant embeddings learned via a deep CNN and triplet loss [16]. Finally, in the realm of end-to-end automated systems leveraging deep learning, FIN-PRINT stands out [5]. This framework for killer whale recognition incorporates dedicated stages for dorsal fin and saddle patch detection, extraction, data enhancement to filter invalid images, and multi-class individual classification. Our work departs from these contour and CNN-based methods by leveraging a Vision Transformer (ViT) architecture and moving beyond a primary reliance on exclusively the fin's edge or any one specific biomarker. We show that the ViT's self-attention mechanism is better suited to capture the subtle, holistic patterns of notches and pigmentation crucial for white shark re-identification, moving beyond a primary reliance on the fin's edge alone.

Beyond recognizing fin morphology, re-identification tasks for other species have broadened algorithm development. HotSpotter is a widely used algorithm based on extracting and matching keypoints or "hotspots" [7]. This approach emphasizes distinctive regions on an animal's pattern and has been successfully applied to zebras, giraffes, and jaguars, among others. For sharks with unique spotting patterns, van Tienhoven et al. introduced the Interactive Individual Identification System [24]. This computer-aided program compares natural pigment marks, relying on human input to highlight key pigmentation regions and then confirm matches from a ranked list. Finally, Zheng et al. proposed a transformer network structure for wild terrestrial animal re-identification based on fur, stripes, and facial features, addressing limitations of traditional CNNs in capturing long-distance feature relationships [29]. These successive works highlight a growing trend towards more advanced, less feature-specific deep learning architectures for both marine and general animal recognition.

A growing trend is the development of species-agnostic

models and comprehensive platforms. Some research focuses on modular pipelines that can be adapted to new species with minimal re-training [17], while others have trained single, large-scale networks on diverse, multi-species datasets [19]. By generalizing to multiple species, these frameworks address the challenge of limited labeled training data for many species and make promising steps in improving the reach of automated re-identification. While we train and evaluate our current model on white sharks exclusively, a promising future area of research involves exploring how our contributions are compatible to the study of multi-species frameworks.

The increasing complexity and volume of photographic data have also driven the development of comprehensive re-identification frameworks and platforms. Wildbook pioneered an open-source web platform that enables citizen scientists to participate in crowdsourcing conservation through its aggregation of species detection models paired with a robust database and web-based user interface [4]. Flukebook, a species specific Wildbook, [6] offers an to support the identification for 15 cetacean species using a human-in-the-loop approach pairing re-identification algorithms and expert judgment to confirm or reject potential matches. Building in parallel with these efforts, we introduce a shark matching platform that provides an interface to our framework, allowing marine biologists to view and maintain their catalog of labeled shark images and efficiently process and match newly captured photographs. Collectively, these systems illustrate the growing integration of automated identification technologies, balancing automation and expert oversight to meet the diverse needs of conservation biology.

3. Methods

Our goal is to construct a neural model that can produce discriminative image embeddings that are highly sensitive to biomarkers on the dorsal fin, such as unique notches and pigmentation patterns, while remaining invariant to confounding variables such as pose, lighting, and image quality. We then apply a retrieval technique to determine the individuals with the greatest likelihood of matching the subject of a query image.

Formally, let $\mathcal{X}$ denote the space of input images, and let $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ represent a dynamic catalog of known individual identities, each associated with a small gallery $\mathcal{G}_y \subset \mathcal{X}$ of associated images. Given a query image $x_q \in \mathcal{X}$, the objective is to identify the correct individual $y^* \in \mathcal{Y}$ such that x_q and the images in $\mathcal{G}_{y^*}$ depict the same individual. Crucially, this is formulated as a *retrieval problem*, rather than a classification problem. In many real-world ecological monitoring scenarios, including shark re-identification, the set of known identities $\mathcal{Y}$ is dynamic: new individuals are frequently discovered and added to the gallery over time. As a result, directly training a model to predict $\mathbb{P}(y \mid x_q)$ on a closed set of labels is infeasible and does not generalize to unseen individuals encountered after training. This task is therefore decomposed into two components: (1) learning an embedding function $f_\theta : \mathcal{X} \to \mathbb{R}^d$ that maps images to a representation space where those depicting the same individual lie close under a distance metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$; and (2) retrieving the identities y^* that best match the query in this space.

Various forms of neural architecture have been leveraged to produce image embeddings, with recent research having shifted towards transformer-based models, which have already shown great success in domains such as natural language processing (NLP). We begin by introducing foundational information about our neural model architecture (Section 3.1), then the following two subsections will describe our optimizations in model training (Section 3.2) and embedding retrieval (Section 3.3).

3.1. Model Architecture

Vision Transformer (ViT) models represent a paradigm shift in computer vision by directly applying the Transformer architecture, previously dominant in NLP, to image recognition tasks [8]. Unlike previous neural network-based computer vision approaches that typically rely on Convolutional Neural Networks (CNNs) that integrate attention mechanisms within convolutional structures [9], ViT models introduce a transformer exclusive architecture that eliminates the inherent reliance on convolution. Although ViT models lack intrinsic understanding of the locality achieved in CNNs through their convolutional structure, this apparent limitation is overcome through the ViT self-attention mechanism. When trained on very large-scale datasets, this self-attention mechanism is able to learn to distinguish and weight significant across any region within the image. The additional ability of ViT's self-attention layers to integrate information globally across the entire image, over all network layers, provides a key benefit over CNNs, whose receptive fields are heavily localized, especially in early network layers.

The core ViT innovation involves treating images as sequences of flattened 2D patches, analogous to how NLP Transformers process sequences of words or tokens. Specifically, an input image is reshaped into numerous equally sized patches, which are then linearly embedded to a lower dimension, forming "patch embeddings". These patch embeddings are combined with learnable position embeddings, which encode spatial information. This resulting sequence of embedding vectors is then fed into a standard transformer encoder, a modular stack of alternating perceptron and self-attention and blocks [25], which correspondingly function to understand patch features and the relationship between different patches. This modular setup enables ViTs to lever-

age the same scalable architecture and benefit from the efficient implementations developed for NLP Transformers.

A key aspect of ViT's exceptional performance and adaptability lies in their training paradigm, which involves extensive pretraining on vast, diverse datasets, optionally followed by fine-tuning for highly specific downstream tasks. When pretrained on sufficiently large datasets, ViT models achieve state-of-the-art results that match or surpass comparable CNNs. We selected ResNet34 [11] and ViT Large Patch16-384 [8] models, two CNN and ViT backbones, both of which were pretrained on large, ImageNet datasets, and compared their performance on our data set. Our evaluation showed a significantly higher accuracy with a ViT backbone; therefore, we focused our efforts on optimizing the performance of this model. To adapt our foundation model for our shark re-identification task, we append a multilayer perceptron (MLP) projection head, which is trained to produce the final embeddings from the ViT's output representations.

3.2. Model Training

Fully training a projection head without adapting our ViT foundation model limits performance on our domain-specific task. We therefore explored fine-tuning our ViT backbone. For computational efficiency, we used low-rank adaptation (LoRA) [13], and introduced 3.5M trainable parameters into the foundation model. These trainable parameters are then updated during fine-tuning while the foundation parameters remain frozen. By training only the LoRA trainable parameters, instead of the foundation model's full 307M, this parameter efficient approach enabled model convergence within 72 hours on a single NVIDIA Quadro RTX 6000 GPU. The combination of training the MLP projection head and fine-tuning the backbone via LoRA resulted in a marked improvement in the model's ability to generate discriminative embeddings, evidenced by significant gains in retrieval performance.

3.2.1. Triplet Loss

Our model aims to produce an embedding space where images of the same individual are grouped together and dissimilar images are pushed apart. We train using a triplet loss function [22] which operates on triplets of images: an "anchor" (A), a "positive" (P) (an image of the same individual as the anchor), and a "negative" (N) (an image of a different individual). The triplet loss is then formally defined as:

$$\mathcal{L}(A, P, N) = \max(d(A, P) - d(A, N) + \alpha, 0)$$

Here d represents the distance between any two images in the embedding space and α is a margin hyperparameter. The function's goal is to minimize anchor-positive distance while ensuring the anchor-negative distance is larger by at least the margin α. Model updates with triplet loss are greatest when training on "hard" triplets where the anchor–negative distance is initially smaller than the anchor–positive distance in the embedding space. Such hard triplets produce informative gradients for updating the model's weights.

3.2.2. Training Enhancements

While our model architecture training with triplet loss achieved strong performance, we identified two key enhancements that significantly improved results: *class-aware triplet sampling* and *data augmentation*.

Class-Aware Triplet Sampling Given the sparse nature of our dataset, where most individuals have only a few images, random batch sampling rarely yields triplets with the necessary positive pairs. This limitation reduces the effectiveness of triplet loss, as many batches fail to form valid anchor–positive–negative relationships. To address this, we implemented a class-aware triplet sampler that explicitly constructs batches containing at least two positive samples for each anchor image. This approach ensures a higher proportion of valid triplets, enabling the model to learn more effectively from each training batch.

Data Augmentation Following fine-tuning with class-aware triplet sampling, we observed increased adaptation to our domain but also signs of overfitting. Qualitative analysis of the embedding space revealed the model sometimes relied on spurious correlations, grouping images by irrelevant attributes such as background color or camera angle rather than individual-specific biomarkers. To mitigate this, we applied augmentations that altered perspective, rotation, scale, and color during training. Generating two randomly augmented views per original image increased dataset diversity, improved robustness to nuisance factors, and reduced reliance on background cues. This augmentation strategy ultimately decreased overfitting and improved generalization to unseen data.

3.3. Retrieval Techniques

Once the embedding function f_θ has been trained, all catalog images are mapped to a shared representation space. Given a query image, retrieval is performed by ranking candidate identities according to their similarity to the query in this space. A well-trained embedding model should organize images into distinct, well-separated clusters, each corresponding to a unique individual.

We evaluate two retrieval strategies: (1), *nearest neighbor* (NN) retrieval compares the query embedding directly to all catalog image embeddings and ranks identities based on the closest matches; (2) *nearest prototype* (NP) retrieval first computes a representative prototype embedding for

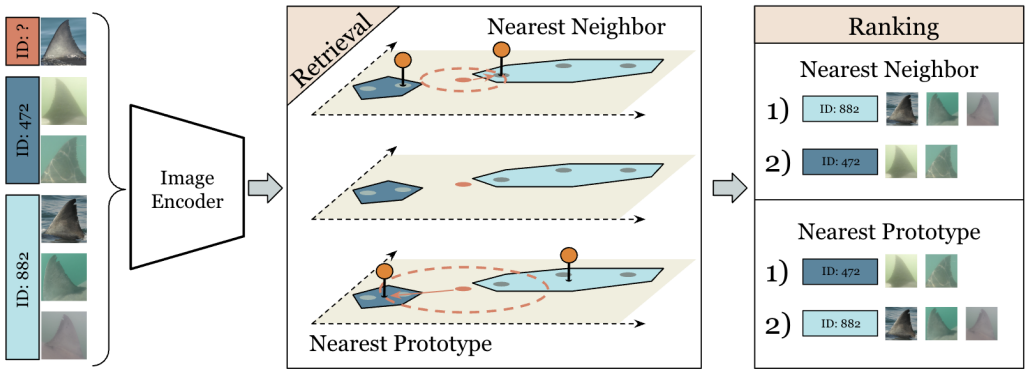

Figure 2. A visual comparison between ranking of shark suggestions using NN vs. NP retrieval. Each image maps to a point in embedding space and two distinct shark clusters emerge. Retrieval of the nearest neighbors or nearest prototypes, indicated by the yellow flags, can impact the resulting ranking.

each individual by averaging all of their image embeddings, then compares the query to these prototypes [21]. Prototype-based retrieval reduces the search space from the total number of images to the number of unique identities and is more robust to intra-individual variation, an example of which can be seen in Figure 2.

Formally, let the embedding of a query image x_q be $z_q = f_\theta(x_q)$, and let the embedding set for identity y be $Z_y = \{f_\theta(x) \mid x \in \mathcal{G}_y\}$. Retrieval ranks candidates using a distance metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$:

Nearest Neighbor (NN) Retrieval. Rank all identities by their closest image embedding:

$$\hat{\mathcal{Y}}_{\text{NN}} = \text{rank}\left(y \in \mathcal{Y} \;\middle|\; \min_{z \in Z_y} d(z_q, z) \right).$$

Nearest Prototype (NP) Retrieval. First compute the prototype for each identity:

$$\mu_y = \frac{1}{|Z_y|} \sum_{z \in Z_y} z,$$

then rank identities by their prototype distance:

$$\hat{\mathcal{Y}}_{\text{NP}} = \text{rank}\left(y \in \mathcal{Y} \;\middle|\; d(z_q, \mu_y) \right).$$

By averaging over all available images for an individual, NP retrieval smooths over nuisance variation and background noise, often improving accuracy in dynamic catalog settings.

4. Experiments

Experiments are conducted on a dataset of 3,083 dorsal fin images of 1,031 unique white sharks, curated by the Block Lab at Stanford University [2]. This dataset contains images captured between 1994 and 2025, with many individuals re-sighted and photographed years apart. In particular, the data set has been curated to contain only a single image of any individual shark encounter. As a result, there are no duplicate images or images with trivial variation which would artificially increase average retrieval scores. To create a held-out test set, we randomly isolated one image from each of the 497 individuals that have two or more photos, resulting in a training set of 2,586 images and a test set of 497 images. Performance is benchmarked using hits@K, which measures the proportion of queries for which the correct individual is present within the top K retrieved matches. We report scores up to K=50, representing a practical upper limit of match results for human review. All reported metrics are averaged over two independent, randomly generated data splits to ensure the robustness of our results.

In addition to evaluating embedding representations of CNN and ViT backbones and our successive training augmentations, we also compare our two distinct retrieval strategies, *nearest neighbor* (NN) and *nearest prototype* (NP) retrieval. The results of our experiments are summarized in Table 1. At a high level, our findings demonstrate the superior capabilities of the ViT backbone, a progressive improvement in retrieval accuracy with each subsequent training enhancement, and the consistent superiority of NP retrieval.

Looking more closely at the results in 1, each applied training optimization yields a measurable performance gain over the entire range of k values evaluated. Improvements in hits@1 values are particularly meaningful in this setting, as they indicate the frequency with which a correct matching individual is presented as the first result for an image queried. Similarly, achieving hits@50 scores above 90%

Model	Nearest Neighbor Retrieval				Prototype Retrieval			
	Hits@1	Hits@5	Hits@25	Hits@50	Hits@1	Hits@5	Hits@25	Hits@50
CNN	0.02	0.05	0.14	0.20	0.02	0.07	0.15	0.22
ViT	0.03	0.09	0.23	0.33	0.05	0.13	0.30	0.38
ViT + LoRA	0.09	0.23	0.40	0.53	0.11	0.23	0.43	0.53
ViT + LoRA + CAS	0.31	0.55	0.74	0.83	0.33	0.56	0.75	0.82
ViT + LoRA + CAS + Aug	0.48	0.67	0.85	0.90	0.51	0.68	0.84	0.91

Table 1. Hits@K results for successive model optimizations (LoRA, class-aware triplet sampling (CAS), and data augmentation) comparing nearest neighbor and prototype retrieval.

indicates that for 90% of the queried images, we present the correct matching individual within an easily reviewable 50 results. In such cases, we are able to reduce the human-in-the-loop effort by more than 95%, where otherwise the full catalog of 1031 sharks would need to be reviewed.

Comparison between NN and NP techniques shows that the latter consistently outperforms the standard nearest neighbor approach, regardless of the underlying embedding model. By aggregating feature representations into a single prototype for each shark, we achieved more stable and accurate shark representations, resulting in a higher accuracy ranking in retrieval. During our evaluation, where test images were held out of the calculation of prototype embeddings, we found that sub-optimally trained embedding models, such as our initial, minimally adapted ViT model, see particularly meaningful improvements with improvements in hits across all K values; as the embedding models improve, NP retrieval continues to boost accuracy, especially improving the ranking of the highest likelihood matches, represented by hits@1. These results highlight that, for fine-grained retrieval tasks, how the embedding space is searched can be as critical as how it is constructed.

Further analysis shows a strong correlation between retrieval performance and the number of images available of a specific individual. When evaluating sharks with at least four total images, nearest prototype hits@1 and 25 results are 73% and 97%, representing 21 and 13 percentage point increases, respectively. This result is particularly encouraging in the nearest prototype setting, where all individuals, regardless of image count, have only one prototype embedding, suggesting that retrieval of the correct prototype is more reliable with more individual data points per shark and not a result of having a higher number of equally labeled embeddings in the search space.

Additionally, we performed an ablation over different categories of image augmentations to assess their impact on retrieval performance. We evaluated three categories: (1) geometric augmentations, including rotation, perspective shift, and reflection; (2) geometric + color augmentations; and (3) geometric + color + erasing augmentations, where random regions of the image were replaced with blank space [30]. Results for these experiments are shown in Table 2.

The findings reveal that while geometric and color augmentations improve generalization, erasing augmentations not only fail to provide benefits but also reset performance gains achieved by other augmentations. This suggests that, in the context of white shark identification, discriminative cues are distributed throughout the dorsal fin image, often in the form of small, localized notches or pigmentation patterns. Removing even a portion of these features can make individuals nearly indistinguishable, underscoring the importance of preserving the complete visual context during training.

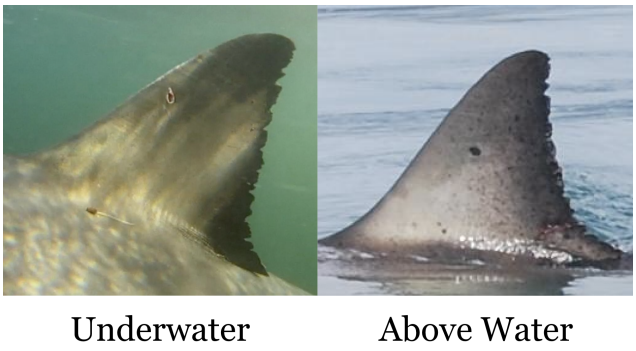

Figure 3. An example of the same individual captured underwater and above water. Matching across these two image capture styles consistently shows lower accuracy.

Finally, analysis of low-performance examples, images where the first hit is not in the top 500 results, reveals that our model has the most difficulty matching across underwater and above water images, an example of which can be seen in 3. A significant majority of our dataset consists of underwater images with an equal proportion of both types being included in the training and test splits. Matching within the same image capture type reveals that accuracy for underwater and above water images is comparable, suggesting that model understanding of either image type in isolation is not the cause for lower performance. Instead,

Augmentations	Nearest Neighbor Retrieval			
	Hits@1	Hits@5	Hits@25	Hits@50
None	0.31	0.55	0.74	0.83
Geometric	0.45	0.65	0.83	0.89
Geo. + Color	0.48	0.67	0.85	0.90
Geo. + Color + Erase	0.35	0.57	0.74	0.83

Table 2. Hits@K results detailing efficacy of augmentation techniques.

we hypothesize that while invariance to position, rotation, and coloration is aided with augmentations, the additional edges, shapes, and contours produced by the water's surface in above water images compared with underwater images falsely contributes towards the final fin embedding. Additional modeling techniques and explicit encoding of this categorical difference to improve performance are areas for future exploration.

5. Shark Matching Platform

To bridge the gap between our deep learning model and the practical needs of researchers, we developed Shark Matcher, a purpose-built platform that streamlines the re-identification workflow. The system replaces the laborious and time-consuming manual process of searching photographic catalogs for individual matches. For each queried image, the interface returns the top-ranked candidate matches, allowing the crucial expert task of visual inspection to begin immediately, and an example of this is Figure 4. The resulting human-in-the-loop system provides an intuitive and efficient interface to verify suggested matches, annotate new individuals, and manage catalog metadata. The deployment of Shark Matcher with our collaborators at the Block Lab has substantially reduced their manual effort required for data curation, enabling a more rapid and scalable analysis of shark populations.

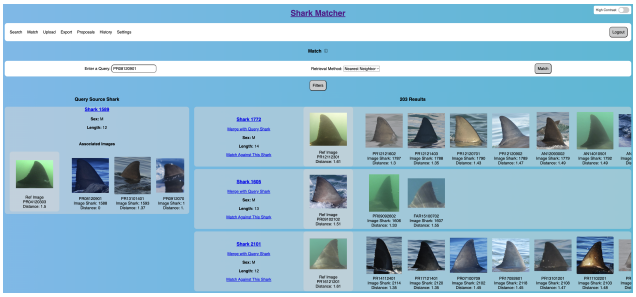

Figure 4. The "Match" tab in the Shark Matcher UI shows a queried shark and a ranked list of matching shark suggestions. Users can then match sharks, and update relevant metadata using Shark Matcher.

6. Conclusion

In this work, we develop and refine a pipeline for re-identification of individual white sharks, specifically addressing the challenge of data sparsity. We present optimizations to model architecture, training, and retrieval technique and quantify their effects within the setting of automated re-identification. Our results confirm that each major optimization presented results in a significant increase in retrieval accuracy. Our final results of 51% hits@1 and 91% hits@50 cumulate in a powerful re-identification tool that provides meaningful benefit to white-shark researchers by significantly reducing the time they must invest in data processing, no longer having to manually compare new images against all sharks in their catalogs.

Beyond white sharks, our approach illustrates a flexible methodology for adapting large vision models to specific, real-world ecological monitoring tasks. As the global community continues to amass increasingly large and diverse databases of individually identified animals, the need for reliable, principled, and scalable tools to manage and leverage these resources becomes ever more critical. Our framework represents a step toward meeting this need by combining state-of-the-art deep learning with domain-specific considerations to ensure both performance and usability at scale. Future work will focus on refining embedding generation and retrieval methods, as well as extending evaluation beyond a single dataset to include additional species and imaging conditions. Such experiments will deepen understanding of our framework's generalization capabilities and guide the development of more universal, species-agnostic identification tools that can support conservation efforts across ecosystems.

References

[1] Scot Anderson, Taylor Chapple, Salvador Jorgensen, A. Klimley, and Barbara Block. Long-term individual identification and site fidelity of white sharks, carcharodon carcharias, off california using dorsal fins. *Marine Biology*, 2011. 1, 2

[2] Samantha Andrzejaczek, Taylor K. Chapple, Salvador J. Jorgensen, Scot D. Anderson, Michael Castleton, Paul E. Kanive, Timothy D. White, and Barbara A. Block. Multi-

decadal high-resolution data reveal the cryptic vertical movement patterns of a large marine predator along the californian coast. *Frontiers in Marine Science*, 2022. 5

[3] Jay Barlow, John Calambokidis, Erin A Falcone, C Scott Baker, Alexander M Burdin, Phillip J Clapham, John KB Ford, Christine M Gabriele, Richard LeDuc, David K Mattila, et al. Humpback whale abundance in the north pacific estimated by photographic capture-recapture with bias correction from simulation studies. *Marine Mammal Science*, 2011. 2

[4] Tanya Berger-Wolf, Daniel Rubenstein, Charles Stewart, Jason Holmberg, Jason Parham, Sreejith Menon, J.P. Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arxiv*, 2017. 3

[5] Christian Bergler, Alexander Gebhard, Jared R. Towers, Leonid Butyrev, Gary J. Sutton, Tasli J. H. Shaw, Andreas Maier, and Elmar Nöth. Fin-print a fully-automated multi-stage deep-learning-based framework for the individual recognition of killer whales. *Nature Scientific Reports*, 2021. 1, 2

[6] Drew Blount, Shane Gero, Jon Van Oast, Jason Parham, Colin Kingen, Ben Scheiner, Tanya Stere, Mark Fisher, Gianna Minton, Christin Khan, Violaine Dulau, Jaime Thompson, Olga Moskvyak, Tanya Berger-Wolf, Charles Stewart, and Jason Holmberg. Flukebook: An open-source ai platform for cetacean photo identification. *Mammalian Biology*, 2022. 3

[7] Jonathan P. Crall, Charles V. Stewart, Tanya Y. Berger-Wolf, Daniel I. Rubenstein, and Siva R. Sundaresan. Hotspotter — patterned species instance recognition. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2013. 1, 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2021. 1, 2, 3, 4

[9] Kunihiko Fukushima, Sei Miyake, and Takayuki Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 1983. 3

[10] Philip Hammond, Tessa Francis, Dennis Heinemann, Kristy Long, Jeffrey Moore, André Punt, Randall Reeves, Maritza Sepulveda, Gudjon Sigurdsson, Margaret Siple, Gísli Víkingsson, Paul Wade, Rob Williams, and Alexandre Zerbini. Estimating the abundance of marine mammal populations. *Frontiers in Marine Science*, 2021. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[12] G.R. Hillman, N. Kehtarnavaz, Bernd Würsig, Babak Araabi, Glenn Gailey, David Weller, S. Mandava, and H. Tagare. Finscan, a computer system for photographic identification of marine animals. In *IEEE Biomedical Engineering*, 2002. 1, 2

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv*, 2021. 4

[14] Benjamin Hughes and Tilo Burghardt. Automated visual fin identification of individual great white sharks. *arXiv*, 2016. 1, 2

[15] Megan Maroen, Mary Rowlinson, Conrad Matthee, and Sara Andreotti. The effectiveness of the semi-automated identifin software to accurately fast-track dorsal fin photographic identifications of sharks. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 2024.

[16] Olga Moskvyak, Frederic Maire, Asia O. Armstrong, Feras Dayoub, and Mahsa Baktashmotlagh. Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. In *Digital Image Computing: Techniques and Applications (DICTA)*, 2021. 2

[17] Ekaterina Nepovinnykh, Ilia Chelak, Tuomas Eerola, Veikka Immonen, Heikki Kälviäinen, Maksim Kholiavchenko, and Charles Stewart. Species-agnostic patterned animal re-identification by aggregating deep local features. *International Journal of Computer Vision*, 2024. 1, 3

[18] Douglas Nowacek, Fredrik Christiansen, Lars Bejder, Jeremy Goldbogen, and Ari Friedlaender. Studying cetacean behaviour: new technological approaches and conservation applications. *Animal Behaviour*, 2016. 2

[19] Lasha Otarashvili, Tamilselvan Subramanian, Jason Holmberg, J. J. Levenson, and Charles V. Stewart. Multispecies animal re-id using a large community-curated dataset. *arxiv*, 2024. 1, 3

[20] Jason Parham, Jonathan Crall, Daniel Rubenstein, Jason Holmberg, Tanya Berger-Wolf, , and Charles Stewart. An animal detection pipeline for identification. *International Journal of Modern Physics B*, 2018. 1

[21] J. J. Rocchio. Relevance feedback in information retrieval. In *The Smart retrieval system - experiments in automatic document processing*, 1971. 5

[22] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4

[23] Jaime Thompson, Victoria Zero, Lori Schwacke, Todd Speakman, Brian Quigley, Jeanine Morey, and Trent Mcdonald. finfindr : Automated recognition and identification of marine mammal dorsal fins using residual convolutional neural networks. *Marine Mammal Science*, 2021. 1, 2

[24] Anna Van Tienhoven, J. Hartog, Renate Reijns, and Victor Peddemors. A computer-aided program for pattern-matching of natural marks on the spotted raggedtooth shark carcharias taurus. *Journal of Applied Ecology*, 2007. 1, 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[26] Hendrik Weideman, Zachary Jablons, Jason Holmberg, Kirsten Flynn, John Calambokidis, Reny Tyson Moore, Jason Allen, Randall Wells, Krista Hupman, Kim Urian, and Charles Stewart. Integral curvature representation

and matching algorithms for identification of dolphins and whales. In *ICCV*, 2017. 1, 2

[27] Randall S Wells and Michael D Scott. Estimating bottlenose dolphin population parameters from individual identification and capture-release techniques. *Reports of the International Whaling Commission*, 1990. 2

[28] Bernd Wuersig and T.A. Jefferson. Methods of photo-identification for small cetaceans. *Methods of Photo-identification for Small Cetaceans*, 1990. 2

[29] Zhaoxiang Zheng, Yaqin Zhao, Ao Li, and Qiuping Yu. Wild terrestrial animal re-identification based on an improved locally aware transformer with a cross-attention mechanism. *Animals*, 2022. 1, 2

[30] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI 2017*, 2017. 6