Optimizing Vision Transformers for White Shark Re-Identification

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

University of California Santa Cruz[1], Stanford University[2]

Introduction and Motivation

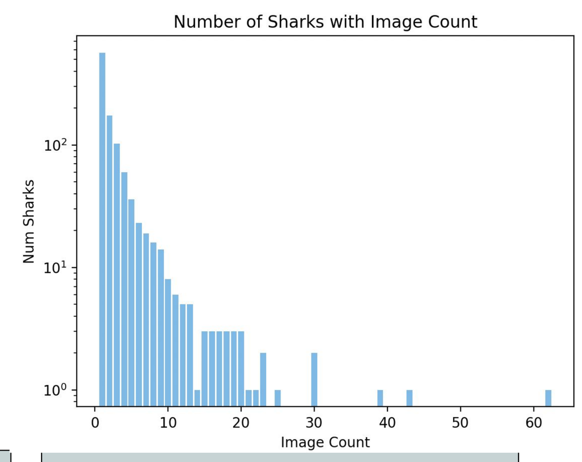
Animal Re-Identification (Re-ID), matching new observations against a catalog of known individuals, is essential for wildlife conservation, however manual re-identification is time-consuming and error-prone.

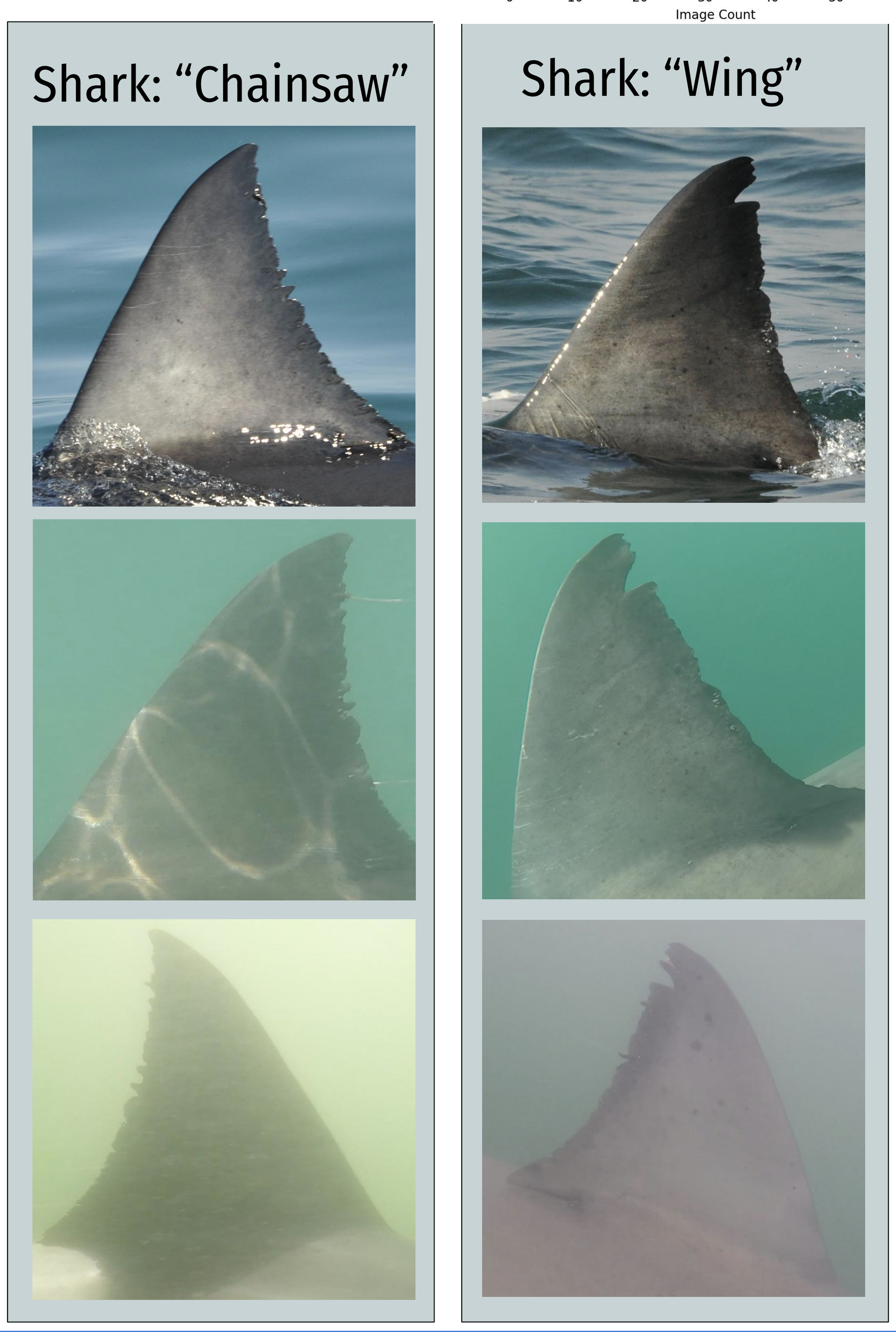
We introduce an automated white shark Re-ID framework designed to accelerate and improve this process.

Our system leverages visual features of dorsal fins [1] while keeping humans in the loop for validation, enabling efficiency and reliability.

Dataset

- Their rarity, vast habitat, and difficulty to photograph creates a sparsely populated database of many sharks, each represented by few images
  - 3083 dorsal fin images
  - 1031 unique white sharks
  - 20+ year timespan
- Images include numerous confounding factors
  - Fin orientation, rotation, angle
  - Background color/lighting
  - Water surface, reflection, splash





Methods

- **Image encoder maps images to embedding space**
  - Google/vit-large-patch16-384 feature extraction backbone [2]
  - Multilayer perceptron (MLP) projection head
  - Trained for:
    - Sensitivity to biomarkers on the dorsal fin (notches and pigmentation patterns)
    - Invariance on confounding factors
- **Retrieval algorithm searches embeddings space**
  - Well-trained model organizes embeddings into distinct clusters for each shark
  - Matching sharks are retrieved by searching the embedding space for nearest neighbors to a query image

Shark Re-Identification Framework

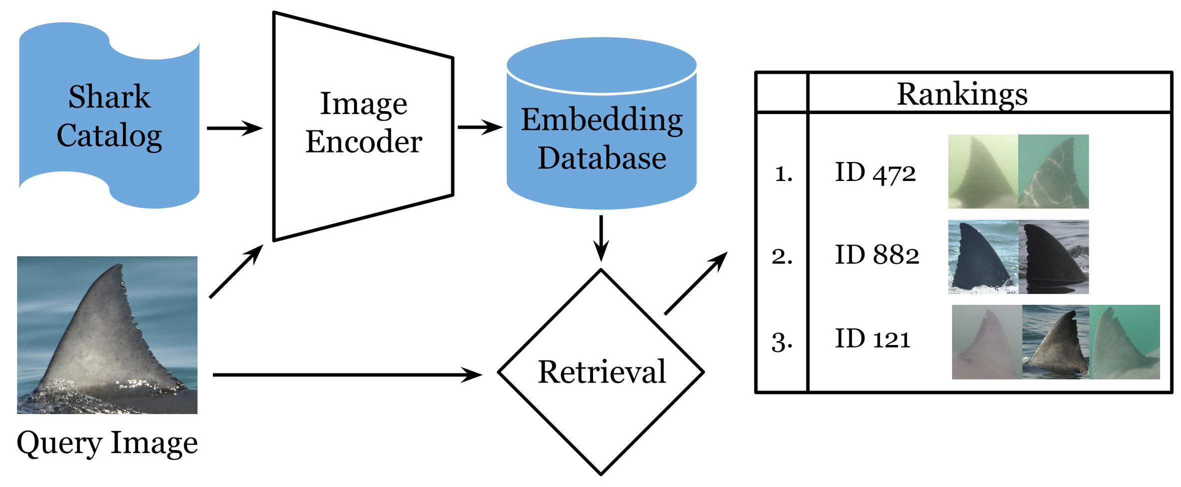


Image Encoder Training

- **Training Approach:**
  - *Triplet loss function* [3]:

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

  - *Fine-tuning with low rank adaptation* [4]
    - Parameter efficient method allows convergence in 72 hours on single GPU
- **Training Enhancements:**
  - *Class-aware triplet sampling*
    - Sparse dataset: many training batches lack anchor-positive samples
    - Explicitly sample two positive samples for each anchor image
  - *Training image augmentation*
    - Augmentations to perspective, rotation, scale, and color during training
    - Increase invariance to confounding traits in images
    - Reduce overfitting

Retrieval Techniques

Nearest Neighbor (NN)

- Rank all identities by their closest image embedding to the query

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

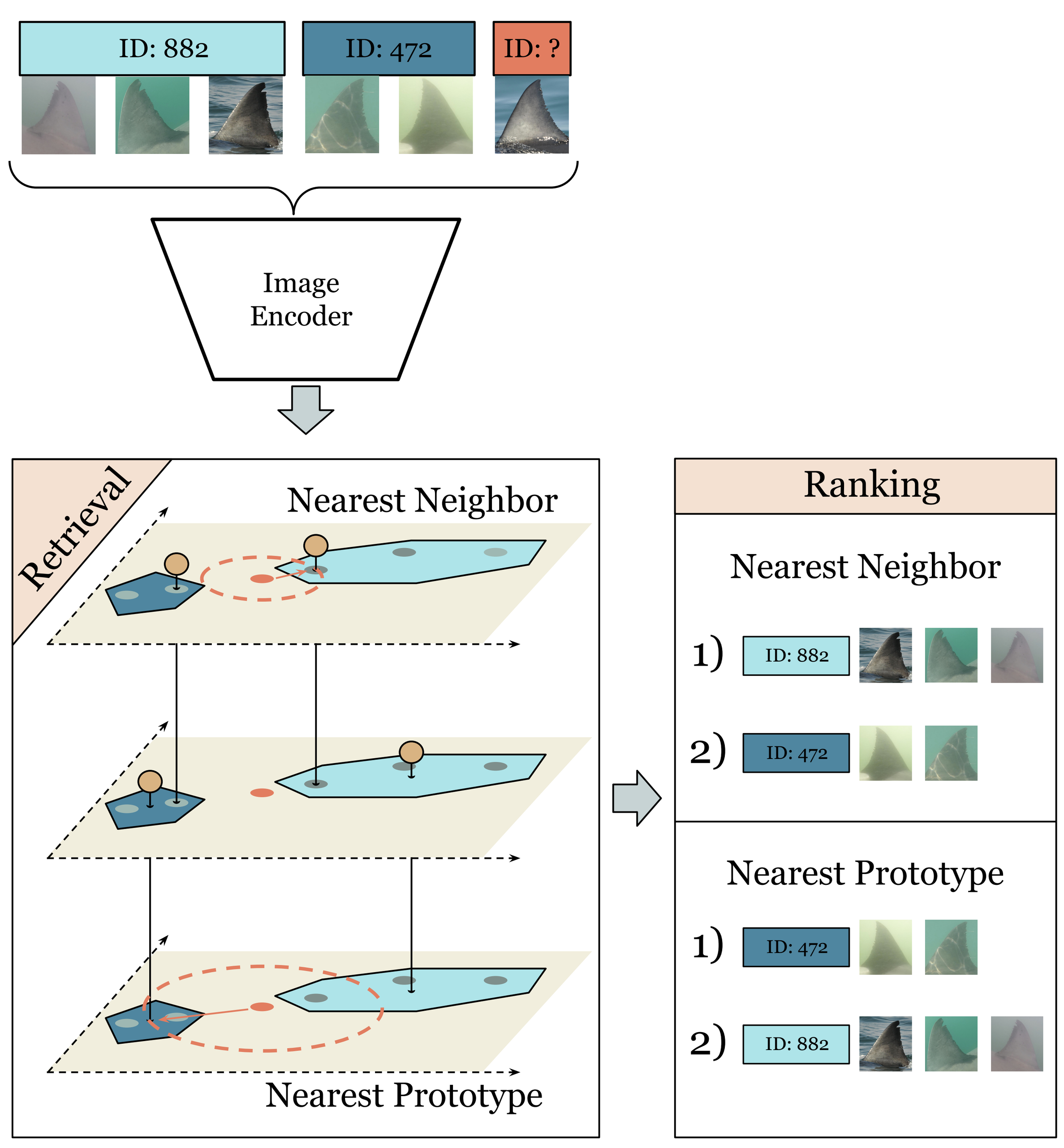
Nearest Prototype (NP)

- Prototype embedding for each ID [5]

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

- Rank by the closest prototype to the query

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



Results

- Hits@K scores
  - Proportion of queries where correct individual is among first k retrieved
  - K=50 represents a practical upper limit for human review
- Test set of 1 randomly selected image from each individual with > 1 image
  - 497 train, 2,586 test images

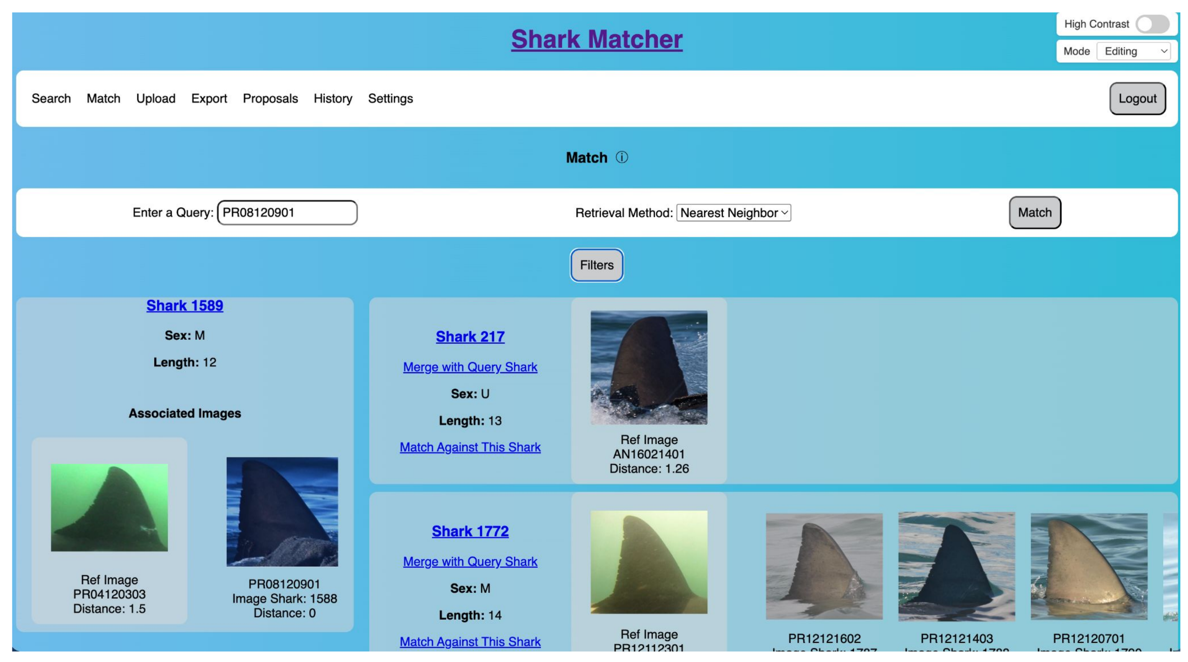
Model	Nearest Neighbor Retrieval			Prototype Retrieval		
	Hits@1	Hits@5	Hits@10	Hits@1	Hits@5	Hits@10
ViT	0.03	0.09	0.13	0.05	0.13	0.18
ViT+LoRA	0.09	0.23	0.29	0.11	0.23	0.30
ViT+LoRA+CAS.	0.31	0.55	0.64	0.33	0.56	0.64
ViT+LoRA+CAS.+Aug.	0.48	0.68	0.76	0.51	0.68	0.75

Ablation of augmentation techniques:

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

Shark Matcher Human-in-the-Loop UI

Collaborating with marine biologists, we developed a UI taylored to streamline their labeling workflow. Our model's match results can be browsed, filtered, reviewed, and approved, committing them to a shark database.



Conclusion and Future Work

We develop a framework for white shark Re-ID, presenting optimizations to model training and retrieval technique, showing their benefits to retrieval accuracy. Paired with our UI, this work has become a valuable tool for dataset de-duplication and matching newly captured shark images for our marine biologist collaborators.

As future work, we aim to:

- Evaluate over different animal species to better understand generalization potential
- Improve model adaptation with online learning of newly added data
- Develop techniques for training accurate models on noisy and mislabeled real-world data

Acknowledgements

This work was partially supported by the NSF grant CCF-2023495.

References

[1] Nowacek, Christiansen, Bejder, Goldbogen, Friedlaender. Studying cetacean behaviour: new technological approaches and conservation apps. (2016).

[2] Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, Dehghani, Minderer, Heigold, Gelly, Uszkoreit, and Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv (2021)

[3] Schroff, Kalenichenko, and Philbin. Facenet: A unified embedding for face recognition and clustering. CVPR (2015).

[4] Hu, Shen, Wallis, Zhu, Li, Wang, Wang, and Chen. Lora: Low-rank adaptation of large language models. arXiv (2021).

[5] Rocchio. Relevance feedback in information retrieval. (1971).

BayLearn

Bay Area Machine Learning Symposium