SeaTurtleID2022: A long-span dataset for reliable sea turtle re-identification

Lukáš Adam†, Vojtěch Čermák†
Czech Technical University
lukas.adam.cr@gmail.com
cermavo3@fel.cvut.cz

Kostas Papafitsoros†
Queen Mary University of London
k.papafitsoros@qmul.ac.uk

Lukas Picek†
UWB and INRIA
picekl@kky.zcu.cz
lpicek@inria.fr

† - Equal contribution in alphabetical order

Abstract

This paper introduces the first public large-scale, long-span dataset with sea turtle photographs captured in the wild – SeaTurtleID2022. The dataset contains 8729 photographs of 438 unique individuals collected within 13 years, making it the longest-spanned dataset for animal re-identification. All photographs include various annotations, e.g., identity, encounter timestamp, and body parts segmentation masks. Instead of standard "random" splits, the dataset allows for two realistic and ecologically motivated splits: (i) a time-aware closed-set with training, validation, and test data from different days/years, and (ii) a time-aware open-set with new unknown individuals in test and validation sets. We show that time-aware splits are essential for benchmarking re-identification methods, as random splits lead to performance overestimation. Furthermore, a baseline instance segmentation and re-identification performance over various body parts is provided. Finally, an end-to-end system for sea turtle re-identification is proposed and evaluated. The proposed system based on Hybrid Task Cascade for head instance segmentation and ArcFace-trained feature-extractor achieved an accuracy of 86.8%.

1. Introduction

Image-based individual animal re-identification, i.e., the process of recognizing individual animals based on their unique stable-over-time external characteristics, is essential for studying different aspects of wildlife, like population monitoring, movements, behavioral studies, and wildlife management [36,43,49]. The increasing sizes of the associated photo databases stemming from the multi-year span of such projects [42,45] have highlighted the need for automated methods to reduce labor-intensive human supervision in individual animal identification.

As a result, a plethora of automatic re-identification methods have been developed during the last years [8,16,29,50]. Evaluation of these methods is performed on benchmark databases, covering several animal groups like mammals [1,30,35,46,47,52], reptiles [4,16], and smaller organisms [10,20,40]. Typically, such databases are split into a reference set – a set of images with individuals whose identity (label) is known – and a query set – the set of images with individuals whose identity needs to be matched to the reference set. In deep learning, these sets are usually called training and test sets.

The quality of datasets influences the objectivity of the method evaluation. Therefore, the dataset and its splitting should mimic a realistic scenario, i.e., the images in the query and reference sets should not originate from the same encounters (burst mode in camera traps, consecutive video frames, multiple photographs taken during an encounter) and/or capture unknown identities. Other factors, e.g., different locations, image capture conditions, and images that reflect changes in animal appearances over time, are also vital. For reference see Figure 1.

Typically, images produced during one encounter share the same factors as the encounter lasts for a short period.
The most efficient way to indicate different encounters and factors in a dataset is by including the capture date and time in metadata, i.e., *timestamps*. Without knowing the time of the observation, datasets are often split into reference and query sets exclusively randomly. Therefore, images in training and test sets often originate from the same encounter/observation, representing unwanted training-to-test data leakage. This might result in overfitting to factors of a particular encounter instead of learning an inner representation of each individual. Thus, a random split implicitly assumes that one will encounter the same factors in the future, which is highly unrealistic. On the other hand, timestamps allow for time-aware splits, where images from a time period are all in either the reference or the query set. This leads to a more realistic case in which new factors are encountered in the future.

Based on our extensive research, just five publicly available datasets contain timestamps (see Table 1). From those, Cows2021 [21] and GiraffeZebraID [37] span only one month, and WhaleSharkID [26] includes timestamps for only 9% of photographs. This leaves only two wildlife datasets with timestamps; with span of at most two years. We introduce a novel dataset with photographs of loggerhead sea turtles (*Caretta caretta*) – the SeaTurtleID2022. The dataset was collected over 13 years and consists of 8729 high-resolution photographs of 438 unique individuals. Each photograph includes various annotations, e.g., identities, encounter timestamps, and body parts segmentation masks. To the best of our knowledge, the SeaTurtleID2022 is the longest-spanned public wild animal image dataset and the only public dataset of sea turtles with photographs captured in the wild. In contrast to existing datasets, the SeaTurtleID2022 allows for two realistic and ecologically motivated splits instead of a "random" split:

- **time-aware closed-set**: with reference images belonging to different encounters than query ones, and
- **time-aware open-set**: with new unknown individuals (i.e., newly introduced to population) in test and validation sets (common in ecology).

### Table 1. Dataset statistics for all publicly available animal re-identification datasets with timestamps; number of photographs, percentage of photographs with timestamps, number of individuals and encounters, and dataset span in days.

| Dataset                  | Images | t-stamp | ind. | enc. | span |
|--------------------------|--------|---------|------|------|------|
| Cows2021 [21]            | 8670   | 100%    | 181  | 3036 | 31   |
| GiraffeZebraID [37]      | 6925   | 100%    | 2051 | 2494 | 12   |
| MacaqueFaces [51]        | 6280   | 100%    | 34   | 494  | 525  |
| BelugaID [1]             | 5902   | 100%    | 788  | 1241 | 785  |
| WhaleSharkID [26]        | 7693   | 9%      | 98   | 424  | 1971 |
| SeaTurtleID2022 (ours)   | 8729   | 100%    | 438  | 1221 | 4390 |

Even though the SeaTurtleID2022 dataset is intended primarily as an animal re-identification benchmark, it can be used for the evaluation and testing of several fundamental problems, including: (i) object detection, (ii) instance segmentation, (iii) fully- and weakly supervised semantic segmentation, (iv) 3D reconstruction, and (v) concept drift analysis.

We stress that SeaTurtleID2022 lacks common drawbacks of other (human) re-identification datasets. In particular, face-id datasets typically contain low-resolution photographs, are restricted to limited poses, have limited time spans, and are either artificially generated [6], or collected by crawling the internet [27], raising privacy concerns.

Apart from the dataset, we provide a baseline performance evaluation for body-part segmentation and re-identification. Based on that, a baseline methodology for wildlife re-identification is proposed and evaluated over the SeaTurtleID2022 and several other well-known datasets using hand-crafted features and metric learning approaches. The best ArcFace-trained feature extractor achieved an accuracy of 69.2% on the SeaTurtleID2022 dataset while using cropped heads. In case no body part detection is done, the use of full images and the same approach resulted in an accuracy of 17.1%, showing that turtle identification is still a challenging task without body parts detection.

Furthermore, we showcase that time-unaware splits can often lead to performance overestimation if compared to time-aware splits. Hence, we recommend evaluating re-identification-focused algorithms over datasets with timestamps and unbiased (e.g., time-aware) splits. Additionally, imaging data collectors and database curators should ensure that time information is included in the metadata.

The main contributions of this paper are as follows:

- We introduce a novel dataset – SeaTurtleID2022 – for animal re-identification with unique characteristics and a wide variety of annotations, e.g., identities, encounter timestamps, segmentation masks, bounding boxes, and orientations for all body parts.
- We provide (i) baseline re-identification performance evaluation using hand-crafted features and metric learning approaches over SeaTurtleID2022 and other established datasets and (ii) baseline performance for body-part segmentation using well-known instance segmentation methods.
- We provide empirical evidence that a time-unaware splitting of the dataset leads to a significant overestimation bias.
- Based on all the above, we have developed and evaluated an end-to-end system for reliable sea turtle identification in the wild that can potentially be transferred to other species as well.
2. The SeaTurtleID2022 dataset

This section describes the data collection process, annotation procedures, and key features of the dataset.

2.1. Data collection

Location and species: All photographs were taken in Laganas Bay, Zakynthos Island, Greece (37°43’N, 20°52’E), from 2010 until 2022; May–October. Laganas Bay is a main breeding site for the Mediterranean loggerhead sea turtles [33]. Female turtles (around 300 annually) are mainly migratory and visit the island to breed every 2–3 years [42]. On the other hand, certain individuals reside on the island, and they can be observed in consecutive years [36, 43]. Loggerheads are long-lived species, and they can have reproductive longevity of more than three decades [32], which can lead to long-span image recordings for specific individuals. Sea turtles are particularly amenable to photo-identification due to their scale patterns [41]. In particular, the polygonal scales in the lateral (side) and dorsal (top) sides of their heads are unique to every individual and remain stable throughout their lives [11], see Figure 1 and additional examples in the supplementary material. Notably, the left and right side patterns differ for a given turtle.

Photographic procedure: All photographs were captured underwater during snorkeling surveys from a distance ranging from 7 meters to a few centimeters using three cameras: (i) Canon IXUS 105 digital compact camera with a Canon underwater housing in 2010–2013, (ii) Canon 6D full-frame DSLR camera combined with a Sigma 15mm fisheye lenses and an Ikelite underwater housing in 2014–2017, and (iii) the same camera with an additional INON Z330 external flash in 2018–2022. The resolution ranges from 4000×3000 (Canon IXUS) to 5472×3648 pixels (Canon 6D) with an average of 5269×3564. The water depth ranged from 1 to 8 meters, with the vast majority of photographs taken less than 5 meters deep.

Photographs taken in 2014–2022 are generally of better quality due to the use of a more advanced camera and a shorter camera-subject distance. On the other hand, due to the use of fisheye lenses, barrel shape distortion can be noticeable, especially for close-up photographs. Finally, more natural colors were acquired using the external flash. In Figure 2, we display three images of the same individual – obtained by the three different camera set-ups – to highlight the resulting visual differences.

2.2. Dataset highlights

Large-scale in the wild dataset: With 8729 photographs and 438 individuals, the dataset represents the most extensive publicly available dataset for sea turtle identification in the wild. The images are in original resolution and with various backgrounds. Approximately 90% of photographs have a size of 5472×3648 pixels, the average photograph size is 5269×3564 pixels, while the head occupies on average 635×554 pixels. Figure 3 shows the number of photographs for each individual. The majority of individuals (370/438) have at least ten photographs (depicted by the dashed line). Similarly, most individuals (369/438) were encountered at least twice. We note that this number is expected to increase in the following years since this dataset is updated annually.

Figure 3. Number of photographs for each of the 438 turtles. The orange line corresponds to 10 photographs.
Long time span & timestamps: The dataset contains photographs continuously captured over 13 years from 2010 to 2022. In contrast to most existing animal datasets that are usually collected in controlled environments and/or over a short time span, the SeaTurtleID2022 dataset includes a timestamp (in dd:mm:yyyy format) for each photograph. Figure 4 (left) shows the number of encounters for each year, with a significantly larger number from 2015 onwards. We note that this is driven by an increasing data collection effort rather than reflecting actual annual recurrence. In Figure 4 (right), we show the number of newly observed individuals. Furthermore, Figure 4 (middle) shows the distribution of the 438 individuals with respect to the total number of observation years. A span of one year means that a turtle was photographed only in one year. Many turtles (180/438) were photographed in at least two different years, and 9 individual turtles spanned more than 9 years.

Segmentation masks and bounding boxes: Almost all photographs in the dataset have a visible head and/or flippers. Therefore we provide body parts annotations photographs as segmentation masks and bounding boxes. Apart from masks, we include orientation (left, right, top, top-right, top-left, front or bottom) for each head mask, and orientation (top or bottom) and location (front left/right or rear left/right) for flipper masks. Such annotations allow further development and evaluation of turtle identification methods or novel methods for object detection and semantic segmentation. All segmentation mask annotations were done semi-automatically using the Segment Anything [28] model integrated within the CVAT.

Multiple poses: The dataset includes multiple images from different angles and, therefore, provides a ground for the challenging task of 3D animal reconstruction.

Comparison with ZindiTurtleRecall [4]: For a better perspective, we compare the SeaTurtleID2022 with the ZindiTurtleRecall dataset, which is the only other publicly available sea turtle dataset. We stress that the latter dataset contains photographs in a controlled environment (a rehabilitation center) with no timestamps. We summarise all comparable aspects of both datasets in Table 2.
2.3. Dataset splits and subsets

Standardly, the re-identification datasets are split into a reference (training) and a query set (test) randomly, which might result in unwanted data leakage and inflated performance. In other words, images from the same encounter might be in both sets. To illustrate the problem, we provide in Figure 6 four images of the same individual turtle, two captured in the same day in 2011 and two in the same day in 2021. While images from the same day are easy to match due to the same background and coloration, images from different days/years do not share it and therefore are significantly more challenging to match. To overcome this issue, we introduce two realistic ecologically motivated splits that utilize timestamps to prevent information leakage from the test set to the training set. We further refer to those splits as time-aware splits. For easier future comparison, we provide predefined training/validation/test splits even though validation in some cases is not mandatory. The construction is further elaborated below. The dataset statistics, including the number of individuals and images, are listed in Table 3.

### Time-aware closed-set split

The closed-set split is similar to a standard closed-set re-identification scenario, as all validation/test sets identities are available for training. Such a scenario is realistic in environments with controlled populations, e.g., zoos or reservations. While constructing the split, we group all the data based on the date of acquisition and split it in a time-aware fashion. Data from approximately 80% days are selected for the development set (training + validation), and the remaining days go to the test set. If an individual turtle was observed just once, it was kept for training. We provide 438 identities for training and 270 for testing. The development set was split into training/validation subsets using the same strategy.

### Time-aware open-set split

The open-set split is based on cutoff time points (specific years). In this setting, each subset (training/validation/test) contains all images within given subsequent periods. Intuitively, this split results in an open-set problem, reflecting the natural population dynamic and growth. During construction, we used the 2010–2018 period for training, the whole year of 2019 for the validation, and the 2020–2022 period for the test set. There are 357 identities in the training set and 151 in the test set. Out of the 151 identities, 51 are newly observed. A similar ratio (new/known) is naturally acquired in the validation set; 38 out of 83 are new identities.

| Subset   | # of images | # of identities |
|----------|-------------|-----------------|
| Training | 4679        | 438             |
| Validation | 1418    | 91              |
| Test     | 2632        | 270             |

Table 3. Provided time-aware datasets split and their statistics.

**Note:** The open-set split is much closer to the real-world re-identification settings than the closed-set problem. Therefore, the open-set split should be preferred for automated method evaluation over all datasets. In case closed-set evaluation is desired, then the time-aware split must be the preferred option over the random split.

### Body-parts subsets

Furthermore, we provide three subsets that cover various body parts, e.g., full-body, flippers, and heads, using crops from the original resolution. The number of data points differs for each body part, as some parts might not be visible. We used the time-aware closed-set and constructed part-based sets with the following number of training/test samples: (i) 6139 / 2650 full turtle bodies, (ii) 14849 / 6237 flippers, and (iii) 5956 / 2583 heads.

3. Sea turtle re-identification baselines

Animal re-identification is generally approached using either (i) traditional methods and local descriptors (e.g., SIFT and SURF) [5, 19, 39], (ii) deep learning [10, 30, 48], or (iii) species-specific methods [7, 22, 50]. To establish a baseline performance on the SeaTurtleID2022 and to propose the system for end-to-end turtle identification, we perform various ablation studies using various traditional and deep learning based methods. In this section, we describe selected methods and all relevant hyperparameters.
3.1. Local feature-based methods

The most popular methods used for wildlife re-identification – Wild ID [9], and Hotspotter [19] – are based on local descriptors. Therefore we study the performance of SIFT and more recent Superpoint [18] descriptors on the proposed dataset. We have developed a straightforward algorithm (inspired by Dunbar et al. [19]) based on local descriptors matching. First, we extract a set of keypoints and their corresponding descriptors for each image. Second, for all possible training-test image pairs, we calculate the distance between their descriptors. Third, all potentially false matches are filtered out using a ratio test and threshold; the optimal values (0.2 for SIFT, 0.6 for Superpoint) for the ratio test thresholds were obtained from the training set. At last, we predict an identity using the training label with maximal similarity score, calculated as an absolute number of correspondences. We opt not to use alternative approaches, such as RANSAC or SuperGlue, as they add significant computational overhead and provide just a small improvement [19, 38].

3.2. Metric learning

Metric learning methods aim to learn a representation function that maps objects into a deep embedding space. Usually, a CNN- or transformer-based feature extractor is trained to group samples within the same semantic category closer and far from other categories. For our experiments, we use two algorithms with state-of-the-art performance in face recognition: ArcFace [17] and Triplet loss [44]. For baseline performance evaluation, we use a Swin-B [31] backbone and default training hyperparameters.

Both metric learning approaches were optimized for 100 epochs using a learning rate of 0.01, the cosine annealing schedule, and a mini-batch size of 128. All images were pre-processed using the Random augment method.

ArcFace loss [17] was designed for face recognition but can be easily repurposed for wildlife re-identification. It extends the cross-entropy loss by placing the embeddings on the hypersphere with radius $s$ and incorporating an angular margin $m$ to improve the learned embeddings’ discriminative capability that ensures high inter-class variety while keeping a high level of intra-class compactness. The similarity of samples is determined using cosine distance. We use the same values for $s = 64$ and $m = 0.5$ as in [17].

In Triplet loss [44], we select triplets $(x_a, x_p, x_n)$ with anchor $x_a$ that has the same label as positive $x_p$ and a different label than negative sample $x_n$. Triplet loss learns a representation that minimizes the distance between $x_a$ and $x_p$ and maximizes the distance between $x_n$ up to a margin $m$. In our experiments, we follow [44] and use $m = 0.1$. Triplet loss tends to be sensitive to triplet selection. Therefore, we follow [25] and select hard triplets using an online mining strategy to improve the training.

**Feature matching:** In our metric learning experiments, we approach animal identification using k-NN classifier in a deep embedding space. For each image $x$ from the test set, we assume its $k$ most similar training set identities, and we take the one with the highest occurrence. The formal definition of k-NN we use is as follows. The set of $k$ nearest neighbors $S_x$ of $x$ is defined as a subset of the training set such that every point in the training set but not in $S_x$ is at least as far away from $x$ as the furthest point in $S_x$, measured in a suitable distance function. We define the classifier as a function returning the most common label in $S_x$.

In the case of a draw, we take an identity from smaller $k$, i.e., $(k - 1)$. For metric learning approaches, the distance function is a cosine distance, i.e.,

$$\text{dist}(x, z) = \frac{x \cdot z}{\|x\| \|z\|}, \quad (1)$$

3.3. Random vs. time-aware splits

To showcase the unwanted performance overestimation when a random dataset split is used, we compare the performance of newly proposed time-aware splits (open and closed) with their random counterparts. The random split is obtained by randomly shuffling the time-aware split for each identity separately. This ensures a fair comparison between the split with the same training/validation/test ratio. We used the entire image and different body parts in this experiment. We use an ArcFace loss with the Swin-B backbone and input size of $224 \times 224$.

4. Baseline Results

In this section, we provide (i) baseline results for body-part segmentation and re-identification achieved over the newly proposed dataset (ii) qualitative and quantitative evaluation to show the importance of the time-aware splits, and (iii) performed ablation studies to select the most viable approach for sea turtle re-identification.

Based on extensive experiments with different $k$ values for k-NN matching (available in Supplementary), we predict an identity using k-NN, with $k = 1$.

**Local vs deep features:** Comparing local descriptors with metric learning approaches showed superior performance of metric learning on our dataset and seven other datasets with patterned species. In most cases, the metric learning approaches outperformed the Superpoints by more than 20%. Furthermore, if we compare local descriptor methods, the Superpoints method is a better fit for animal re-identification. A detailed comparison is listed in Table 4.
Table 4. Local and deep feature methods performance comparison (accuracy) for full images. Time-aware closed-set split. Input size $224 \times 224$. For metric learning, a Swin-B backbone was used.

| Dataset                | SIFT | Superpoint | ArcFace | Triplet |
|------------------------|------|------------|---------|---------|
| BelugaID [1]           | 1.1  | 2.4        | 18.2    | 20.5    |
| HumpbackWhaleID [2]    | 11.7 | 11.8       | 52.5    | 43.9    |
| NDD20 [46]             | 17.1 | 30.0       | 59.1    | 29.9    |
| NOAARightWhale [3]     | 6.5  | 15.3       | 23.5    | 5.4     |
| WhaleSharkID [26]      | 4.3  | 22.9       | 28.6    | 32.5    |
| ZindiTurtleRecall [4]  | 17.9 | 25.7       | 45.8    | 19.1    |
| SeaTurtleID2022 (ours) | 8.4  | 20.2       | 34.7    | 25.7    |

Table 4. Local and deep feature methods performance comparison (accuracy) for full images. Time-aware closed-set split. Input size $224 \times 224$. For metric learning, a Swin-B backbone was used.

Body parts performance: In addition to setting overall full-body turtle performance, we explored the importance of various body parts, revealing their relative significance. In contrast to the findings of [34], our results highlight the key role of the turtle’s head in sea turtle identification. Focusing solely on the head increased the absolute performance by 34.5% compared to the full body. Furthermore, we show that the flippers appear as the less influential body part for in-the-wild identification using metric learning. The full comparison is provided in Table 5.

Encounter based prediction: Available timestamps allow combining all image-based predictions into so-called encounters. Rather than identifying each image separately, one identity is predicted for each set of images belonging to one individual. Using just majority voting to combine the image-based predictions, we significantly increased the performance for all body parts (see Table 5). In the case of heads, the accuracy was increased by 19.2%.

4.1. Random vs time-aware splits

The performance comparison of two ArcFace-trained feature extractors on the random and time-aware splits of the SeaTurtleID2022 dataset validated our hypothesis about unwanted performance inflation related to training-to-test data leakage. Results listed in Table 5 demonstrate that the random split results (in terms of accuracy) were higher by 42.2%, 53.8%, 45.8%, and 18% for full image, and flippers, body, and head crops, respectively.

Table 5. Random split accuracy inflation on SeaTurtleID2022 (closed-set). Encounter- vs image-based; Swin-B + ArcFace; $224 \times 224$.

| Split          | Full image | Flippers | Turtle | Head |
|----------------|------------|----------|--------|------|
| Images         | 17.1       | 12.2     | 34.7   | 69.2 |
| Encounters     | –          | 21.4     | 48.6   | 88.4 |
| Images Random  | 59.4       | 66.0     | 80.5   | 87.2 |

Table 6. Random split accuracy inflation on SeaTurtleID2022 (closed-set). Swin-B + ArcFace; $224 \times 224$.

| Split          | Full image | Background | Foreground |
|----------------|------------|------------|------------|
| Random         | 59.4       | 45.1       | 59.5       |
| Time-aware     | 17.1       | 3.9        | 14.3       |
| $\Delta$       | +42.2      | +41.2      | +45.2      |

Table 6. Random split accuracy inflation on SeaTurtleID2022 (closed-set). Swin-B + ArcFace; $224 \times 224$.

Furthermore, we qualitatively demonstrate overfitting to the background using Grad-CAM++ [12] and visualizing identity activations based on the cosine similarity between the embeddings of the two images. We selected two similar images with noticeable backgrounds from the same encounter that are in the test set for both random and time-aware splits. In Figure 7, we illustrate that the model trained on the random split learns to utilize background features, whereas the model trained using the time-aware approach concentrates on the turtle’s features.

Figure 7. Qualitative evaluation demonstrating overfitting to the background on random split using Grad-CAM++. Identity-based activations for (left) time-aware and (right) random split.

2For the flippers performance evaluation, we choose the closest (based on cosine similarity) identity using all available flippers on a given image.
4.2. Body-parts segmentation baselines

The SeaTurtleID2022 dataset comes along with instance segmentation annotations; thus, it might be used as a benchmark for instance segmentation or object detection. To set the baseline performance for the turtle body parts (head, flipper, and full-body) segmentation, we evaluate three distinct architectures, including the standard Mask R-CNN [23], the Hybrid Task Cascade (HTC) [13], and the state-of-the-art transformer-based Mask2Former [15]. We combine the three detection methods with two backbones, ResNet-50 [24] and Swin-B transformer [31] using the MMDetection [14] framework. While training, both backbones were initialized from publicly available ImageNet-1k weights using the default implementation and hyperparameters setting. All models were fine-tuned for 12 epochs with a step-wise learning rate schedule. Experiments are conducted on both time-aware splits.

Generally, all selected methods evaluated on the SeaTurtleID2022 achieved a competitive performance (in terms of coco mAP) suitable for the following task, i.e., turtle re-identification. While the best-performing model – Mask2Former with Swin-B backbone – achieved a coco mAP of 0.896, the worst-performing model – Mask R-CNN with ResNet-50 backbone – achieved an mAP of 0.865. Even though the Mask2Former approach showed better overall performance, the HTC method performed better on heads that are important for accurate re-identification.

The full performance comparison on both time-aware splits (open and closed) is available in Table 7 and supplementary materials.

| Method         | mAP | head   | turtle | flippers |
|----------------|-----|--------|--------|----------|
| Mask R-CNN     | 0.865 | 0.838  | 0.910  | 0.848    |
| HTC            | 0.868  | 0.842  | 0.912  | 0.852    |
| Mask2Former    | 0.892  | 0.822  | 0.977  | 0.876    |

Table 7. Instance segmentation performance of selected backbone and head architectures over the SeaTurtleID2022. Closed-set split.

5. Recommended end-to-end system

Following the insights from our baseline experiments allowed us to create a reliable end-to-end system that takes sets of images as input and returns identity predictions. The system within the pipeline and the performance of the system are fully described below.

First, we find all head region bounding boxes on high-resolution images (20MP) using the Hybrid Task Cascade instance segmentation model (with Swin-S backbone). We focus primarily on turtle heads as they allow the best re-identification capability. Second, we crop all heads from the high-resolution photographs and re-scale them to 224 × 224 to match the expected input size for the feature extractor, i.e., the Swin-B ArcFace-trained re-identification model. Third, all head-based crops are feed-forwarded into the feature extractor to obtain feature vectors for matching. For images without a head segmentation, we do not provide any identity prediction. Fourth, for each image, we predict an identity using k-NN (with k = 1) with the training set’s head embeddings. Finally, we group all images based on time and create the so-called encounters. The identity of each image within an encounter is retrieved by majority voting.

The proposed end-to-end system for sea turtle re-identification achieved an accuracy of 86.8% on the SeaTurtleID2022. Notably, it shows a significant improvement over the 17.1% accuracy achieved by a naive approach that analyzes full images without utilizing body parts or harnessing encounter knowledge.

6. Conclusions

This paper introduced the SeaTurtleID2022 dataset, the longest-spanned publicly available wildlife re-identification dataset with various annotations, e.g., identities, encounter timestamps, and body parts segmentation masks. The dataset can be used for benchmarking re-identification algorithms and several other computer vision tasks, including instance and semantic segmentation and object detection. Instead of a standard "random" split, we highlight the necessity to use realistic and ecologically motivated splits: (i) time-aware: with reference data from different encounters, and (ii) open-set: with new unknown individuals (i.e., newly introduced to population) in test and validation sets.

Furthermore, (i) we provided a baseline performance of various methods, for instance segmentation and animal re-identification, (ii) provided qualitative and quantitative evidence that time-unaware (random) splits of the dataset lead to a significant performance overestimation bias, and (iii) we proposed, described, and evaluated an end-to-end system for sea turtle identification in the wild, that could potentially be transferred to other species as well.

7. Acknowledgments

This research was supported by the Czech Science Foundation (GA CR), project No. GA22-32620S and by the Technology Agency of the Czech Republic, project No. SS05010008. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic and by the OP VVV project “Research Center for Informatics” (No. CZ.02.1.01/0.0/0.0/16_019/0000765).
References

[1] Beluga ID 2022. https://lila.science/datasets/beluga-id-2022. Accessed: 4-11-2023. 1, 2, 7

[2] Humpback whale identification. https://www.kaggle.com/c/noaa-right-whale-identification. Accessed: 4-11-2023. 7

[3] Right whale recognition. https://www.kaggle.com/c/right-whale-recognition. Accessed: 4-11-2023. 7

[4] Turtle recall: Conservation challenge. https://zindi.africa/competitions/turtle-recall-conservation-challenge. Accessed: 4-11-2023. 1, 4, 7

[5] William Andrew, Sion Hannuna, Neill Campbell, and Tilo Burghardt. Automatic individual holstein friesian cattle identification via selective local coat pattern matching in RGB-D imagery. In 2016 IEEE International Conference on Image Processing (ICIP), pages 484–488. IEEE, 2016. 5

[6] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jiingiing Shen. Digitface-1M: 1 million digital face images for face recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 3526–3535, 2023. 2

[7] Anka Bedetti, Cathy Greyling, Barry Paul, Jennifer Blondeau, Amy Clark, Hannah Malin, Jackie Horne, Ronny Makukule, Jessica Wilmot, Tammy Eggelting, et al. System for elephant ear-pattern knowledge (seek) to identify individual african elephants. Pachyderm, 61:63–77, 2020. 5

[8] Drew Blount, Shane Gero, Jon Van Oost, Jason Parham, Colin Kingen, Ben Scheiner, Tanya Stere, Mark Fisher, Gianna Minton, Christin Khan, Violaine Dulau, Jaime Thompson, Olga Moskvyk, Tanya Berger-Wolf, Charles V Stewart, Jason Holmberg, and J Jacob Levenson. Flukebook: an open-source AI platform for cetacean photo identification. Mammalian Biology, pages 1–19, 2022. 1

[9] Douglas T Bolger, Thomas A Morrison, Bennet Vance, Derek Lee, and Hany Farid. A computer-assisted system for photographic mark-recapture analysis. Methods in Ecology and Evolution, 3(5):813–822, 2012. 6

[10] Joakim Bruslund Haurum, Anastasija Karpova, Malte Pedersen, Stefan Hein Bengtsson, and Thomas B Moeslund. Re-identification of zebrafish using metric learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, pages 1–11, 2020. 1, 5

[11] Alice S Carpentier, Claire Jean, Mathieu Barret, Agathe Chassagneux, and Stéphane Ciccone. Stability of facial scale patterns on green sea turtles chelonia mydas over time: A validation for the use of a photo-identification method. Journal of Experimental Marine Biology and Ecology, 476:15–21, 2016. 3

[12] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE winter conference on applications of computer vision (WACV), pages 839–847. IEEE, 2018. 7

[13] Kai Chen, Jiayao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4974–4983, 2019. 8

[14] Kai Chen, Jiaqi Wang, Jiayao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiariu Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155, 2019. 8

[15] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Perpixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems, 34:17864–17875, 2021. 8

[16] Jonathan P Crall, Charles V Stewart, Tanya Y Berger-Wolf, Daniel I Rubenstein, and Siva R Sundaresan. Hotspotter-patterned species instance recognition. In 2013 IEEE Workshop on Applications of Computer Vision (WACV), pages 230–237. IEEE, 2013. 1

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4690–4699, 2019. 6

[18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 224–236, 2018. 6

[19] Stephen G Dunbar, Edward C Anger, Jason R Parham, Colin Kingen, Marsha K Wright, Christian T Hayes, Shahnaj Safi, Jason Holmberg, Lidia Salinas, and Dustin S Baumbach. Hotspotter: Using a computer-driven photo-id application to identify sea turtles. Journal of Experimental Marine Biology and Ecology, 535:151490, 2021. 5, 6

[20] André C Ferreira, Liliana R Silva, Francesco Renna, Hanja B Brandl, Julien P Renoult, Damien R Farine, Rita Covas, and Claire Doutrelant. Deep learning-based methods for individual recognition in small birds. Methods in Ecology and Evolution, 11(9):1072–1085, 2020. 1

[21] Jing Gao, Tilo Burghardt, William Andrew, Andrew W Dowsey, and Neill W Campbell. Towards self-supervision for video identification of individual holstein-friesian cattle: The cows2021 dataset. arXiv preprint arXiv:2105.01938, 2021. 2

[22] Andrew Gilman, Krista Hupman, Karen A Stockin, and Matthew DM Pawley. Computer-assisted recognition of dolphin individuals using dorsal fin pigmentation. In 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ), pages 1–6. IEEE, 2016. 5

[23] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017. 8
A. Additional Experiments and Evaluation

A.1. Dependence of identification accuracy on $k$ in $k$-NN classification

We conducted additional experiments to find an optimal $k$ value for the animal re-identification using the $k$-NN classifier. Besides SeaTurtleID2022 (head and full-body versions), we evaluated the experiments on BelugaID, NDD20, WhaleSharkID, HumpbackWhaleID, NOAARightWhale and ZindiTurtleRecall datasets. We used embeddings from the ArcFace-trained model.

Our findings indicate that opting for a smaller $k$ value yields better results, with $k=1$ being a reasonable choice in any case. This discovery is consistently supported by results in various other datasets we considered. We attribute this phenomenon to the significant class imbalance present in wildlife datasets. As $k$ increases, identities with higher prior probability overwhelm the classification results, i.e., for larger $k$ values, there are often just a few samples for the less frequent identities. On the SeaTurtleID2022 dataset (head and full-body) the performance in terms of accuracy significantly decreased from 69.2% at $k = 1$ to 55.0% at $k = 100$. A similar, though less severe, drop in performance was also noticeable in other datasets. We depict the relationship between accuracy and values of $k$ in Fig. 8.

A.2. Time-aware vs random split: Additional experiment with cross-entropy learning

To further elaborate the performance inflation related to random split, we have tested various deep learning backbone architectures optimized using softmax cross-entropy. In Tab. 8, we provide the performance of five architectures on two splits of the SeaTurtleID2022 dataset: time-aware and a random split. In Tab. 9, we perform a similar experiment on 3 other datasets that allow time-aware splitting, showcasing that this inflation is not a characteristic of the SeaTurtleID2022 dataset, but it occurs in other datasets as well. We employed a 50/50 training-test split; therefore, results are directly not comparable to results in Section 4.1. In all experiments, all images were resized to match the pre-trained model input size of $224 \times 224$.

### Table 8. Performance inflation (accuracy) with different backbones fine-tuned with softmax cross-entropy.

| Backbone      | Time-aware close-set | Random split  |
|---------------|----------------------|---------------|
| ResNeXt-50    | 38.6%                | 63.4%         |
| EfficientNet-B0 | 39.9%                | 76.5%         |
| ConvNeXt-B    | 47.2%                | 78.5%         |
| ViT-Base/p32  | 45.2%                | 82.5%         |
| Swin-B/p4w7   | 47.6%                | 83.2%         |

### Table 9. Performance inflation (accuracy) with different datasets.

| Dataset            | Time-aware close-set | Random split |
|--------------------|----------------------|--------------|
| BelugaID           | 7.8%                 | 12.1%        |
| GiraffeZebraID     | 2.1%                 | 30.1%        |
| MacaqueFaces       | 91.1%                | 98.9%        |

A.3. Body-part instance segmentation

We further present additional baseline instance segmentation experiments for different turtle body parts (head, flipper, and full-body). We provide an evaluation of three architectures, e.g., Mask R-CNN, the Hybrid Task Cascade (HTC), and the state-of-the-art transformer-based Mask2Former, on the time-aware open-set split. We used the same training strategy, i.e., backbones were initialized from publicly available ImageNet-1k checkpoints using the default implementation and hyperparameters setting and fine-tuned the models for 12 epochs with a step-wise LR schedule.
The comparison of selected methods utilizing well-known CNN- and transformer-based backbone architectures on the time-aware open-set split of the SeaTurtleID2022 dataset validated findings from the initial experiment with closed-set split, i.e., that the Mask2Former (both backbones) approach showed better overall performance but underperformed in the on heads. See Table 10 for detailed performance evaluation.

| Method       | mAP | head | turtle | flippers |
|--------------|-----|------|--------|----------|
| Mask R-CNN   | 0.827| 0.735| 0.907  | 0.840    |
| HTC          | 0.833| 0.740| 0.909  | 0.849    |
| Mask2Former  | 0.850| 0.708| 0.975  | 0.866    |

Table 10. Instance segmentation performance of selected backbone and head architectures over the SeaTurtleID. Open set split.

A.4. The importance of time-aware splitting

We further test and demonstrate the need for time-aware splits on other datasets that include timestamps using the Swin-B/p4w7 with the same setting as in the previous section. In Tab. 9, we show that in all cases, the results on the random split are undesirably inflated and much better than the ones of the time-aware split. We get further insight by considering all pairs of images of the same individuals with the same head orientation and see how their matching probability (proportion of correctly matched pairs) is affected by the time between them. Fig. 9 shows that the probability of correctly matching such image pairs decreases as the time between them increases. For instance, while this probability is 53.5% for images taken on the same day, it decreases to 2.5% for images taken more than one year apart.

Further insights: We further interpret Fig. 9 with the specific example of turtle “t298”, which was observed only on two days: 01/07/2016 and 12/07/2020. The random split has images from both dates in both reference and query sets, while the time-proportion split contains all images from 2016 in the reference set and all images from 2020 in the query set. While there were 26 matches for 2016-2016 images and 140 matches for 2020-2020 images, there were only 2 matches for 2016-2020 images. This further implies that there are many matches between the reference and query sets for the random split but almost no such matches for the time-proportion split. Therefore, the random split unrealistically simplifies the real-world re-identification problem.

B. Additional figures about the SeaTurtleID2022 dataset

Fig. 10 displays photographs of seven individuals (one individual per row) showing the variability of the unique facial scale patterns of loggerhead sea turtles. The scales on the left and right sides of the head are different in a given individual, making it impossible to match them without any intermediate images.

Fig. 11 shows further examples of different visual appearances of the same individual sea turtles over long periods of time due to different factors like camera capture conditions and animal aging. The shapes of the facial scales remained stable, but other features have changed over time, like coloration, pigmentation, shape, and scratches.

Fig. 12 shows sample images from the SeaTurtleID2022 dataset, highlighting the variety of photographs (poses, orientations, backgrounds, etc.).
Figure 10. Examples of 7 individuals (one individual per row) that show the variability of unique facial scale patterns of loggerhead sea turtles. From left to right: right lateral facial scales, left lateral facial scales, dorsal head scales.

Figure 11. Further examples of different visual appearances of the same individual sea turtles over long periods of time due to different factors like camera capture conditions and animal ageing. The shapes of the facial scales remained stable, but other features have changed over time, like colouration, pigmentation, shape, and scratches.
Figure 12. Examples of original photographs from the SeaTurtleID2022 dataset.