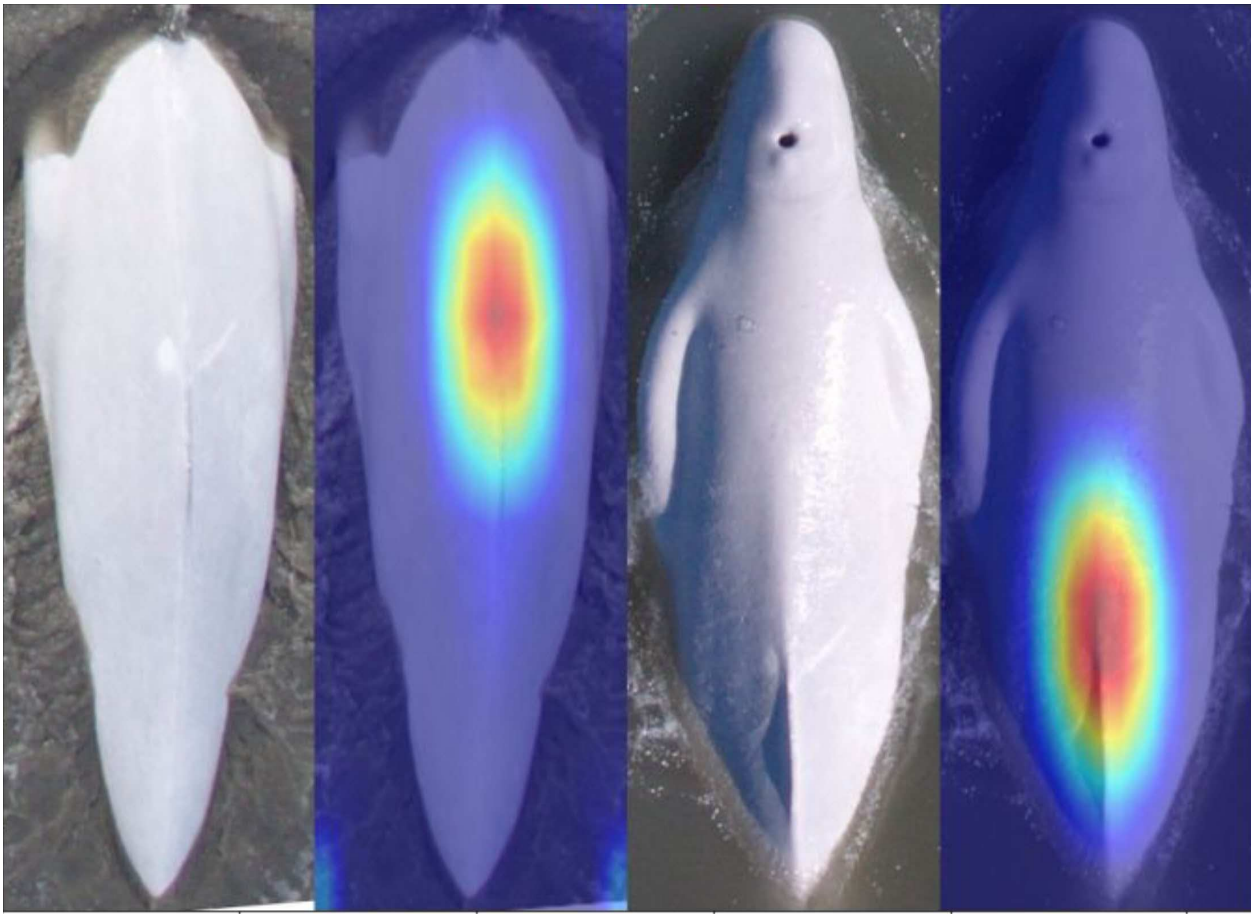
OCS Study
BOEM 2024-021

“Where’s Whale-Do?” Evaluating and Implementing Competitor Approaches to Machine Learning-based Re-ID of Belugas



U.S. Department of the Interior
Bureau of Ocean Energy Management
Sterling, VA

BOEM
Bureau of Ocean Energy Management

OCS Study
BOEM 2024-021

“Where’s Whale-Do?” Evaluating and Implementing Competitor Approaches to Machine Learning-based Re-ID of Belugas

May 2024

Authors:
Jason A. Holmberg
Lasha Otarashvili
Jamison Smith

Prepared under Contract 140M0121D0004
By
Wild Me
1726 N Terry Street
Portland, OR 97217

Blue World Research Institute
728 West Ave., #174
Cocoa, FL 32927

U.S. Department of the Interior
Bureau of Ocean Energy Management
Sterling, VA

BOEM
Bureau of Ocean Energy Management

DISCLAIMER

Study concept, oversight, and funding were provided by the U.S. Department of the Interior, Bureau of Ocean Energy Management (BOEM) under Contract Number 140M0121D0004. This report has been technically reviewed by BOEM, and it has been approved for publication. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the opinions or policies of BOEM, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

REPORT AVAILABILITY

Download a PDF file of this report at https://espis.boem.gov/Final%20Reports/BOEM_2024-021.pdf. To search other studies completed by BOEM's Environmental Studies Program, visit https://www.boem.gov/environment/environmental-studies/environmental-studies-information/.

CITATION

Holmberg JA, Otarashvili L, Smith J. 2024. "Where's whale-do?" Evaluating competitor approaches to machine learning-based re-ID of belugas. Sterling (VA): U.S. Department of the Interior, Bureau of Ocean Energy Management. 34 p. Report No.: OCS Study BOEM 2024-021.

ABOUT THE COVER

Our cover photo displays an example of Gradient-weighted Class Activation Mapping (GRAD-CAM) rendering of corresponding visual patterning in dorsal photographs of a previously matched beluga. Grad-CAM (Selvaraju et al. 2017) heatmaps were used to highlight the pixels that a machine learning model uses in identifying a beluga as a particular individual. The image was generated for whale 340 as part of the Explainability Bonus Round of the "Where's Whale-Do?" machine learning competition.

ACKNOWLEDGMENTS

Wild Me would like to thank the many competitors of the "Where's Whale-Do?" machine learning competition. Critical input to this report came from original project contributions from Dr. Jason Parham as well as the final competition reports produced by Greg Lipstein and Peter Bull of DrivenData. Dr. Paul Wade and Christy Sims of NOAA provided data and fundamental, real-world input into the formulation of the competition.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI ..... Artificial Intelligence
BOEM ..... Bureau of Ocean Energy Management
CE ..... cross-entropy
CNN ..... convolutional neural networks
GeM ..... generalized mean
GPU ..... graphics processing unit
Grad-CAM..... Gradient-weighted Class Activation Mapping
ML ..... Machine Learning
PIE..... Pose Invariant Embeddings (a matching algorithm used by Wild Me)
TTA ..... test-time-augmentation

1 Executive Summary

Wild Me (wildme.org) reviewed the top-ranking solutions in the "Where's Whale-do?" machine learning competition, which was launched as a collaboration between BOEM, the National Aeronautics and Space Administration, HeroX, Wild Me, and the National Oceanic and Atmospheric Administration (NOAA) to apply artificial intelligence (AI) to speed and scale photo-based, mark-recapture studies of the Cook Inlet beluga population. The team's joint effort engaged a significant number of competing teams to build machine learning to match individual beluga whales (*Delphinapterus leucas*) across photos and years in support of population studies conducted by NOAA. Considered a hard problem in both by-eye photo matching and in status quo computer vision due to subtle and sometimes changing patterning as well as variable lighting and water conditions, matching individual belugas is a worthy challenge for a machine learning competition in which new techniques can be explored to advance the state-of-the-art for a real-world mark-recapture problem. Wild Me explored the winning solutions and discovered several common techniques were used, and we were able to show that the competitors exceeded the capabilities of the current state-of-the-art (Pose Invariant Embeddings [PIE] v2; rank-1 ID prediction 44.6%) using ArcFace, EfficientNet, and other tools and techniques. Rather than implementing exactly a winning solution, which all represent 12+ ensembles of models meant to achieve even the slightest edge over other competitors, we demonstrated that the bulk of the benefit (rank-1 ID prediction 69.1%) can be achieved with a base ArcFace+EfficientNet implementation, which can also be cross-applied to other species. We subsequently developed a new matching algorithm named "MiewId" that is based on the ArcFace+EfficientNet combination used by competitors, and we deployed it in the Flukebook.org platform for production use in matching aerial imagery. Flukebook.org is now fully operational for beluga matching by NOAA with the new algorithm, and several points of feedback and over 100 code improvements have been subsequently made. In addition to belugas, we have deployed MiewId for a broader array of finned whales and dolphins, demonstrating that our work achieved success for belugas and much broader impact for additional marine species.

2 Completed Task Summary Table

The following tasks were completed under BOEM Contract 140M0121D0004 (Table 1).

Table 1. Completed Tasks Summary

Task	Computer Vision Techniques	Status
Task 2.2.2.2 Competition results evaluation In addition to providing technical requirements for competitors, Wild Me staff will assist BOEM in evaluating algorithm performance metrics and technical foundation, as well as providing input on algorithm and winner selection from among those submitted challenge participants.	PIE v2 (Moskvyak et al. 2019), ArcFace (Deng et al. 2019).	**COMPLETE:** Wild Me generated this report in March 2023 as its completed work product for Task 2.2.2.2. We evaluated the machine learning solutions of the "Where's Whale-Do?" competition and recommended an implementation pathway that was approved by NOAA in April 2023.
2.2.2.3 Catalog Integration	N/A	**COMPLETE:** Wild Me integrated the base NOAA catalog of individuals and has directly supported additional rounds of bulk upload of new data.
2.2.2.4 Integration of the automated matching algorithm into the database platform	ArcFace (Deng et al. 2019) as implemented in the new algorithm MiewId.	**COMPLETE:** Wild Me deployed the synthesized MiewId algorithm to Flukebook where it is being used by NOAA on new data.
2.2.2.5 Identify areas for automated matching process improvement	ArcFace (Deng et al. 2019) as implemented in the new algorithm MiewId.	**COMPLETE:** Wild Me made several rounds of improvement to the MiewId algorithm that resulted from the competition. These included: • Support for rotation matching, ensuring real-world photographs did not need pre-cropping • NOAA request: implement Grad-CAM rendering from competition bonus prize results • Graphics processing unit (GPU) memory reduction in Grad-CAM rendering of MiewId suggested matches • NOAA request: allow collaborator access to a user's bulk imports
2.3.3 Draft and final summary report	-	**COMPLETE**
2.3.4 Database platform	-	**COMPLETE:** Wild Me has maintained and improved the Flukebook.org platform throughout the project, and MiewId has been deployed into production for belugas and several other cetacean species as a result of this project.

3 Competition Results Technical Review

Wild Me reviewed the top-rank competition solutions to understand how their technical implementation could be synthesized into a highly accurate and maintainable machine learning (ML) solution for beluga whale (*Delphinapterus leucas*) reidentification.

3.1 Challenge Data Overview

Data used in the competition were provided by NOAA and included the following:

- 9,303 beluga images of top and lateral viewpoints, representing 978 individual whale IDs over 3 years (2017–2019) as captured from Alaska's Cook Inlet. Lateral viewpoints were collected by NOAA by boat while aerial imagery was collected via overhead drone. Images were pre-cropped and had a standardized rotation to ensure only one whale was present in each image.
- Metadata with whale ID, date, encounter ID, viewpoint

Data was split into train set (~65%) and a hidden test set (35%) that competitors were not allowed to see. Wild Me also coordinated with lila.science to make the data available for further AI exploration in the future.

3.2 Challenge Competitors Overview

The "Where's Whale-Do?" competition ran April 28, 2022, to July 31, 2022. It engaged 442 participants from 61 countries. 63 competitors made it through to submission, contributing a total of 1,112 submissions.

3.3 Challenge Evaluation Setup

All competition submissions were scored against a test dataset that consisted of 10 scenarios, as listed in Table 2, and scored by mean average precision (mAP) with a maximum score of 1 representing perfect predictions.

Table 2. "Where's Whale-Do?" Competition Scenarios

Scenario	# Queries	# Database Annotations	Notes	Avg Finalist Score (mAP)
1	1,208	3,290	Top-view test images against all top-view images	0.45
2	592	633	Top-view new whales with multiple images against top-view new whales	0.50
3	616	2,041	Top-view new images of training set whales against all top-view training images	0.50
4	300	1,110	Top-view 2017 test images against all 2017 top-view images	0.50
5	346	971	Top-view 2018 test images against all 2018 top-view images	0.68
6	562	1,209	Top-view 2019 test images against all 2019 top-view images	0.52
7	562	403	Top-view 2019 test images against all 2018 top-view images with same whales	0.62

Scenario	# Queries	# Database Annotations	Notes	Avg Finalist Score (mAP)
8	562	338	Top-view 2019 test images against all 2017 top-view images with same whales	0.57
9	102	318	Lateral-view new whales against top-view new whales with lateral images	0.22
10	309	111	Top-view new whales with lateral images against lateral-view new whales	0.29

3.4 General Overview of Highest Performing Solutions

The top-scoring team achieved ~0.495 mean average precision (mAP), and only eight participants surpassed > 0.4. The top four finalists (see Table 3) were separated by < 0.01 mAP, and the winning competition methodologies had significant overlap in approach. These included:

- Pre-trained EfficientNet (Tan et al. 2019) backbones
- Facial recognition loss functions, in all cases some variation on ArcFace (Deng et al. 2019), which help enforce "inter-class separability" and "intraclass compactness" within the features of the trained model
- Ensembles of around a dozen models, trained with k-fold cross validation
- Stratification by whale ID (i.e., local testing on whales their model had not seen before)
- Image flipping was a useful test-time-augmentation (TTA) for multiple finalists
- Query expansion: 1st place finalist used database matching images to search for other matches

In their write-ups, competitors highlighted the matchability of the dorsal ridge and scars/marks (if present). Color was also considered but not thought to be a strong feature.

Table 3. Summary of Top 4 Competition-Winning Submissions

Place	Team or User	Public Score	Private Score	Summary of Model	Bonus Round
[1]	Ammarali32	0.4902	0.4954	Ensemble of pre-trained EfficientNet backbones, trained with k-fold cross validation and ArcFace loss. Matching database images are used for re-ranking, and horizontal flip augmentation is applied during inference.	Grad-CAM heatmaps
[2]	qwerty64	0.4936	0.4953	Ensemble of pre-trained EfficientNet backbones, trained with k-fold cross validation and sub-center ArcFace with adaptive margin loss.	-
[3]	sheep	0.4846	0.4910	Ensemble of pre-trained ConvNext and EfficientNet backbones, trained with k-fold cross validation and ArcFace combined with Focal Loss.	-
[4]	karelds	0.4838	0.4871	Ensemble of pre-trained EfficientNet backbones, trained with k-fold cross validation and sub-center ArcFace with adaptive margin loss. Horizontal flip augmentation during inference.	Grad-CAM heatmaps

The top four winning solutions are available in this Github repository.

3.5 Evaluation of Top-performing Results

3.5.1 General Overview

The most prevalent method used in the winning solutions involved an EfficientNet (Tan et al. 2019) backbone paired with a variation of the ArcFace (Deng et al. 2019) loss function. The competition's evaluation process was thorough, with diverse evaluation scenarios and a hidden test set utilized to prevent competitors from overfitting to a narrow distribution. The approach seems to have been successful in controlling overfitting, as the top solutions did not feature pseudo-labeling (incorporating unlabeled data with interim model-generated labels) or leaderboard distribution probing/tuning (purposely submitting models or predictions to "probe" the distribution of the private/withheld test set) techniques. These techniques, while common in competitions, are known to have limited generalization when applied outside a specific dataset.

3.5.2 Preprocessing

Generally, the competitors found success with image resolutions in the mid to higher range (around 512 pixels). Solution ensembles utilized a diverse range of resolutions, starting from 256 pixels up to 1440 pixels. The images are mostly rectangular, although most competitors opted to resize them to a square. It is worth exploring whether this was a deliberate choice for increased performance, as even though the samples are rectangular, their (limited) variation in aspect ratio could still cause undesired behavior during training.

Surprisingly, pre-filtering individuals with less than four samples was also a successful strategy, as used by the 4th place solution. This approach dramatically reduces the size of the training dataset while improving solution performance.

3.5.3 Augmentations

A large variety of augmentations were employed by competitors. Mostly, the augmentation hyperparameters were not too aggressive, probably to ensure sample diversity while maintaining a reasonable clarity for the identifiable features. Augmentations used included:

- Horizontal flip
  Some other ID competitions have utilized the Horizontal flip technique to generate new individuals synthetically, considering a flipped sample as belonging to a different individual than the original one. In this competition's top solutions, the image was flipped, and the label of the individual was kept unchanged. Hypothetically, this setup could have forced the model to learn flip invariance. Additionally, the ambiguity could have been amortized by sub-center ArcFace loss or resolved later by averaging the results on the original+flipped samples at test time.
- Color augmentations like Sharpen, Grayscale, CLAHE, ColorJitter, Posterize
- Shift, Scale, and Rotation augmentations
- Gaussian blur and noise
- RandomGridShuffle, Cutmix, Cutout, triangle (custom), and Coarse dropout
  These are some of the more ‘aggressive’ forms of augmentation. They involve some form of ‘cutting’ and deleting or permuting geometrical pieces in the image. They were used only in some of the winning solutions.

3.5.4 Optimizers

Optimizers are algorithms or methods implemented to 1) define appropriate weights for training at each epoch and 2) minimize a loss function. Optimizers are used during model training for the general purpose of improving the performance of the final ML model. The two optimizers most prevalent in competitor solutions were the following:

- Adam (Adaptive Moment Estimation) (Kingma & Ba et al. 2014)
- AdamW (Adam + decoupled Weight decay) (Loshchilov & Hutter et al. 2019)

The Adam optimizer has been shown to be generally a good fit for a diverse range of tasks. A more recent extension is AdamW. Both Adam and AdamW were employed successfully by the top solutions.

3.5.5 Scheduler

While an optimizer defines an 'update rule' for changing each weight on a given training step, a "scheduler" defines the global rate according to which the optimizer change is calculated. Competition solutions used a scheduler with warmup and some sort of decay. Variations included:

- Warmup and exponential decay
- Cyclic cosine decay with warmup

Despite the Adam optimizer's ability to adjust learning rates for each optimized parameter, our experience suggests that it is still beneficial to control the maximum learning rate for Adam at each training set. Using a scheduler with stages of raising (warmup) and decaying learning rate has been shown to result in faster convergence of the model with better end results.

3.5.6 Models

The following model architectures and configurations were implemented by competitors.

3.5.6.1 Backbones

A "backbone" is a feature-extracting network that transforms input data into a feature representation. Backbones can be used as stand-alone networks on simpler ML tasks, but here competitors utilized them as the feature-extracting part of their more complex entries. EfficientNet (Tan & Le 2019) backbones dominated in the top solutions, including the following:

- EfficientNet-b4 [1] [2]
- EfficientNet-b7 (With or without noisy student weights) [2]
- Efficientnetv2-M

3.5.6.1.1 Embedding Layer

A model's embedding layer transforms input data in to a denser, more efficient data representation. The choice of the embedding size in the competition was mostly kept to standard 512/768, which is the default for the chosen backbones. Most of the approaches experimented with expanding the feature dimension at the final stage using stacked pooling layers before reducing it back down to the intended size with a linear layer. The reasoning could be that the expansion will increase expressiveness and the following reduction will act as a bottleneck and implicitly regularize the learned features.

In other ID competitions, some of the top solutions concatenated the embeddings from the last two layers. The reasoning is that in convolutional neural networks (CNN), the lower layer is expected to capture

more local characteristics of the input image. A combination of activations at different levels of scale could increase the expressiveness of the final embeddings.

3.5.6.2 Neck

Object detecting ML models often contain a backbone network designed for feature extraction, a neck model for feature aggregation, and a head model for inference. The rough architecture of the model neck and head for the competition's top solutions was as follows:

Generalized Mean (GeM) Pooling → Batchnorm1d → (Linear) →ArcFace module

3.5.6.2.1 Generalized Mean (GeM) Pooling

GeM pooling has been used universally in almost all of the top solutions, such as the 1st Place Winner (DrivenData 2022). GeM pooling is a generalization of average pooling and max pooling, both of which are prevalent in the diversity of neural network architectures. The average pooling and max pooling happen to be special cases of the GeM pooling at the extremes of the pooling parameter range. The benefit of GeM pooling is that the pooling parameter is differentiable and learnable by the model. Although, some approaches from other ID competitions found success in keeping this parameter fixed at a specific value p=3.

Some of the competitors found success by setting up multiple branches of GeM pooling and running the backbone output in each separately. The idea is that each branch receives the same input from the backbone, but each will learn slightly different pooling parameters. Concatenating the pooled result of each branch results in a more expressive representation which can then be passed to the head module.

3.5.6.2.2 Batchnorm1d

The batch normalization layer before the ArcFace head is common among all solutions in this and other ID competitions. Layer normalization is another method known to outperform BatchNorm in certain tasks and could be worthwhile to explore, although it was not used by any of the top solutions.

3.5.6.3 Head

The ArcFace (Deng et al. 2019) loss function module and its derivatives are prevalent in all solutions. The complete ArcFace loss can be broken down into a margin function + a loss function. The margin function forces the inter-class separability, while the loss function derives the final value for model error. The derivatives used are:

- Sub-center ArcFace (Deng et al. 2020)
  ArcFace compares the class predictions to the first nearest class cluster center. Sub-center ArcFace relaxes this comparison to K sub-centers. The underlying theory makes this variant a better choice for datasets where there is label noise or unmatchable samples by implicitly filtering noisy samples in separate sub-centers.
- ArcFace with adaptive margins
- Elastic ArcFace
  ArcFace enforces a fixed amount of separation (margin) between each class cluster. This is suboptimal since the inter and intraclass variation between the specific classes is not always the same. In contrast, both of the above variants vary the margins for each class which allows for more flexibility in class separation.

3.5.6.4 Loss Functions Paired with ArcFace

Two loss functions were paired with ArcFace in winning solutions.

- Cross-entropy (CE) loss
  By default, the ArcFace margin function is followed by a CE loss. This has been the choice for most of the top solutions.

- Focal loss
  One of the competitors opted to combine Focal Loss with the ArcFace head. Focal Loss is a variation of the CE loss that incorporates a factor to diminish the impact of high-confidence (easy) samples and enhance the impact of lower-confidence (hard) samples. This formulation has proven effective in mitigating the effects of imbalanced sample distribution in classification tasks. By extension, the long-tailed distribution of annotations-per-sample could similarly benefit from the properties of Focal Loss.

3.5.6.5 Ensembling

All competition finalists used an ensemble of around a dozen models. Ensembling is a technique that combines the predictions of multiple models. The idea is that by combining the outputs of different models of significant diversity, the ensemble can achieve better performance than any individual model. However, ensembling comes at the cost of increased inference complexity, as well as increased training and inference times. As such, it is important to carefully consider the benefits and drawbacks of ensembling for a given task.

In considering transitioning competition models to real-world applications, it is important to note that there is a point of diminishing returns in the ensemble score as a function of the number of models combined. The point at which this occurs may vary depending on a number of factors. For example, the type of models being combined, the size of the dataset being used, and the specific problem being addressed can all impact the point at which the benefit of adding models becomes marginal. While even the marginal gains are crucial for securing the competitors' position on the leaderboard, the added computational complexity is often not warranted in the real-life deployment scenario, and ensembled models focused on narrow-margin, leaderboard optimization may not generalize well to new, real-world data.

3.5.6.5.1 Test Time Augmentation (TTA)

Another variant of ensemble is TTA, which combines the outputs of the same model on multiple augmented versions of the input. A horizontal flip augmentation at training time, paired with averaging the predictions of a flipped- non-flipped input at test time has been shown to be successful in the competition's top solutions.

3.5.6.6 Horizontal Flipping

Horizontal flipping is a simple form used almost universally by almost all of the competition's top submissions, such as winning solutions for the 1st, 2nd, and 4th place winners (DrivenData 2022). Coupled with applying the same augmentation during the training time, this simple trick can potentially give a boost to performance. Although, it does come at the cost of increasing the inference time by requiring successive model inference for each augmented sample. The tradeoff (performance versus inference time) has to be warranted through experimentation.

Although, a light form of an ensemble (like the combination of two distinct models (Blount and Holmberg 2022)) or TTA could be beneficial. It could also be possible that two lighter models might

yield a better result than a one large model of comparable complexity. We recommend measuring the performance of light ensembles for making the ultimate decision.

3.5.6.7 Other Details

The 4th place solution [4] used the Deep Orthogonal Local and Global (DOLG) framework (Yang et al. 2021) to obtain the descriptors. The DOLG framework modifies the neck of the CNN with two branches that generate descriptors corresponding to local and global features. This approach has also been applied in other ID competitions. The rest of the pipeline steps are similar to other solutions, including the loss function (ArcFace). DOLG has been shown to be capable of producing high-ranking solutions but is still being outperformed by the CNN backbones with simpler pooling mechanisms. Some competitors used it with high success while others stated that it did not work for them, probably because it is more sensitive to the dataset and setup specifics. Due to this, it is advised to hold experiments with DOLG as secondary to the simpler backbone + neck alternatives.

3.5.7 Visualization Challenge

In the competition's Explainability Bonus Round, two solutions [1][4] presented visualizations of how their solutions matched individual belugas. This visualization—often lacking in ML ID systems like PIE (Moskvyak et al. 2019) but present in older computer vision techniques like HotSpotter (Crall et al. 2013)—is critical for real-world, human validation of developed models. The approach taken for visualization implemented Grad-CAM [P].

In simple terms, Grad-CAM utilizes the gradients of the final model layer to create a heatmap that highlights regions in the image that contribute most to the excitation of neurons in the layer. It is interesting to note the differences in how the two available competitor solutions approached visualization.

The first competitor utilized Grad-CAM and measured activation of the final classification layer with respect to a certain output class (i.e., individual). This interpretation highlights the regions that contribute to the model identifying the image as belonging to a specific individual, without measuring pairwise relationships. The specific individual activation is possible because ArcFace is trained by classifying the individuals in the training set. The classifier layer weights are only available for the individuals on which the model was trained on.

The second competitor employed Grad-CAM++, a minor iteration over Grad-CAM. Importantly, they did not reference the activation of the classification layer. Instead, the target function was the cosine similarity between the embeddings of two images. This approach highlights the regions that contribute to the embeddings of two chosen images being similar. Thus, a pairwise relationship is measured, and the embedding comparison can be compared freely for individuals outside the training set.

There is also work done in exploring pairwise relationships can be queried in a point-specific manner by decomposing the activations (Zhu et al. 2021). The decomposition makes it is possible to choose a specific pixel point on the image and visualize the regions on both images that correspond to the 'matchiness' of the query point. However, the feasibility and computational complexity of this approach is yet to be evaluated.

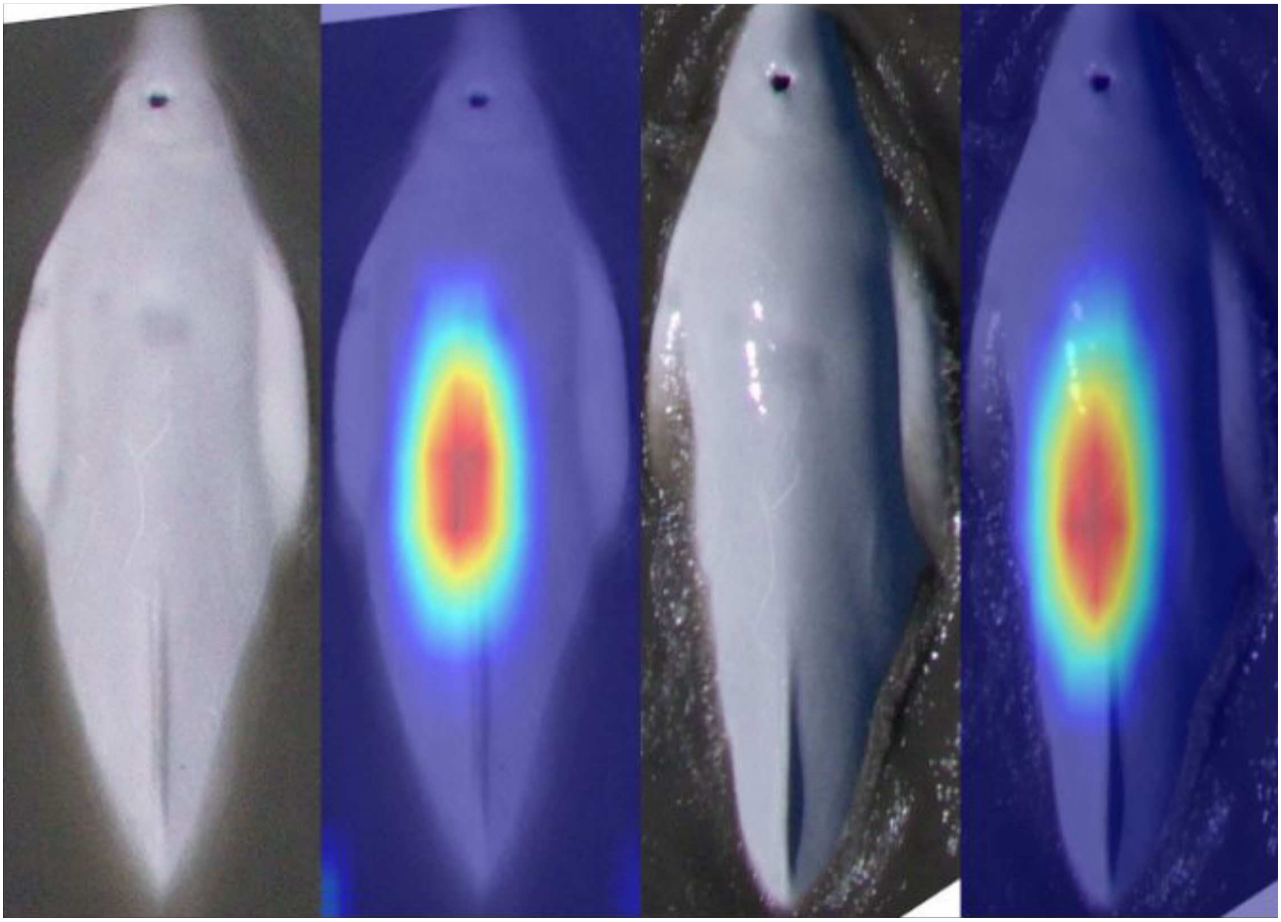


Figure 1. Grad-CAM-based Visualizations of Matched Beluga Images

4 Recommendations for Implementation

The usefulness of a wildlife identification pipeline is its ability to be widely applicable in various contexts, especially in real-world data collection and in future conditions (e.g., new data and previously unseen individuals). The main benchmarks for evaluating a new technique and pipeline—in addition of course to the accuracy of inference—are the ease of application and maintenance as real-world datasets invariably grow. The new observations can include previously unseen individuals and individuals with only a few annotations. The pipeline should continue to be useful with new observations, without the need for retraining or tweaking. Additionally, it will be beneficial if the pipeline is generalizable to other species with minimal modifications, demonstrating a much larger impact from the competition effort.

4.1 Comparing Competition Results to the Current State-of-the-Art

In order to evaluate the net impact of the competition versus the state-of-the-art in reusable re-ID algorithms, Wild Me trained a baseline PIE v2 model (Moskvyak et al. 2019). PIE has been shown to give strong individual matching results and has very good generalization ability across a diverse range of species and datasets. PIE is a contrastive learning pipeline with a CNN backbone trained using a triplet loss function with a hard-mining strategy. In contrast, "Where's Whale-do?" competitors used a similar capacity CNN backbone and an ArcFace loss with dynamic margins. We tested both PIE and the proposed simple baseline on the datasets of beluga whales and bottlenose dolphins (*Tursiops truncatus*), controlling for factors related to input preprocessing and postprocessing.

Data distribution for PIE training on belugas is shown in Figure 2.

```
** cross-set stats **

- Counts:
number of individuals in train:  388
number of annotations in train:  2343

number of individuals in test:  255
number of annotations in test:  896

average number of annotations per individual in train: 9.19
average number of annotations per individual in test: 3.51
```

Figure 2. Data Distribution for Baseline PIE Model Training

After 600 epochs of training, PIE v2 achieved mAP 31.7% on the test set and mAP 32.5% on a held-out validation set, with rank-1 ID accuracy averaging in the low 40th percentile and rank-20 in the low 70th percentile, as shown in Table 4.

Table 4. Baseline PIE v2 Results on Beluga Training Set

Scoring Metric	PIE v2 Test Results (%)	PIE v2 Validation Results (%)
mAP	31.7	32.5
Rank-1 ID	42.7	44.6
Rank-5 ID	58.1	58.8
Rank-12 ID	66.7	68.2
Rank-20 ID	70.7	72.2

These PIE v2 results are most comparable to scenario 1 in Table 2, and the winning mAP values from the competition, with the top-scoring team achieving ~0.495 mean average precision. Generally speaking, this means that the competitors achieved a clear advancement over the state-of-the-art by 15–20% mAP. This baseline test provides good justification to review the competition results and synthesize technical choices into a new approach for beluga (and potentially other species) re-ID.

4.2 Foundations for a New Re-ID Algorithm from Competitor Success

Wild Me's proposed approach to competition implementation involves adapting the following concepts (described in Section 1) in a new algorithm.

- Higher image size ~512 x 512px usage in the ML network
- Filter out individuals with less than *N* annotations, with the value of *N* to be determined
- A larger variety of augmentations
- Warmup and exponential decay scheduler
- EfficientNet Backbone
- GeM pooling + batchnorm1d head
- ArcFace loss derivatives (Sub-center ArcFace with dynamic margins)

4.2.1 Initial Testing

Partially implementing some of the winning elements above, especially ArcFace loss plus EfficientNet, Wild Me was able to demonstrate significant improvement over the state-of-the-art. These initial experiments were conducted with the purpose of acquiring insights into the feasibility of a single model approach trained through a pipeline that was built as a simplified version of the proposed experiment space. The pipeline establishes concrete initial results and serves as a starting point for further development within the experimental space.

After 30 epochs of training, the ArcFace baseline achieved mAP 45.9% on the test set and mAP 48.2% on a held-out validation set, with rank-1 ID accuracy averaging in the high 50th percentile and rank-20 in the low 80th percentile, as shown in Table 5.

Table 5. Initial Testing of Winning Elements without Ensembling on Belugas (aerial viewpoints only)

Scoring Metric	Competition Initial Implementation Test Results (%)	Competition Initial Implementation Validation Results (%)
mAP	45.9	48.2
Rank-1 ID	57.7	61.5
Rank-5 ID	70.9	71.2
Rank-12 ID	75.5	76.1
Rank-20 ID	80.9	80.8

It should be noted in comparing Table 4 and Table 5 (i.e., PIE vs new techniques), PIE was trained on an image resolution of 256 x 256, while the new baseline in Table 5 used a resolution of 440 x 440 pixels. We also trained PIE on the resolution of 440 x 440, but the results were worse compared to 256 x 256. It is possible that the weaker learning mechanism of PIE is not able to handle the extra complexity introduced. This limitation might partially explain the substantial difference between the outcomes of the two algorithms. Distinctive features of beluga whales are often subtle, and the model can vastly benefit from increased resolution.

There are strong indications of the superiority of the newly formulated approach. The new approach achieves better results in a significantly lower number of training epochs. The large difference between the number of training steps required can be explained by the nature of the core algorithms behind each pipeline. At each iteration, PIE only considers geometric distances between the embeddings of three specific annotations: the anchor, a positive (matching), and a negative (non-matching). This can be roughly interpreted as PIE v2 only learning in small increments and at each step taking into account only an extremely small local space while not considering the information of data outside the chosen triplets. To alleviate this drawback, PIE v2 uses a hard sample mining strategy and an auxiliary classification head but still takes a large number of epochs (200–400+) to approach an asymptote. In contrast, ArcFace projects the embeddings into an angular representation of vectors on a unit hypersphere. At each step, the comparison is done between the sample feature and the global class centers. In the past years, this formulation and its derivatives have been shown to dominate various identification tasks, as is the case here.

Furthermore, our findings suggest that removing individuals with a low number of annotations in the training set yields improved test results. These results resonate with the findings of one of the competitors. The testing set was kept unfiltered in the comparisons. The ideal minimum sample threshold is yet to be determined. Even though ArcFace can learn from as little as only one sample per individual, we have found significant benefits from filtering min-3 during training, just like for PIE v2.

There is an important caveat with ArcFace: the performance is very sensitive to the loss hyperparameters—margin and scale—during training. The choice of these parameters can explain up to 10% relative difference in evaluation rank 1. Optimal hyperparameter choice varies on a per-dataset basis and has to be tuned accordingly. In comparison, we have found PIE v2 to be a lot more robust to hyperparameter choice. Various papers following ArcFace have proposed a few rules of thumb with formulas for scaling the parameters based on the number of classes, but their effectiveness has yet to be thoroughly evaluated. From initial experiments, these guidelines prove to be suboptimal. Another approach could be to tune the hyperparameters on a pipeline using a scaled-down model and input size for faster iteration with hyperparameter optimization software (e.g., Optuna). This method has been shown to be effective for tuning in other identification competitions. Although this may increase computational demands, the total pipeline training run time required should still be lower than that of PIE.

4.2.2 Testing Potential Cross-Application to Other Species

While the “Where’s Whale-do?” competition focused exclusively on beluga re-ID, there is always the potential that competitors will develop solutions that can be generalized and cross-applied to other species. We were very pleased to see this potential exists also in at least the ArcFace+EfficientNet combination used by competitors, as we found when training a basic implementation on individually identified bottlenose dolphin fins. Dolphin fins are matchable by lateral photographs of their distinctive shapes, including nicks, notches, and sometimes research “brands” (Moore et al. 2022). Figure 3 provides a visual example of a matched fin. Dolphins also represent a very different visual matching challenge than the more subtle aerial images of belugas and therefore provide opportunity for comparison.

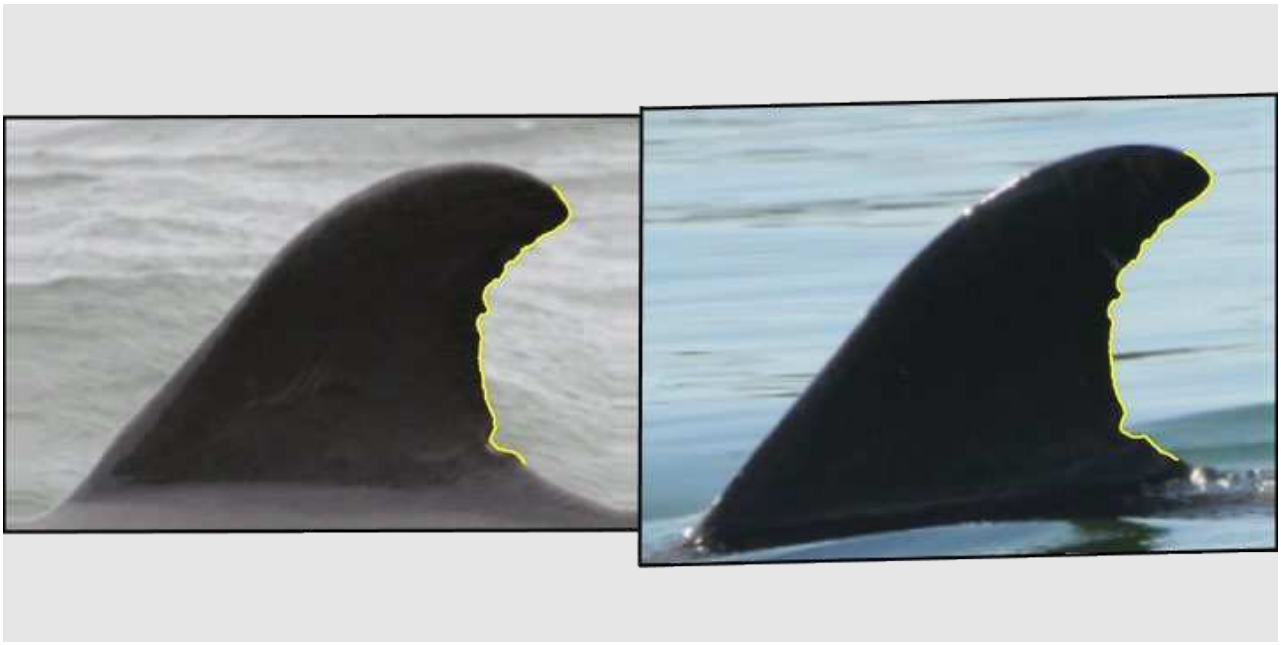


Figure 3. A Potentially Matched Bottlenose Dorsal Fin Visualized with a Simple Edge Trace

Note: The edge may or not be part of the network activation suggesting a match.

The EfficientNet+ArcFace baseline set up for the beluga experiments (Table 5) was retrained on an extant dolphin dataset provided separately by the Sarasota Dolphin Research Project. This quick baseline experiment provided significant improvement for bottlenose dolphin matchability too, yielding a 6–7% absolute increase in Rank-1 compared to PIE, as shown in Figure 4. Moreover, the results are achieved in significantly less epochs: 30 for ArcFace compared to 350 for PIE v2.

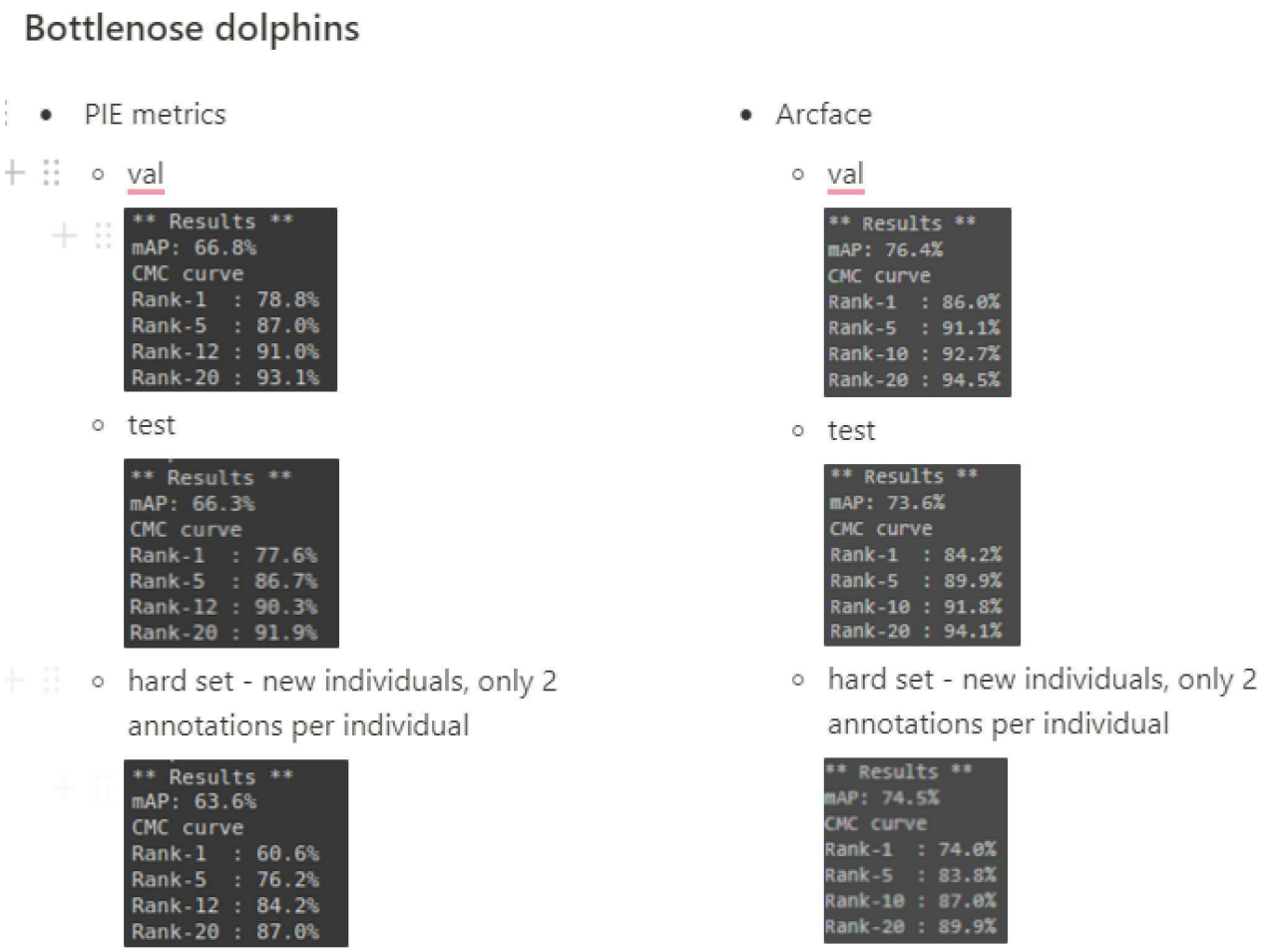


Figure 4. ArcFace+EfficientNet Test and Validation (val) Set Matching

Note: Includes performance analysis when only one other image is matchable in a catalog (hard set), for bottlenose dolphin fins.

Our initial concern was about the generalization of the ArcFace loss function to new individuals, due to learning from classification on the train set individuals. From the tests on the hard evaluation set with only new individuals, and a low number of annotations per individuals, the ArcFace baseline does even better, providing a 14% increase in Rank-1 matching—proving to be effective for generalizing individuals unseen in the training set.

4.3 Evaluation in Training

The metrics of evaluation chosen for this competition are generally accepted for information retrieval systems. The precision at K (P@K, rank-k) is an easily interpretable value. The P@K signifies the ratio of relevant results among the top-K retrieved documents. This is highly relevant for the reviewing process, where the reviewers have to make decisions from the K top-ranked results. Average precision at K (AP@K) is the mean of P@K over a range of values K. The AP@K function is a discrete approximation of the area under the precision-recall curve. The value is less directly interpretable than P@K. Although, by averaging the P@K on increasing cutoff points, this metric favors placing the correctly retrieved

documents higher up on the ranking. In contrast, P@K does not make that distinction, and the score relies only on whether the correct document happens to be at any place within the cutoff point. The additional information from AP@K is favored for reliable comparison of experiments and should be used in evaluating implementation results.

It is also recommended to control the number of annotations per individual during evaluation. Based on our experience, the model's performance increases with the number of samples per individual in the database, even when only taking the first matched annotation into account when assigning a name. Fine-grained model statistics can be generated by evaluating on subsets with a specific number of annotations per individual. The size of each subset with threshold N can be increased by randomly subsampling the annotations of individuals with more than N annotations.

Aside from the sets of specific number of samples per individual, there are other important scenarios to monitor. Since the visual characteristics of beluga whales change noticeably over time, cross-evaluation between annotations captured at different time gaps is recommended, similar to what is used in the competition. To control for individual memorization, a part of or a full evaluation set must consist of individuals not seen during training. The metrics can be monitored for these individuals separately.

Additionally, effects of the model/pipeline overfitting can be controlled by optimizing for the validation set metrics and reserving a holdout set of withheld individuals for final evaluation. Another valid approach is to set up a k-fold cross validation to enable using a larger subset of data for training. Ultimately, we tested with different variations of splits, sampled with different seeds. We also did k-fold cross validation for new individual threshold tuning. Those experiments showed that the test metrics remained stable, given the similar ratios for new individuals and average annotations per individual.

4.4 Inference—Transforming an Embedding to ID

During the inference process, matching the embeddings of output models to a specific name (ID) from the database is necessary. The matching can be done by leveraging a single closest annotation, or a group of annotations. However, during the evaluation between query and database sets, we explicitly impose a constraint that each query in the query set should not rely on the existence of other entries in the query set. This constraint is aligned with the competition rules. This limitation ensures closer alignment to the real-world query scenario, where annotations are added incrementally over different time periods.

One effective method is to assign the name of a database annotation whose embedding is most similar to the embedding of the query annotation. This straightforward approach has demonstrated robustness and consistently favorable results in ID competitions and real-world scenarios.

An alternative method involves aggregating the top-k retrieved annotations and assigning the most commonly occurring name within the results. This approach is reasonable when the distribution of annotations per individual is near-uniform. However, with an imbalanced distribution, the results may be biased toward individuals with a greater number of annotations in the database, leading to subpar outcomes compared to the simple approach of assigning the name of the first retrieved annotation.

The aggregated matching idea can be driver further using the embedding space. Once the similarity is calculated between all database samples, the individual cluster with the highest average similarity to the query can be chosen. From there, the sample with the highest similarity with the query can be presented for pairwise verification.

The first-place solution also used an approach leveraging the embedding space. For each query they retrieved the annotations of top 2 best matches. Then, the similarities were calculated for each of the three (query and two retrieved images) with the other database images. The calculated similarities are raised to

the power of 7. There is no specific reasoning given for the choice of exponent, but the end result is that exponentiation suppresses lower similarity scores more aggressively. Then, the scores of the three images (query, top1, top2) are combined using a weighted average with weights (0.4, 0.4, 0.2), yielding the final similarity score for ranking the possible candidates to be retrieved. They used the same method for doing inference with top-to-lateral models by first retrieving lateral matches with a top query and combining their similarities, and vice versa for lateral-to-top.

4.5 New Individual Classification

The geometry of embedding space could be utilized to classify individuals as previously unseen using a simple heuristic: if the annotation's cosine similarity is lower than a certain threshold to all of the annotations in the database, it can be classified as a new individual.

Hyperparameter optimization software can be used for finding the threshold for classification. To mitigate effects of overfitting, tuning will be done on the k-fold validation sets and performing and verifying the results on the holdout set. The statistical performance of such a classification method has to be thoroughly evaluated on holdout sets on multiple runs.

An important point to be determined is the expected universality of the found threshold: whether the threshold found on a specific dataset is generalizable across datasets and species or whether the threshold has to be tuned individually for each dataset. Also, how the thresholding approach holds up to a changing distribution of new and existing individuals.

4.6 Top View to Lateral Matching

Generally, we found that Scenarios 9 and 10 of the competition provided interesting results, but the lower mAP scores suggested yet another generational increase in AI capability may be needed for more accurate matching on the current dataset. Although, the fact that the competitors were able to achieve around 0.25 mAP for top-to-lateral evaluation cases, suggests that the algorithm is capable of learning such a relationship. To achieve such results, the competitors used sampling schemes which ensured that for each sample, both the lateral and top-view samples were present at each training step to the model. Given the low number of lateral samples—around 250 for each side, there is a possibility that the matching performance can get a significant boost by using the existing approaches on a dataset with a larger lateral sample size.

5 Implementation

Wild Me started formal algorithm development and testing as a new plugin to our ML platform (Parham et al. 2018) (WBIA 2023) under "Task 2.2.2.4 Integration of the automated matching algorithm into the database platform" in March 2023. Our proposed approach as outlined in Section 2 successfully captured the generational improvements in reidentification of individual animals that the competition successfully achieved in the context of belugas. We specifically retrained the new algorithm on the Cook Inlet data and deployed the winning solution for ongoing usage.

5.1 Catalog Integration

Under *Task 2.2.2.3 Catalog Integration*, Wild Me successfully imported the NOAA base beluga catalog into Flukebook.org in March 2023, preserving dates, locations, images, and individual ID. There were 7393 Encounter of 894 individuals imported.

All NOAA encounters:
https://www.flukebook.org/encounters/searchResults.jsp?genusField=Delphinapterus+leucas

All imported NOAA marked individuals:
https://www.flukebook.org/individualSearchResults.jsp?genusField=Delphinapterus+leucas

The numbers above have increased as NOAA has imported new data into Flukebook already. This foundational data was also used in the MiewId model training.

5.2 MiewId Algorithm Integration

With NOAA's approval of the next project steps provided in a meeting on April 7, 2023, Wild Me implemented and deployed this new algorithm with the new code base available at:

https://github.com/WildMeOrg/wbia-plugin-miew-id

Entitled "Matching and Interpreting Embeddings for Wildlife Identification" or "MiewId", the algorithm incorporates an ArcFace loss function and EfficientNet backbone.

5.2.1 Model Performance

In addition to implementing the new plugin for Wildbook Image Analysis, which is used by Flukebook.org for ML, Wild Me trained three MiewId AI models:

- Aerial-to-aerial matching for belugas
- Aerial-to-aerial+aerial-to-lateral for belugas
- Fin-matching for bottlenose dolphins (outside comparison to demonstrate future multispecies application)

5.2.1.1 Aerial-to-Aerial Matching Performance (Default)

Aerial-to-aerial image matching is the primary use case of NOAA for Cook Inlet Belugas. Wild Me trained an aerial-to-aerial matcher using MiewId and obtained results exceeding competition performance, as shown in Figure 5.

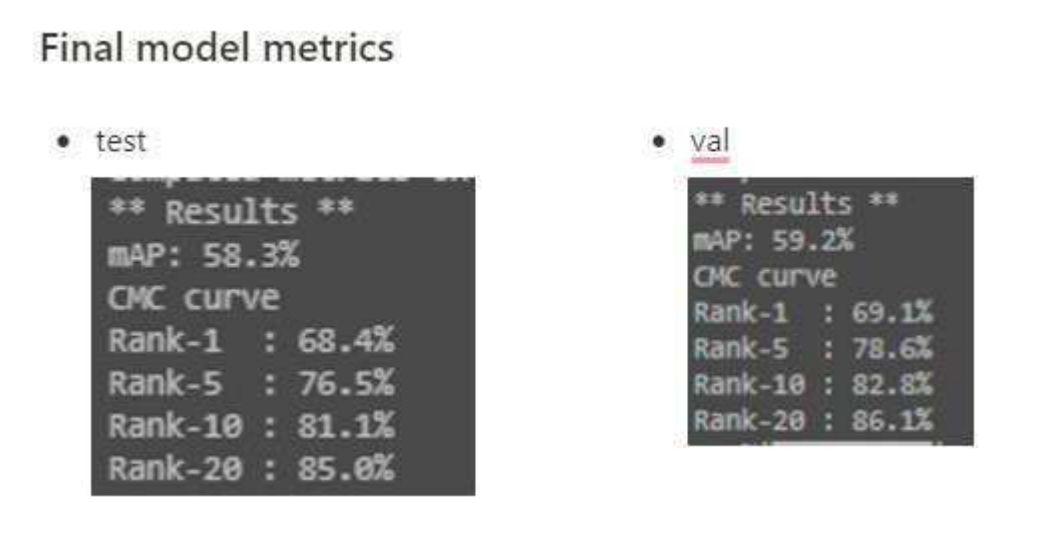


Figure 5. Aerial-to-aerial Top-k Matching Performance for Belugas

This model is now fully deployed in Flukebook, as shown in Figure 6.

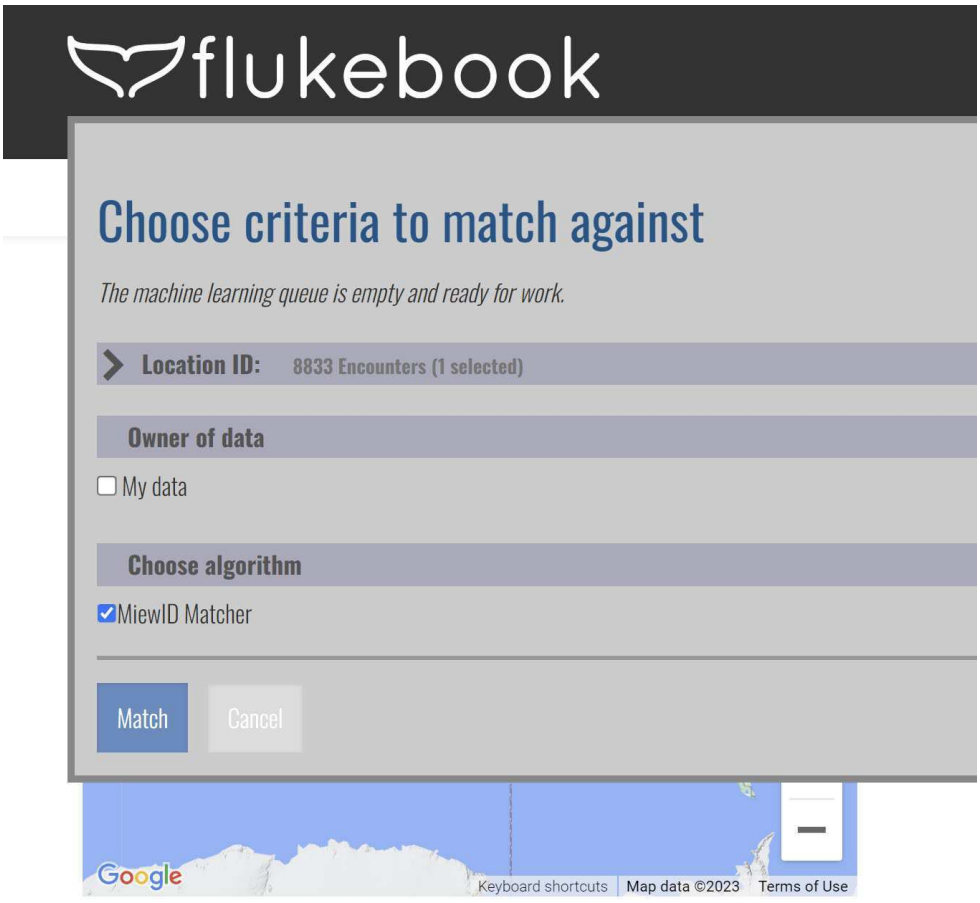


Figure 6. The MiewId Model for Belugas—Now Accessible in Flukebook from the Encounter Page

Note: The "Choose criteria to match against" dialog box allows for algorithm and location selection.

Figure 7 presents an example match in Flukebook as a ranked list of ID predictions. This result can also be accessed at Flukebook.org.

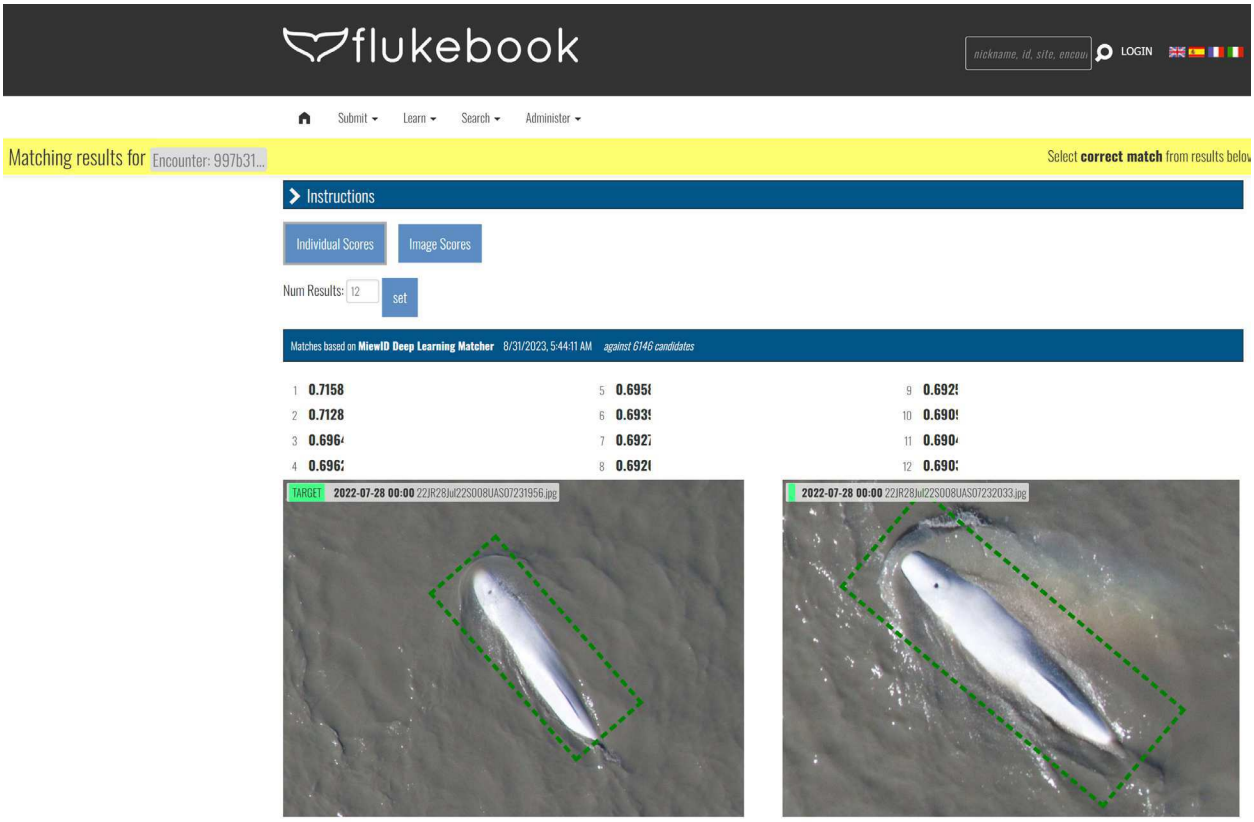


Figure 7. The Match Results Page for a MiewId Match

Note: Predicted identity results are ranked in order of prediction confidence.

Since only one MiewId model per species can be deployed currently in Flukebook, the aerial-to-aerial model is deployed for immediate use and to match the primary use case of NOAA.

5.2.1.2 Aerial-to-Lateral+Aerial-to-Aerial Matching Performance

For aerial-to-lateral matching, Wild Me trained a MiewId model to handle both aerial-to-aerial and aerial-to-lateral matching. Aerial-to-lateral matching in the combined model achieves accuracy similar to that of the competition, as shown in Figure 8.

- top to lateral

```
** Results **
mAP: 24.98%
CMC curve
Rank-1  : 20.14%
Rank-5  : 35.59%
Rank-12 : 47.59%
Rank-20 : 56.69%
```

- top to lateral

```
** Results **
mAP: 22.06%
CMC curve
Rank-1  : 21.89%
Rank-5  : 36.48%
Rank-12 : 48.50%
Rank-20 : 54.94%
```

Figure 8. Aerial-to-lateral Matching Top-k Performance in the Combined Viewed Model

Note: This model is not currently deployed.

In the combined model, aerial-to-aerial (top view) matching suffers significantly as model balancing required reducing available aerial-to-aerial training data. Figure 9 demonstrates the accuracy loss for aerial matching in the combined model.

- top-to-lateral model for top view only

```
** Results **
mAP: 44.6%
CMC curve
Rank-1  : 55.2%
Rank-5  : 69.1%
Rank-10 : 75.7%
Rank-20 : 80.8%
```

Figure 9. Top-k Matching Performance for Aerial Imagery Matching in the Combined Model

5.2.1.3 Bottlenose Fin ID

For comparison and to demonstrate the generalizability of the MiewId algorithm, Wild Me also retrained bottlenose dolphin fin ID matching, demonstrating Viewed as a generational improvement over PIE v2 in matching dolphin fins, as shown in Figure 10.

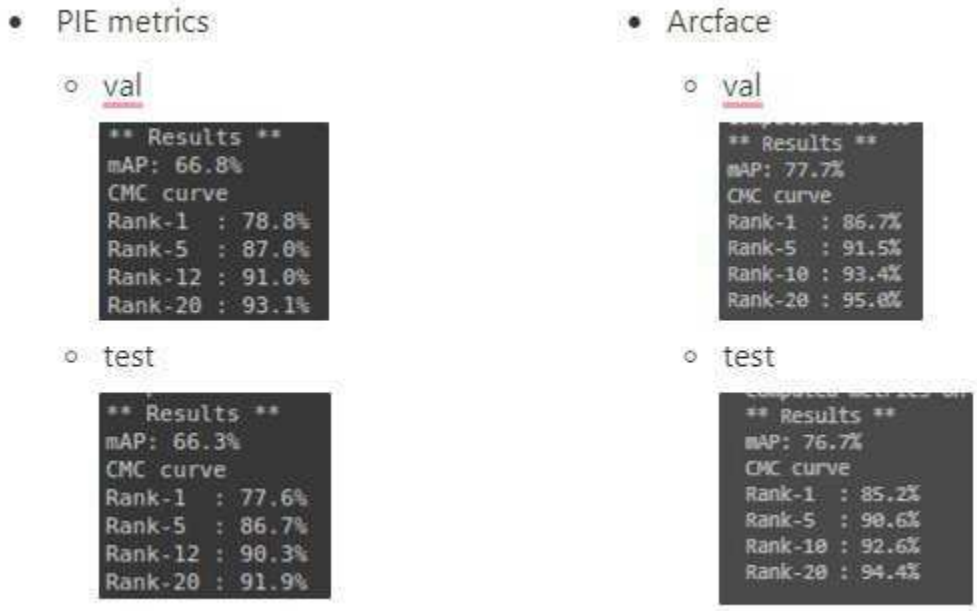


Figure 10. MiewId Significantly Outperforms PIE v2 Individual ID Matching in Top-1 (Rank 1) Results when Trained on the Same Data

6 Feedback and Iterative Improvements

With successful deployment of MiewId into Flukebook.org for belugas (contracted species) and bottlenose dolphins (broader impact and benefit for NOAA Hawaii researchers), Wild Me shifted to supported and iterative improvement based on NOAA feedback under *Task 2.2.2.5 Identify areas for automated matching process improvement*. Feedback from NOAA generated the following now completed work items.

6.1 NOAA Request: Implement Grad-CAM Rendering from Competition Bonus Prize Results

In our April 2023 meeting, NOAA approved Wild Me's recommendation to implement Grad-CAM-based rendering of MiewId result. Grad-CAM visualizations help users understand which areas of the body the MiewId algorithm was looking at in order to suggest its ranked list of matched encounters and individuals. These are now generated for each MiewId matching job for the top 20 results. Figure 11 and Figure 12 demonstrate how these can be accessed from the match page's **inspect** button.

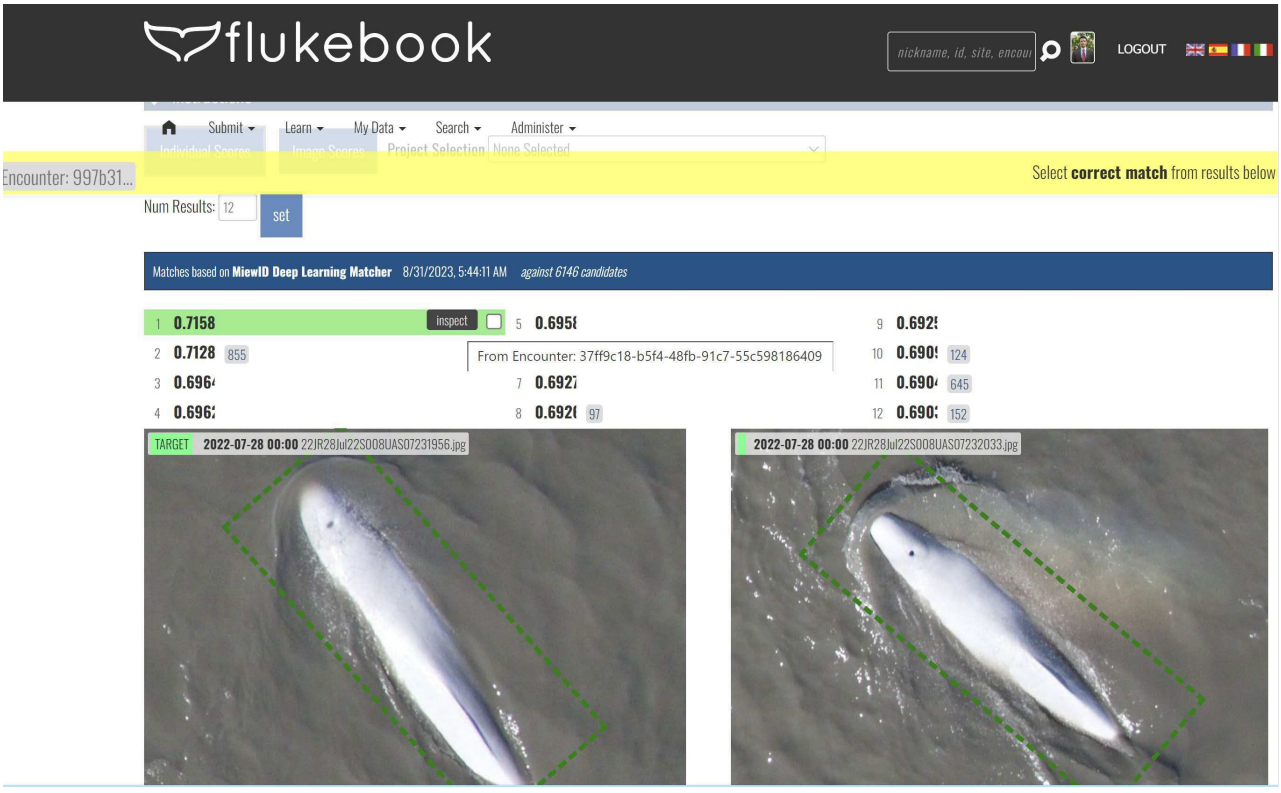


Figure 11. Inspect Button Results

Notes: The *Inspect* Button on the Match Results Page allows a Flukebook.org User to further inspect MiewId's suggested match result by accessing a Grad-CAM visualization of network activation as potential evidence of "why" the algorithm suggested the match.

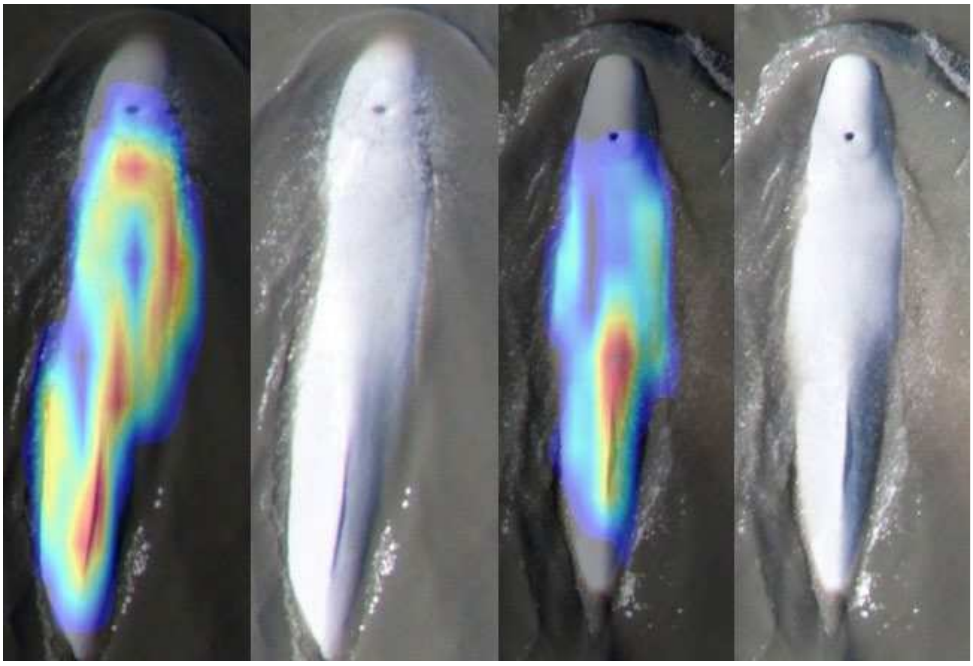


Figure 12. A Grad-CAM Visualization of the Image Areas that Activated the Network Behind a MiewId Beluga Identity Prediction

Care must be taken in interpreting these results: Grad-CAM visualizations are pairwise renderings of the potentially best matched annotation of an individual. However, MiewId matching is not based on pairwise annotation consideration but rather the total evidence of the cluster of annotations for an individual. Therefore, Grad-CAM visualizations are not an exact match rendering but rather closer to an interpretation of "Why did this match?" We noticed in MiewId development and training that what humans may focus their attention on in matching (e.g., scars) is not necessarily what MiewId found to be the most distinguishing set of pixels for an individual (e.g., dorsal ridge).

6.2 GPU Memory Reduction in Grad-CAM Rendering of MiewId Suggested Matches

During production use of MiewId, Wild Me engineers noticed high GPU usage by Grad-CAM when rendering MiewId match visualizations, limiting Flukebook.org's scalability to handle larger volumes of AI tasks for users. Wild Me made successive improvements to MiewId to reduce the memory footprint required by Grad-CAM renderings.

6.3 Support for Rotation Matching, Ensuring Real-World Photographs Did Not Need Pre-cropping

The base catalog of belugas provided to Wild Me included only pre-cropped and rotated images for AI training. However, real-world photography shows belugas at arbitrary angles. Wild Me modified the MiewId algorithm in August 2023 to ensure it could handle arbitrary beluga rotations in imagery, correcting detected annotations of belugas to the vertical orientation before matching and in Grad-CAM visualization. Figure 11 and Figure 12 demonstrate how Flukebook.org can detect a beluga in its real-world orientation in an image and match it at a standardized, corresponding vertical rotation, improving the algorithm's ability to find similarity.

6.4 NOAA Request: Allow Collaborator Access to a User's Bulk Imports

In August 2023, NOAA requested a security modification to Flukebook.org's handling of bulk imports, which were previously limited to only be visible to the uploading user and administrators. NOAA requested that a user's collaborators also have access to the import, allowing multiple NOAA staff to review a bulk import for data accuracy and match results. Wild Me completed this modification to Flukebook, which is of broad benefit to other users as well.

7 Broader Impacts

Wild Me completed over 100 GitHub commits of maintenance and improvements to ensure continuous improvement of MiewId since April 2023:

https://github.com/WildMeOrg/wbia-plugin-miew-id/commits/main

We have also successfully cross-applied MiewId to these cetacean species in addition to belugas:

- Bottlenose dolphins (NOAA user species)
- Spinner dolphins (NOAA user species)
- Long-finned pilot whales
- Short-finned pilot whales
- Minke whales
- Harbour porpoises
- Pacific white-sided dolphins
- Fraser's dolphins
- Risso's whales
- False killer whales
- Guiana dolphins

Because Wild Me works in support of an international cadre of marine and terrestrial research efforts, and because the MiewId algorithm emerging from the BOEM-funded competition was so successful in advancing the state-of-the-art of animal photo ID, MiewId models have also been trained and deployed for:

- African lion
- Cougar/mountain lion
- African leopard

While the above are not all species of interest, we recommend BOEM engage the U.S. public by presenting how AI research for belugas can have broad benefits in the study of other species.

Wild Me is continuing MiewId advancement beyond this project under separate BOEM funding, exploring multispecies training variants and how they improve algorithm accuracy for existing species and new species (i.e., "zero shot learning").

8 Works Cited

Blount A and Holmberg J (2022) Phase II of the Gray Whale Study: Development of an Ensemble Algorithm Foundation and a Gray Whale-specific Ensemble Algorithm. (Report No. 2022-064) Bureau of Ocean Energy Management (BOEM), Department of the Interior.

Brock A, De S, Smith SL, and Simonyan K (2021) High-Performance Large-Scale Image Recognition Without Normalization. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

Crall JP, Stewart CV, Berger-Wolf TY, Rubenstein DI, and Sundaresan SR (2013) HotSpotter-Patterned species instance recognition. In 2013 IEEE Workshop on Applications of Computer Vision, WACV 2013 (p. 230-237). [6475023] (Proceedings of IEEE Workshop on Applications of Computer Vision). https://doi.org/10.1109/WACV.2013.6475023.

Deng J, Guo J, Xue N and Zafeiriou S (2019) ArcFace: Additive Angular Margin Loss for Deep Face Recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4685-4694, doi: 10.1109/CVPR.2019.00482.

Deng J, Guo J, Liu T, Gong M, Zafeiriou S (2020) Sub-center ArcFace: Boosting Face Recognition by Large-Scale Noisy Web Faces. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12356. Springer, Cham. https://doi.org/10.1007/978-3-030-58621-8_43.

DrivenData (2022) Beluga's Winners: The Data Science for Good Challenge Winners. DrivenData Blog. https://drivendata.co/blog/belugas-winners. Accessed 2023-10-14.

Kaggle (2021) Happywhale - Whale and Dolphin Identification. https://www.kaggle.com/competitions/happy-whale-and-dolphin/discussion/320310. Accessed 2023-10-14.

Kingma D and Ba J. (2014) Adam: A method for stochastic optimization. arXiv:1412.6980.

Loshchilov, I. and Hutter, F. (2019) Decoupled Weight Decay Regularization. 7th International Conference on Learning Representations, New Orleans, 6-9 May 2019.

Moore R, Urian K, Allen J, Cush C, Parham J, Blount D, Holmberg J, Thompson J, Wells J (2022) Rise of the Machines: Best Practices and Experimental Evaluation of Computer-Assisted Dorsal Fin Image Matching Systems for Bottlenose Dolphins. Front. Mar. Sci. 07 April 2022.

Moskvyak O, Maire F, Armstrong AO, Dayoub F, Baktashmotlagh M (2019) Robust Re-identification of Manta Rays from Natural Markings by Learning Pose Invariant Embeddings. https://arxiv.org/pdf/1902.10847.pdf

Parham J, Stewart C, Crall JP, Rubenstein D, Holmberg J, and Berger-Wolf T (2018) An Animal Detection Pipeline for Identification. 1075-1083. 10.1109/WACV.2018.00123.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

Tan M. and Le QV (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, 9-15 June 2019, 6105-6114. http://proceedings.mlr.press/v97/tan19a.html

Wildbook Image Analysis (WBIA) Pipeline. https://docs.wildme.org/docs/researchers/ia_pipeline. Accessed 2023-10-14.

Yang M, He D, Fan M, Shi B, Xue X, Li F, Ding E, Huang J (2021) DOLG: Single-Stage Image Retrieval with Deep Orthogonal Fusion of Local and Global Features. https://doi.org/10.48550/arXiv.2108.02927

Zhu S, Yang T, and Chen C (2021) Visual Explanation for Deep Metric Learning. arXiv, 28 Aug. 2021. arXiv.org, http://arxiv.org/abs/1909.12977.

U.S. Department of the Interior (DOI)

DOI protects and manages the Nation's natural resources and cultural heritage; provides scientific and other information about those resources; and honors the Nation's trust responsibilities or special commitments to American Indians, Alaska Natives, and affiliated island communities.

Bureau of Ocean Energy Management (BOEM)

BOEM's mission is to manage development of U.S. Outer Continental Shelf energy and mineral resources in an environmentally and economically responsible way.