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. 1.

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. 1, we provide the performance of five architectures on two splits of the SeaTurtleID2022 dataset: time-aware and a random split. In Tab. 2, 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 pretrained model input size of 224 × 224.

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 1. Performance inflation (accuracy) with different backbones fine-tuned with softmax cross-entropy.

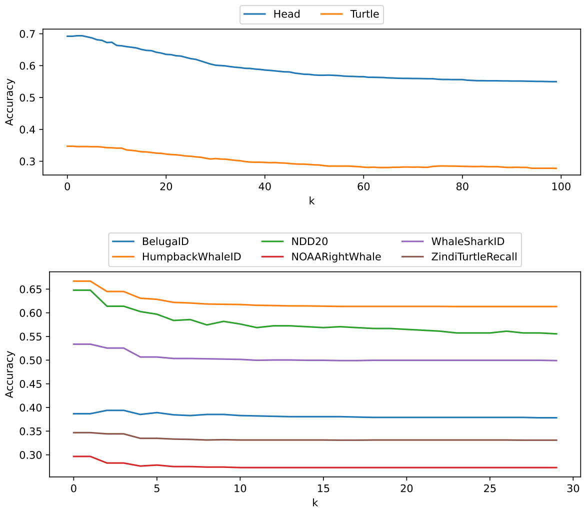

Figure 1. **Effect of k on performance**. We display the classification accuracy of k-NN classifier with ArcFace embeddings for various k values. Different body parts (e.g. head and full-body) performance on the SeaTurtleID dataset (top) and selected wildlife re-identification datasets (bottom).

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

Table 2. Performance inflation (accuracy) with different datasets.

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 3 for detailed performance evaluation.

	Method	mAP	*head*	*turtle*	*flippers*
ResNet-50	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
Swin-B	Mask R-CNN	0.833	0.743	0.913	0.844
	HTC	0.839	0.740	0.921	0.856
	Mask2Former	0.855	0.714	0.977	0.874

Table 3. 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. 2, 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. 2 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.

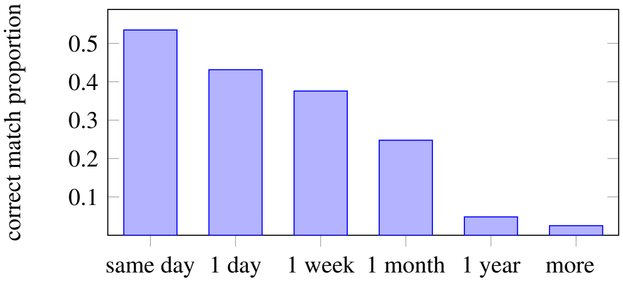

Figure 2. The probability of the correctly matched pairs of images of the same individuals with the same head orientation (left side) decreases as the time between the two images increases.

Further insights: We further interpret Fig. 2 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 unnaturally simplifies the real-world re-identification problem.

B. Additional figures about the SeaTurtleID2022 dataset

Fig. 3 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. 4 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. 5 shows sample images from the SeaTurtleID2022 dataset, highlighting the variety of photographs (poses, orientations, backgrounds, etc.).

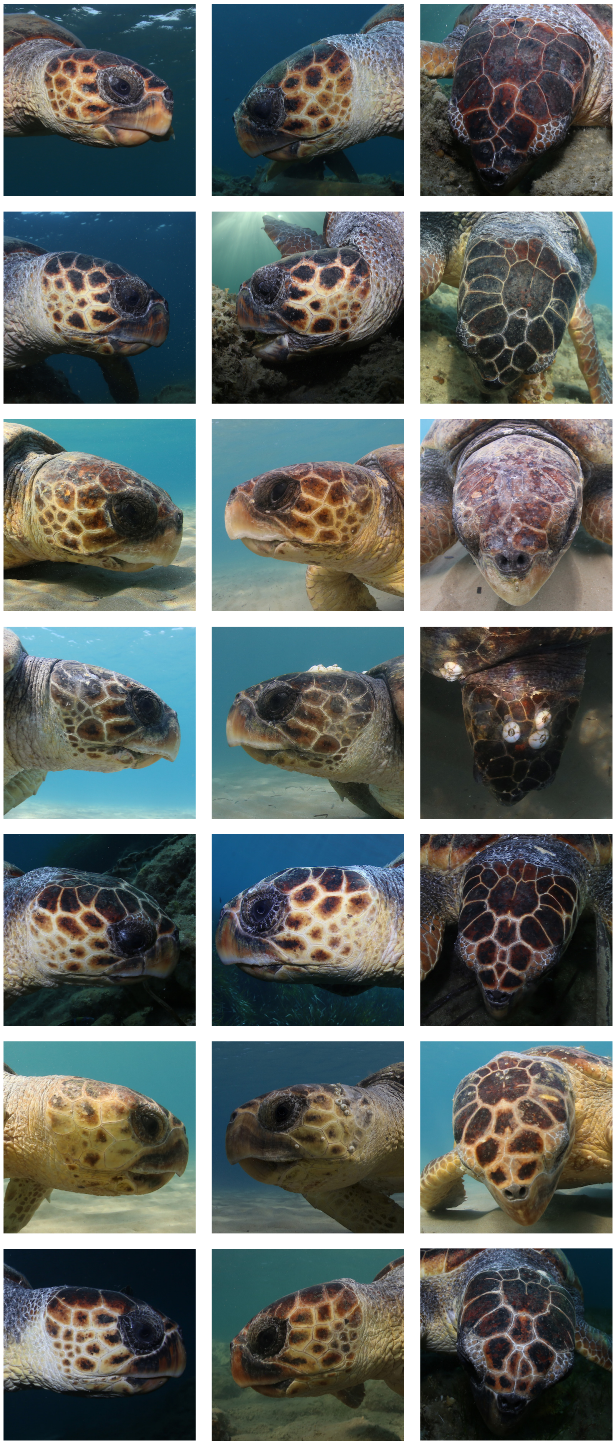

Figure 3. Examples of 7 individuals (one individual per row) that show the variability of unique facial scale patterns of loggerhead sea turtles. From left to right: right lateral facial scales, left lateral facial scales, dorsal head scales.

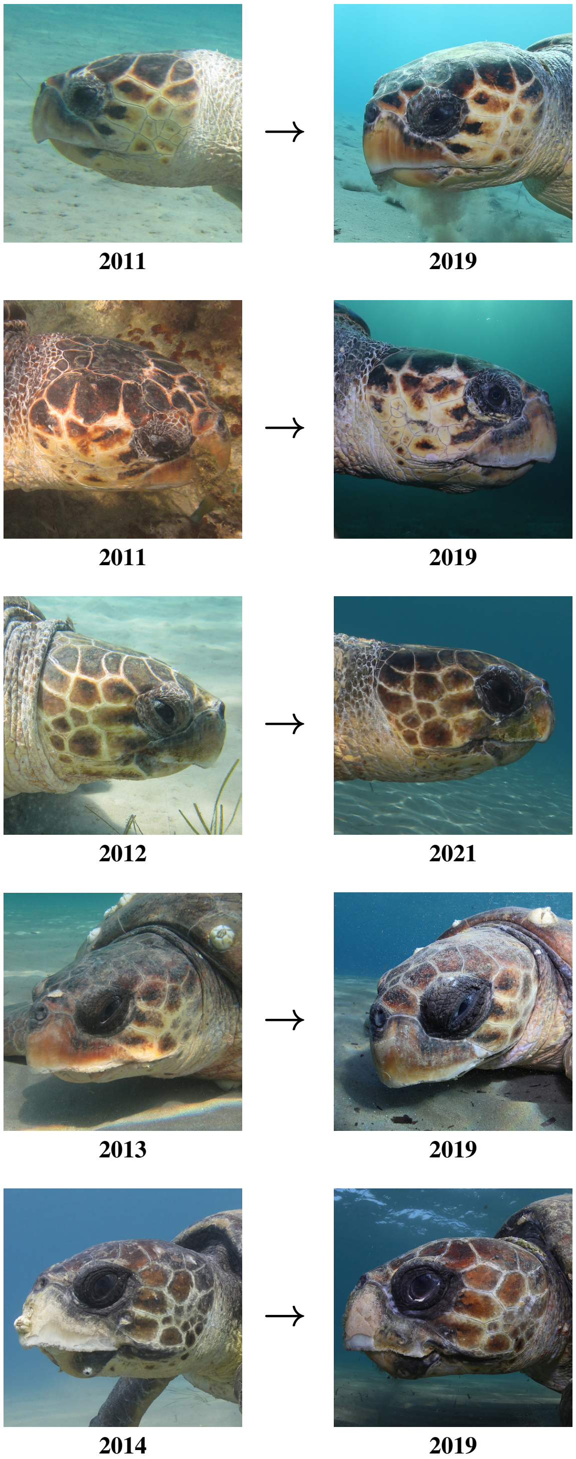

Figure 4. Further examples of different visual appearances of the same individual sea turtles over long periods of time due to different factors like camera capture conditions and animal ageing. The shapes of the facial scales remained stable, but other features have changed over time, like colouration, pigmentation, shape, and scratches.

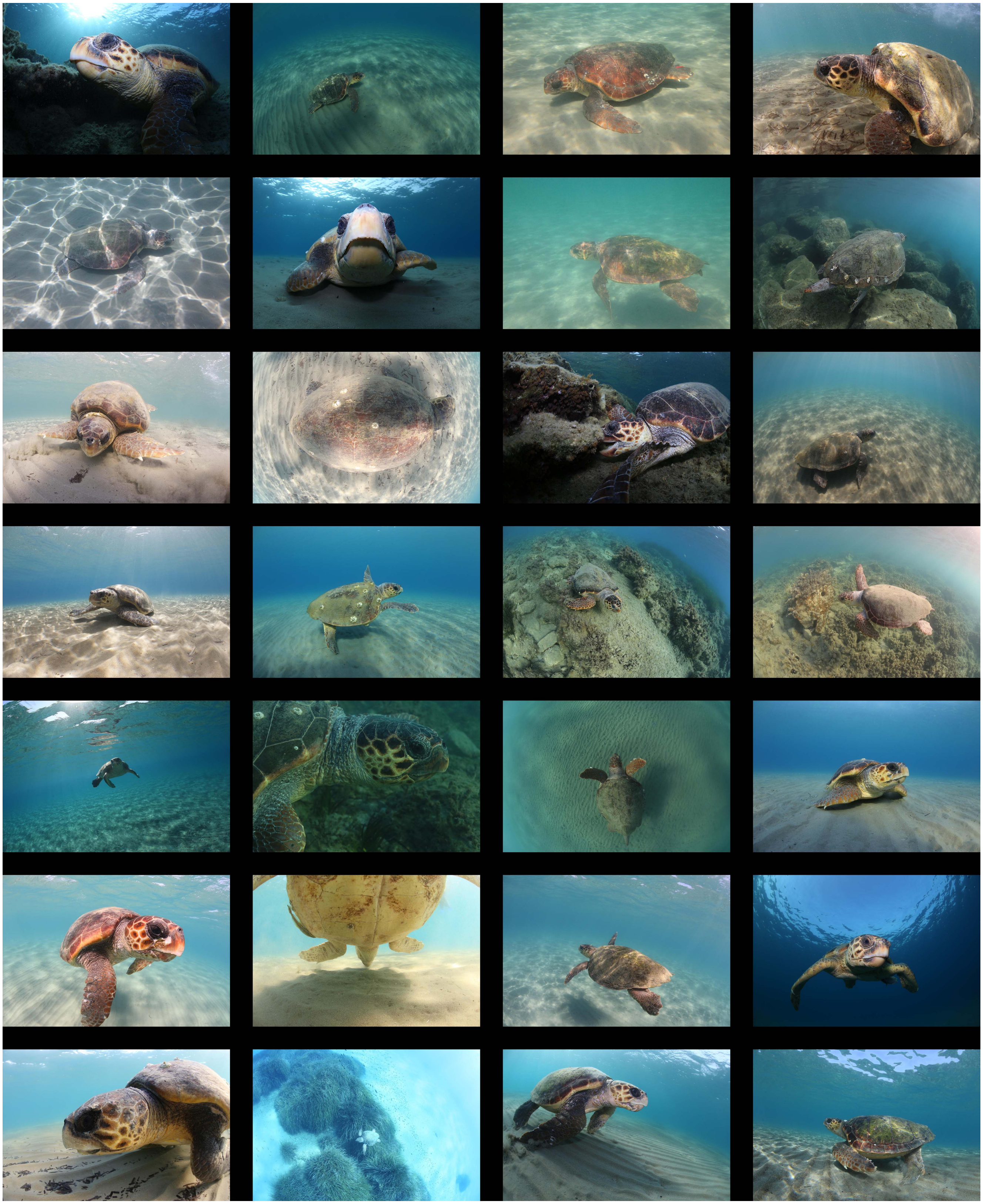

Figure 5. Examples of original photographs from the SeaTurtleID2022 dataset.