WildFusion: Individual Animal Identification with Calibrated Similarity Fusion

Vojtěch Cermak[1], Lukas Picek[2,3], Lukáš Adam[4], Lukáš Neumann[1], and Jiří Matas[1]

[1] Czech Technical University in Prague, FEL, CMP, Prague, Czechia
[2] University of West Bohemia, FAS, Department of Cybernetics, Pilsen, Czechia
[3] Inria, LIRMM, University of Montpellier, CNRS, Montpellier, France
[4] University of West Bohemia, FEE, RICE, Pilsen, Czechia
{cermavo3,matas}@fel.cvut.cz, lukaspicek@gmail.com, adamluk3@fel.zcu.cz

Abstract. We propose a new method – WildFusion – for individual identification of a broad range of animal species. The method fuses deep scores (e.g., MegaDescriptor or DINOv2) and local matching similarity (e.g., LoFTR and LightGlue) to identify individual animals. The global and local information fusion is facilitated by similarity score calibration. In a zero-shot setting, relying on local similarity score only, WildFusion achieved mean accuracy, measured on 17 datasets, of 76.2%. This is better than the state-of-the-art model, MegaDescriptor-L, whose training set included 15 of the 17 datasets. If a dataset-specific calibration is applied, mean accuracy increases by 2.3% percentage points. WildFusion, with both local and global similarity scores, outperforms the state-of-the-art significantly – mean accuracy reached 84.0%, an increase of 8.5 percentage points; the mean relative error drops by 35%. We make the code and pre-trained models publicly available[5], enabling immediate use in ecology and conservation.

1 Introduction

Identifying individual animals is essential in various domains of wildlife research. It help us understand the complexities of species dynamics [46,62], which is necessary for developing efficient conservation strategies. Besides, it can improve the accuracy of population density estimation, which is important in problems like disease monitoring and control [45], the role of the animal in the ecosystem [51], monitoring invasive species [11] and measuring the involvement of humans in the animal's habitat and ecological restoration [8]. Accurate identification requires domain knowledge and is extremely time-consuming due to the need for manual data processing. Therefore, considerable progress has been made in the development of methods for automating this process. Even though identifying animal individuals from images is challenging, machine learning and computer vision methods applied to species with unique patterns already enhance ecological research [67]. The automation of animal re-identification is typically based on (i) *deep learning*, (ii) *local feature matching*, or (iii) *species-specific methods*.

[5] https://github.com/WildlifeDatasets/wildlife-tools

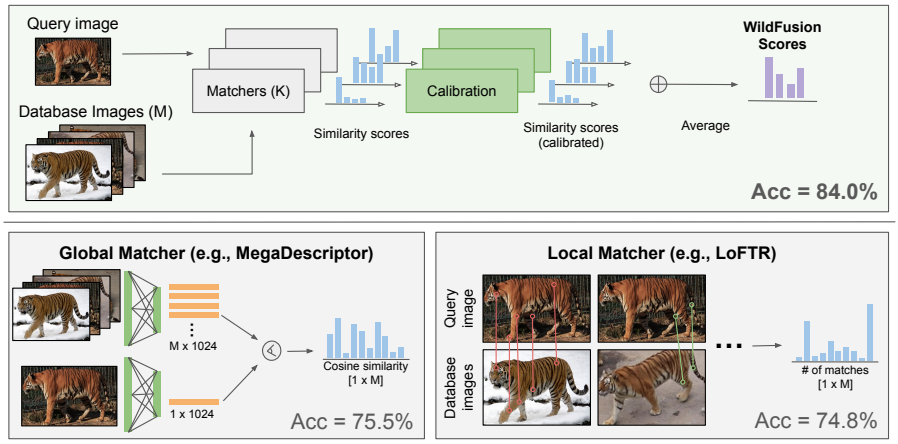

Fig. 1: Calibrated similarity fusion. Fusing local (in the $[0, \mathcal{R}]$ range) and global matching scores (e.g., cosine similarity) is not possible without calibration. By calibrating the outputs of any local and global matcher, we can easily fuse them and achieve better performance. In terms of accuracy and evaluated on 17 datasets, we increased the performance by 8.5% on average and reduced relative error by 35%.

The *deep learning*-based approaches [10,16,23,34,40,61] use either a standard classification-like approach or metric learning. Even though those approaches perform well, the models need relatively large annotated datasets, and fine-tuning requires considerable computational resources. On the other hand, methods based on *local descriptors* (e.g., SIFT [39], SuperPoint [18]) can be used without fine-tuning. [6, 22, 47, 50]. Indeed, the overall accuracy of matching local descriptors does not achieve a performance of deep learning methods [13], but those approaches are still very popular due to the existence of open-source tools (e.g., HotSpoter, WildMe) that are based on local descriptors. Additionally, matching requires a pairwise comparison between all query and database samples. As the identity database grows in size, the computational time quickly becomes unfeasible; therefore, local feature matching remains a viable option only for moderately sized datasets. The *species-specific* methods are usually tailored to suit species without any visual characteristics [5, 7, 28, 31, 65]. Existing methods focus, for example, on the shape of an elephant's ear, the facial characteristics of primates, or the fluke shape of whales. However, due to their idiosyncratic nature, these methods are difficult to transfer to other species.

In light of that, we propose WildFusion, a new state-of-the-art approach to zero-shot animal re-identification. It fuses calibrated deep similarity functions (i.e., MegaDescriptor and/or DINOv2 feature similarity) and local matching scores (i.e., number of matches from descriptors such as LoFTR and LightGlue) to select an identity from a database. For reference, see the illustration in Figure 1. With this straightforward approach, WildFusion significantly outperforms the current state-of-the-art without domain adaptation or fine-tuning.

The main contributions of this paper are:

- A new ensembling framework (WildFusion) that allows a combination of deep- and local-feature matching scores.
- A state-of-the-art performance on a set of animal identification problems, outperforming current methods by 8.5% on average; measured on 17 datasets.
- Comprehensive evaluation of selected state-of-the-art deep-learning and local feature-matching methods for image-matching and animal re-identification.
- Showing that WildFusion works without the need for fine-tuning and provides state-of-the-art performance out of the box, even in a zero-shot setting.

2 Related work

Local feature matching methods: Early work on animal re-identification used hand-crafted features such as cheetah's spots [31] or zebra's stripes [33]. Since these approaches suffer from poor performance and are non-transferable to other species, methods extracting local patterns such as SIFT [39] or ORB [52] were soon widely used. They extract descriptors from a database of images and match the descriptors from image pairs. Popular SW, e.g., WildID [9], HotSpotter [15] and I^3S are using such approach for years. Recently, the focus moved to local features extracted by deep networks such as ALIKED [71], DISK [60] or SuperPoint [18]. The classical matching of local descriptors could be simply replaced by deep methods such as LightGlue [36], Superglue [53], and LoFTR [57] that allow both extracting and matching of the local features. These matching methods return potential matches and their confidence scores. They require manual thresholding to determine which features are matched. In animal re-identification, deep local features are slowly coming into focus; for example, [48] used a combination of the SuperPoint features with the SuperGlue matching.

Deep embedding methods: The applications of deep methods in animal re-identification are relatively new [12]. The simplest use case consists of extracting embeddings from a neural network and feeding them to an SVM classifier [14, 32, 41]. This approach has low computational demands, but the network cannot be fine-tuned. Another simple approach involves fine-tuning a pre-trained neural network [23, 54]. These approaches usually require a fixed number of classes (individuals), which is not realistic. For this reason, metric learning methods (e.g., ArcFace [13], Siamese networks [30], and Triplet loss [19, 40]) became popular. Instead of classifying images into a pre-determined set of classes, they measure differences between images and are, therefore, able to generalize into new individuals. Another approach is to use publicly available large-scale, foundational models pre-trained on large datasets such as BioCLIP [56], DINOv2 [44], and MegaDescriptor [13]. Since these models are primarily designed for general computer vision tasks, they are not adapted for the nuances of wildlife re-identification, which heavily relies on fine-grained patterns. This was addressed by MegaDescriptor [13], the Swin-based foundational model for animal re-identification that was trained on over 30 datasets (collected using WildlifeDatasets) using ArcFace loss [17].

Species-specific methods are usually tailored to a particular species or closely related species and involve pre-processing steps such as extracting patches from regions of interest or accurately aligning images. Besides, they are not transferable to other species. Examples of these methods are Amphident [21] which find matching pixels within newt patterns or CurvRank v2 [42] and finFindR [66], which match the fin curvatures to identify mantas, dolphins, or whales.

3 Methodology

In this section, we describe the similarity scores based on deep embeddings and local feature matching and introduce the proposed WildFusion, a method for finding an image from a database of images $x_1, \ldots, x_D$ closest to query image x_q.

***Note**: For wildlife re-identification, we employ a standard setting inspired by practical applications in animal ecology, widely used in automated animal re-id studies [4, 13]. This setting corresponds to the image retrieval problem, where the goal is to find the most visually similar images (whose identity is used as prediction) in the database for a given query image based on a similarity metric.*

3.1 Global similarity score

Given an image x, we use a neural network $f(x)$ to extract a fixed-length embedding. The network $f(x)$ is a complex function that maps images into embedding space where the representations of images depicting the same animal are closer together, while those of different individual animals are distinctively separated. Common architectures of neural networks include convolutional [38, 68] or transformer-based [20, 37] architectures and are often trained with metric learning, e.g., ArcFace [17] and Triplet loss [55], to promote separability in the embedding space. The similarity between images is calculated as the similarity between their representation in the embedding space. Formally, we define the *global similarity* between two images x_0 and x_1 as the cosine similarity between their corresponding deep embeddings extracted by neural network f:

$$s_G(x_0, x_1) = \frac{f(x_0) \cdot f(x_1)}{\|f(x_0)\|\|f(x_1)|}. \tag{1}$$

3.2 Matching based similarity score

We derive a similarity metric based on local feature matching as the number of found significant matches. The feature matching methods return a list of potential matches and their confidence score. We declare a match to be significant if its confidence score is above some threshold μ. Formally, given two images x_0 and x_1 with the number of matches $M(x_0, x_1)$ with confidence scores $c_m(x_0, x_1)$, we define the *local similarity* metric as

$$s_L(x_0, x_1) = \sum_{m=1}^{M(x_0,x_1)} I\big(c_m(x_0, x_1) > \mu\big), \tag{2}$$

where I is the counting (0/1) function. This approach requires the tuning of the thresholding hyperparameter μ, where large values of μ allow for a small number of high-quality matches, while low values of μ result in a larger amount of matches with potentially lower quality. As we empirically show in Section 6.1, all considered feature matching methods are robust to μ selections with $\mu = 0.5$ being reasonable choice in most scenarios.

3.3 Score calibration

Calibration refers to rescaling model outputs so that they could be interpreted probabilistically. In our case, it is required to normalize the outputs of multiple models to the common range $[0, 1]$. Predictions of well-calibrated model reflect confidence in the given class predictions [27].

We apply calibration to the predicted similarity scores. The similarity scores are used for comparison and ranking in image retrieval, with the magnitude of the scores having no direct interpretation. We use calibration to ensure that the predicted similarity score corresponds to the probability that the images in a pair have the same identity. We construct calibrated scores from either global or matching-based scores using a calibration function $f_{\text{cal}} : \mathbb{R} \to [0, 1]$.

$$\hat{s}(x_0, x_1) = f_{cal}(s(x_0, x_1)). \tag{3}$$

A common approach for constructing f_{cal} is Platt scaling [49], which involves fitting a single-variable logistic regression with uncalibrated scores as inputs. Another widely used method is isotonic regression [70], a variant of binning regression with a monotonicity constraint. Given uncalibrated scores, it learns a non-decreasing piecewise constant function. However, for our application in ranking and image retrieval, we require a strictly increasing function to handle ties in scores. To achieve this, we first apply the isotonic regression and second interpolate the bin centers using a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) [25], which performs cubic interpolation while preserving monotonicity. This procedure results in the required strictly increasing calibration function.

3.4 WildFusion – Calibrated score ensembling

To construct an ensemble, we consider K models with similarity scores $s_i(x_0, x_1)$ for a pair of images (x_0, x_1). The calibrated scores $\hat{s}_i(x_0, x_1)$ are interpreted as estimates of same true probability $P(\text{id}(x_0) = \text{id}(x_1) \mid x_0, x_1)$. To denoise the prediction, we assume the probabilities to be independent observations with additive noise. The WildFusion score is thus a weighted average of n calibrated scores:

$$s_F(x_0, x_1) = \sum_{i=1}^{K} w_i \hat{s}_i(x_0, x_1), \tag{4}$$

where the weights should reflect the variance of the additive noise in the score. If we assume that all scores have similar variance, equal weights $w_i = \frac{1}{n}$ are selected, and the weighted average reduces to the simple average.

Fig. 2: Distinct animal features for re-identification. Based on the natural visual appearance, the most distinguishable features for animals are spots, stripes, facial landmarks, and the shape of body parts (e.g., ears for elephants and fin for whales).

4 Datasets

We evaluate WildFusion on 17 datasets[6] that include diverse species of animals. The datasets were acquired with the help of the recently developed library Wildlife Datasets [13], which allows easy access to the datasets and provides unified dataset splits. We selected subsets of the datasets that are saturated in performance or have a large number of images. Sample images from selected datasets are shown in Figure 2. For basic statistics of the datasets, see Table 1.

To construct the appropriate training (*database*) and test (*query*) datasets, we followed the methodology proposed in [13]. However, while analyzing this procedure, we discovered inconsistencies caused by the incorrect loading of images for ATRW and NDD20 datasets due to multiple identities in one image. For these datasets, we fixed the loading by applying the appropriate bounding box or segmentation mask. Therefore, as loaded images are not exactly the same, the achieved results for these two datasets are higher than in the original work.

Table 1: Characteristics of selected datasets. †Used in zero-shot scenario.

	category	# of images	# of individuals
ATRW [34]	tigers	5,415	182
CowDataset† [26]	cows	1485	13
GiraffeZebraID [47]	giraffes, zebras	6,925	2,056
Giraffes [40]	giraffes	1,393	178
HyenaID2022 [59]	hyenas	3,129	256
LeopardID2022 [59]	leopards	6,806	430
NyalaData [19]	nyalas	1,942	237
SealID [43]	seals	2,080	57
SeaStarReID2023† [63]	starfish	2187	95
SeaTurtleID [4]	sea turtles	7,774	400
WhaleSharkID [29]	whale sharks	7,693	543
ZindiTurtleRecall [3]	sea turtles	12,803	2,265
BelugaID [2]	belugas	5,902	788
CTai [24]	chimpanzees	4,662	71
IPanda50 [64]	pandas	6,874	50
NDD20 [58]	dolphins	2,657	82
NOAARightWhale [1]	whales	4,544	447

[6] For zero-shot, we use only two that were not used in the MegaDescriptor training.

5 Experiments

Evaluation protocol. We consider the same closed-set splits as in [13], meaning that all individual animals are both in the database (training) and query (test) sets. We approach the problem as image retrieval: for each image in the query set, we find an image in the database and make the query prediction have the same identity as the image from the database. Performance in all experiments is measured as top-1 accuracy.

WildFusion relies on finding hyper-parameters and fitting calibration models. The standard approach involves splitting the training set into development and validation parts, which are used for the selection of the best hyper-parameters and fitting calibration models. However, this approach is not applicable in our case because the MegaDescriptor was trained on the whole training set and cannot be used for validation. We addressed this issue by splitting the original test set into a validation set and a new, smaller test set using a 0.5 ratio. We estimated both μ and the calibration function on the validation set and utilized them for the final prediction on the test set. Due to this change in test set, our results are not directly comparable to the results reported by [13].

Technical details. To construct the global scores, we use embeddings extracted by MegaDescriptor [13] and DINOv2 [44]. For local matching scores, we use LightGlue [36] feature matching with local descriptors ALIKED [71], DISK [60], and SuperPoint [18]. We use at most 512 keypoints and their appropriate descriptors, extracted from images resized to 512×512. For matching with LoFTR [57], we use the outdoor variant trained on the MegaDepth [35] dataset. On input, we use image pairs with both images resized to 512×512. A total of four local feature matching methods were considered to construct matching-based scores. All these methods were taken off-the-shelf, and none were fine-tuned or retrained. We perform the experiments on the datasets described in Section 4. In the baseline WildFusion, we search for optimal hyperparameter μ from Equation 2 separately for each dataset. The calibrated scores are given equal weights w_i. The summary of settings is in Table 2.

Table 2: WilfFusion settings overview. We test a variety of state-of-the-art local and global methods for animal re-identification and image retrieval. The calibration is done using Logistic or Isotonic regression.

Components:	Local matching methods: – LoFTR – LightGlue + SuperPoint – LightGlue + Disk – LightGlue + Aliked	Global similarity methods: – MegaDescriptor-L-384 – DINOv2-512
Calibration:	Isotonic regression with PCHIP interpolation Logistic regression	
Fusion:	Average with equal weights w_i	

5.1 Baseline Performance

WildFusion clearly outperforms MegaDescriptor in most scenarios. When using all available scores, WildFusion shows superior performance on 16 out of 17 datasets, with only one dataset, ZindiTurtleRecall, showing a slight decrease in accuracy. The average accuracy improvement is substantial, with WildFusion (all) achieving 84.0% compared to MegaDescriptor's 75.5%, representing a notable average gain of 8.5 percentage points. The most significant improvements are seen in datasets like NDD20, WhaleSharkID, SeaStarReID2023, and SealID, where WildFusion shows accuracy gains of over 14 percentage points.

Interestingly, even when using only local matching scores, WildFusion maintains competitive performance. It outperforms MegaDescriptor on 11 out of 17 datasets with average accuracy (78.5%), which is better than MegaDescriptor by 3.0 percentage points. This suggests that the local matching scores are quite powerful on their own, without any need for fine-tuning on animal datasets. More details about the results, including per dataset performance, are in Table 3. Besides, we provide a qualitative evaluation in Figure 3

Table 3: WildFusion's performance in comparison with MegaDescriptor. On average, WildFusion, outperforms MegaDescriptor, even with just *local* descriptors. WildFusion with *all* local and deep descriptors ranks the best on all but two datasets.

	MegaDescriptor Large-384	(*all*) WildFusion	Δ	(*local*) WildFusion	Δ
ZindiTurtleRecall	74.24	71.90	-2.34	45.62	-28.62
CTai	91.86	92.08	+0.21	81.80	-10.06
ATRW	97.96	98.51	+0.56	98.33	+0.37
CowDataset	98.66	100.00	+1.34	100.00	+1.34
SeaTurtleIDHeads	91.18	95.00	+3.82	93.82	+2.63
IPanda50	85.76	89.68	+3.92	81.40	-4.36
NyalaData	41.59	46.26	+4.67	25.23	-16.36
BelugaID	67.61	72.46	+4.85	63.07	-4.54
NOAARightWhale	43.25	49.25	+6.00	42.18	-1.07
Giraffes	91.04	99.25	+8.21	98.51	+7.46
HyenaID2022	78.41	90.48	+12.06	88.25	+9.84
GiraffeZebraID	82.98	95.74	+12.77	94.81	+11.84
LeopardID2022	77.82	90.93	+13.11	89.40	+11.58
SealID	78.47	92.82	+14.35	90.91	+12.44
SeaStarReID2023	82.24	99.53	+17.29	100.00	+17.76
WhaleSharkID	62.04	80.33	+18.28	77.68	+15.64
NDD20	38.35	63.53	+25.19	63.16	+24.81
Average	*75.50*	*83.99*	*+8.49*	*78.48*	*+2.98*

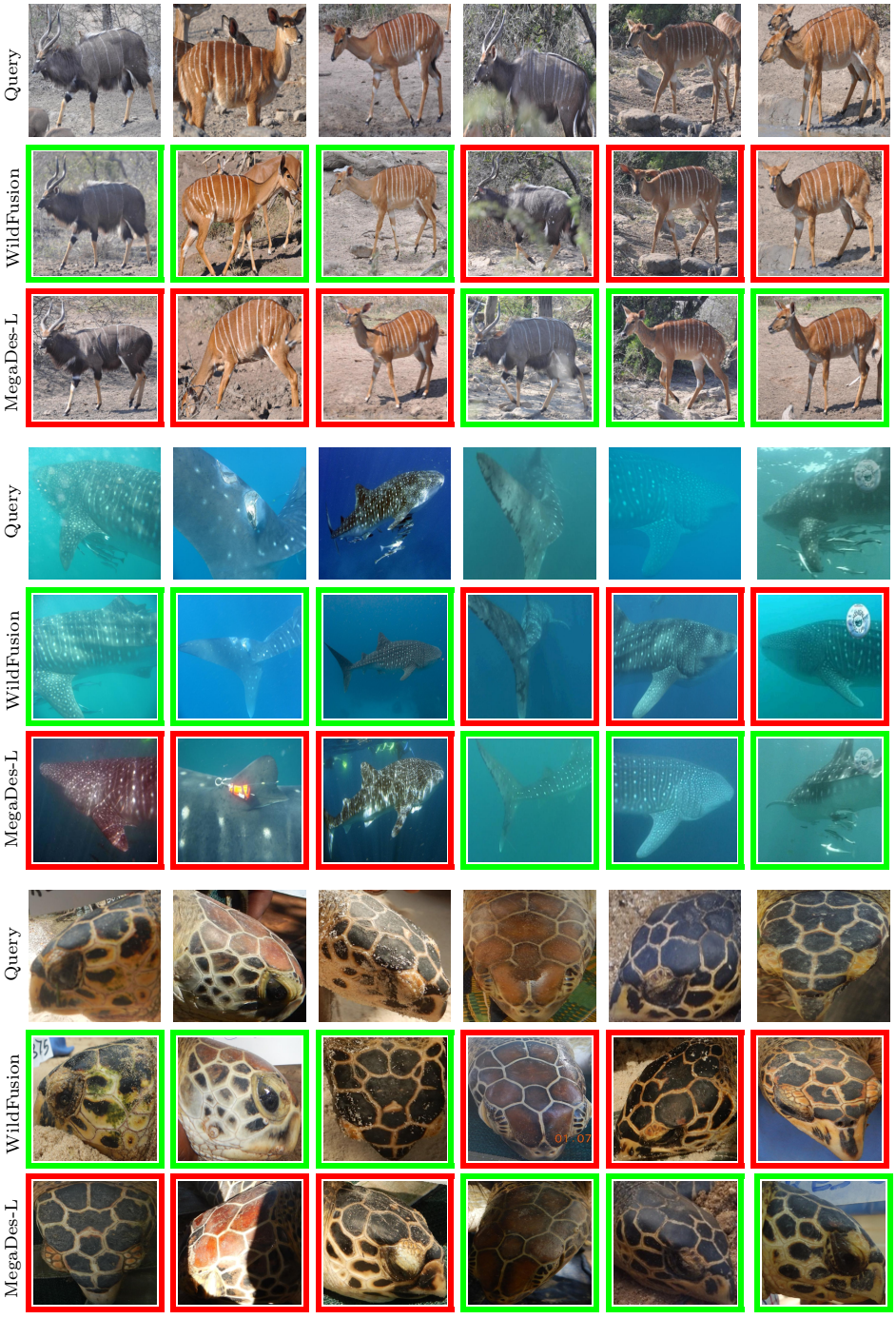

Fig. 3: Qualitative performance. Selected examples where WildFusion changed the decision of the MegaDescriptor-L on NyalaData, WhaleSharkID, and ZindiTurtle; three correct and false samples. We suspect that some wrong matches are mislabeled data.

6 Ablation Studies

This section presents a set of ablation studies to empirically validate the design choices behind the WildFusion.

6.1 Effect of local matching score threshold

The hyperparameter μ controls the trade-off between low-quality matches and fewer high-quality matches. When μ is low, the score is influenced by many low-quality matches, often presented in the background. When μ is very high, it filters out most of the matches, leading to a loss of information and resulting in a zero score for nearly all pairs.

Comparing performance scores with constant μ and μ selected based on the validation set suggests that local methods are robust to μ selection, and selecting any μ values between [0.4, 0.6] is a good choice. Interestingly, local methods perform better with μ=0.45 fixed for all datasets than searching for optimal μ on the validation set. When local matching scores are combined with global scores, the range of suitable μ values is wider and extends from 0.4 to 0.8. This shows that adding global scores to the ensemble reduces the downside of having zeros in the score for large μ values (see Figure 4).

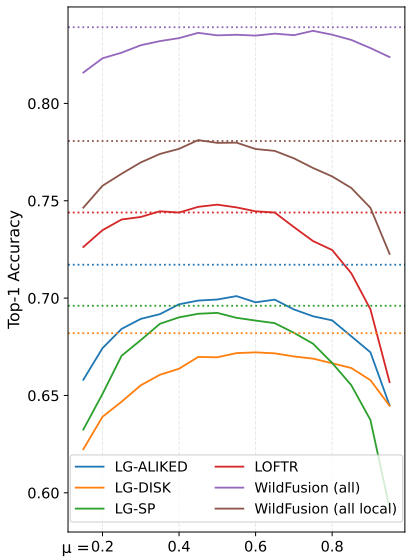

Fig. 4: Effect of μ on performance. Full lines represents constant μ, and dotted lines optimal μ found on validations set for each dataset. Fixing $\mu = 0.5$ provides comparable results to the best μ based on validation set.

6.2 Effect of score selection

WildFusion's versatility allows it to fuse any score. As mentioned, using WildFusion only with all local matching scores outperforms the MegaDescriptor-L global score. When we included the global score from the general-purpose feature extractor DinoV2, performance improved only marginally, highlighting the importance of fine-tuning the deep embedding model.

Using the MegaDescriptor-L global score with at least one local matching score significantly outperforms using MegaDescriptor-L alone. Combining it with LG-ALIKED achieves the highest accuracy of 83.0%, followed by LoFTR at 81.4%. LG-SuperPoint and LG-DISK also show comparable performance with accuracies of 80.6% and 81.1%, respectively. Combining the global score with all local matching scores further improves performance, suggesting that the local matching scores are mostly uncorrelated and perform well in the ensemble. More details can be found in Table 4.

Table 4: Ablation on local and global score fusion. We report WildFusion's performance using various local and global methods. Combining local methods with fine-tuned global scores of MegaDetector-L achieves the best results.

Global \ Local	*None*	LG-DISK	LG-SP	LG-ALIKED	LoFTR	all
None	-	68.9	70.1	72.2	74.8	78.5
DINOv2	47.5	70.4	71.6	73.7	74.8	78.8
MegaDescriptor-L	75.5	81.1	80.6	83.0	81.4	84.0

6.3 Effect of calibration

Using isotonic regression for calibration yields marginally better results on average compared to logistic regression (83.9% accuracy). However, there is a discrepancy in performance between the datasets. For example, using logistic regression was better on NDD20 (+3.0%) and ZindiTurtleRecall (+ 2.7%), but it significantly underperformed on NyalaData (-6.5%) compared to the isotonic regression. This suggests that the poor performance of WildFusion on ZindiTurtleRecall can be related to incorrect calibration.

How much data do we need for calibration? To test this, we create variously sized subsets of labeled images from database and validation sets, such that at least 2 positive and 2 negative pairs can be created. Pairs created from this subset are used for calibration and finding μ. We perform additional experiments with μ fixed to 0.5 to isolate the effect of low data calibration from finding μ. As visualized in Figure 5, isotonic regression performs better than logistic regression in low data scenarios, both with optimized and fixes μ. Calibration with fixed μ is significantly better for a smaller number of samples but yields marginally worse results when a lot of labeled data is available. In general, calibration is very data efficient. For example, fixing μ to 0.5 and calibrating each dataset using only 10 labeled images still gives a reasonable 79.5% accuracy. Adding more data to the calibration further increases the performance of up to 200 samples, where additional data only gives marginal improvements. Our results suggest that WildFusion is viable even with very few labeled samples.

6.4 Constraining number of comparisons

Given a database with M samples and a query with N samples, methods based on local features often need to perform pairwise comparisons, needing $M \times N$ comparisons. Since many modern matching algorithms are based on neural networks $M \times N$, neural network inferences are required. With the increasing database size, the computational time quickly becomes unfeasible; therefore, the calculation of all local scores remains a viable option only for moderately sized datasets.

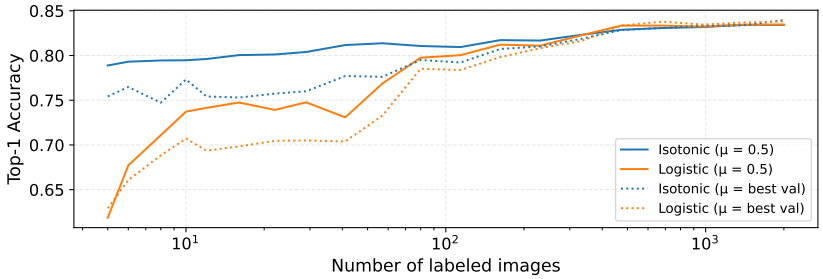

Fig. 5: Ablation on optimal number of images for model calibration. Isotonic regression calibration with fixed $\mu = 0.5$ for all datasets outperforms other approaches in low data scenarios.

Shortlist strategy. We consider a scenario where we have two types of scores, one cheap to calculate, such as global score s_G from MegaDescriptor or DinoV2, and one expensive, such as WildFusion with local matching scores s_W. We follow the shortlist strategy [69] and use cheap global scores to filter candidate samples. The expensive scores are calculated for a restricted size shortlist to validate and re-rank the top matches. The running time is controlled by a computational budget B in terms of the number of expensive score evaluations per query image.

Results. Using the shortlist strategy, we are efficiently able to utilize the WildFusion scores s_W, which are costly to calculate. On average, budget $B = 300$ is enough to reach accuracy comparable to calculating all scores. For example, on SeaTurtleIdHeads, WildFusion needs only about 200 s_W calculations to reach its peak performance. With a database size of 6063, this results in more than a 30-fold increase in inference speed. Interestingly, performance at $B = 10^3$ is slightly better than using all comparisons. It indicates that local matching scores in WildFusion are prone to some degree of false positive matches when applied to all images in the database. A more detailed visualization of the speed-up is in Figure 6.

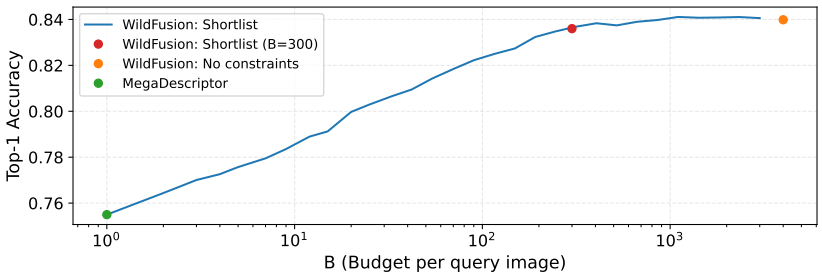

Fig. 6: Rate of performance improvement with increasing budget. The shortlist strategy allows adding more computational resources to improve performance, up to a budget of $B = 200$.

7 Zero shot performance

Encouraged by the fact that calibration works well even with a low number of data points to achieve reasonable performance, we conducted an experiment in a zero-shot setting, meaning no data is needed prior to inference. We split datasets into disjoint subsets. For each score, we trained a single calibration model on one subset and evaluated it on a different subset, with $\mu = 0.5$ fixed for all local matching scores. This differs from the default setting, where subsets of the same dataset were used for calibration and evaluation.

For the zero-shot experiment, we evaluated WildFusion using only local matching scores without incorporating MegaDescriptor's global scores, as the latter had already been trained on the data. Our zero-shot WildFusion approach achieved an average accuracy of 76.2%, which is 2.3 percentage points lower than the accuracy obtained with dataset-specific calibration. Notably, this performance is also 0.7 percentage points higher than the state-of-the-art fine-tuned model MegaDescriptor-L-384, demonstrating the effectiveness of our method in a zero-shot setting without fine-tuning or dataset-specific calibration. For more detail about performance, see Table 5.

For both novel datasets not included in the MegaDescriptor training set, WildFusion with local scores achieved a perfect 100% accuracy. In contrast, for the CowsDataset, DINOv2 reached an accuracy of 96.0%, while MegaDescriptor achieved 98.7%. Similarly, for the SeaStarReID2023 dataset, DINOv2 obtained an accuracy of 82.2%, whereas MegaDescriptor reached 88.8%.

Table 5: WildFusion performance in zero-shot setting. No data from the evaluated dataset was used prior to test time (except Wahltinez et al. [63], which used standard classification setting).

	MegaDescriptor-L	DINOv2	(*local*) WildFusion	[63] Wahltinez et al.
CowsDataset	98.7	88.8	100.0	–
SeaStarReID2023	82.2	96.0	100.0	99.9
17 datasets	–	47.5	76.2	–

8 Conclusion

In this paper, we presented WildFusion, a novel approach to individual animal identification that leverages a calibrated similarity fusion of deep and local matching scores. By combining deep features extracted from MegaDescriptor or DINOv2 with local matching descriptors (e.g., LoFTR and SuperPoint), WildFusion achieves state-of-the-art performance across a wide range of datasets. Our method is easy to use in real applications as it does not require training and is usable out of the box with any pre-trained deep embedding models and local feature-matching methods. Besides, the code was made public.

Even though the best results were obtained with dataset-specific calibration, we have empirically shown that using WildFusion of only local similarity score and with generic calibration still gives good performance, with mean accuracy dropping only by 2.3% and still reduced the relative error of MegaDescriptor by 44 percentage points. WildFusion's flexibility was also further proven by its strong performance in zero-shot settings tested on species "never seen before."

The scalability and generalization potential of WildFusion makes it suitable for application across different species and environments, contributing significantly to the field of animal re-identification.

Limitations: WildFusion leverages off-the-shelf local matching methods like LoFTR and LightGlue, which were originally trained on datasets featuring static objects. Therefore it is not optimized for matching animals, where the same animal can be observed in various poses, lighting conditions and with occlusions. Therefore, our work can be extended by adapting or retraining these local feature-matching models specifically for animal identification tasks, potentially improving the accuracy and robustness of the WildFusion approach. The same applies to deep descriptors such as MegaDescriptor and DINOv2, which, if not trained on that species, will most likely underperform.

WildFusion is best suited for offline analysis of existing databases. While we introduced a method to address scalability, it remains insufficient for real-time identification. Future research could explore the development of more efficient algorithms to enable real-time processing and online identification.

Acknowledgements

The authors were supported 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.

References

1. Right whale recognition (2015), https://www.kaggle.com/c/noaa-right-whale-recognition
2. Beluga ID 2022 (2022), https://lila.science/datasets/beluga-id-2022
3. Turtle recall: Conservation challenge (2022), https://zindi.africa/competitions/turtle-recall-conservation-challenge
4. Adam, L., Čermák, V., Papafitsoros, K., Picek, L.: Seaturtleid2022: A long-span dataset for reliable sea turtle re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7146–7156 (2024)
5. Anderson, C.J., Da Vitoria Lobo, N., Roth, J.D., Waterman, J.M.: Computer-aided photo-identification system with an application to polar bears based on whisker spot patterns. Journal of Mammalogy **91**(6), 1350–1359 (2010), https://academic.oup.com/jmammal/article/91/6/1350/888329

6. Andrew, W., Hannuna, S., Campbell, N., Burghardt, T.: 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). pp. 484–488. IEEE (2016), https://ieeexplore.ieee.org/abstract/document/7532404
7. Bedetti, A., Greyling, C., Paul, B., Blondeau, J., Clark, A., Malin, H., Horne, J., Makukule, R., Wilmot, J., Eggeling, T., et al.: System for elephant ear-pattern knowledge (seek) to identify individual african elephants. Pachyderm **61**, 63–77 (2020), https://pachydermjournal.org/index.php/pachyderm/article/view/65
8. Blount, J.D., Chynoweth, M.W., Green, A.M., Şekercioğlu, Ç.H.: Covid-19 highlights the importance of camera traps for wildlife conservation research and management. Biological Conservation **256**, 108984 (2021)
9. Bolger, D.T., Morrison, T.A., Vance, B., Lee, D., Farid, H.: A computer-assisted system for photographic mark–recapture analysis. Methods in Ecology and Evolution **3**(5), 813–822 (2012)
10. Bruslund Haurum, J., Karpova, A., Pedersen, M., Hein Bengtson, S., Moeslund, T.B.: Re-identification of zebrafish using metric learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. pp. 1–11 (2020), https://ieeexplore.ieee.org/document/9096922
11. Caravaggi, A., Zaccaroni, M., Riga, F., Schai-Braun, S.C., Dick, J.T., Montgomery, W.I., Reid, N.: An invasive-native mammalian species replacement process captured by camera trap survey random encounter models. Remote Sensing in Ecology and Conservation **2**(1), 45–58 (2016)
12. Carter, S.J., Bell, I.P., Miller, J.J., Gash, P.P.: Automated marine turtle photograph identification using artificial neural networks, with application to green turtles. Journal of experimental marine biology and ecology **452**, 105–110 (2014)
13. Čermák, V., Picek, L., Adam, L., Papafitsoros, K.: Wildlifedatasets: An open-source toolkit for animal re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5953–5963 (2024)
14. Clapham, M., Miller, E., Nguyen, M., Darimont, C.T.: Automated facial recognition for wildlife that lack unique markings: A deep learning approach for brown bears. Ecology and evolution **10**(23), 12883–12892 (2020)
15. Crall, J.P., Stewart, C.V., Berger-Wolf, T.Y., Rubenstein, D.I., Sundaresan, S.R.: Hotspotter—patterned species instance recognition. In: 2013 IEEE workshop on applications of computer vision (WACV). pp. 230–237. IEEE (2013)
16. Deb, D., Wiper, S., Gong, S., Shi, Y., Tymoszek, C., Fletcher, A., Jain, A.K.: Face recognition: Primates in the wild. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–10. IEEE (2018), https://ieeexplore.ieee.org/abstract/document/8698538/
17. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
18. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)
19. Dlamini, N., van Zyl, T.L.: Automated identification of individuals in wildlife population using siamese neural networks. In: 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI). pp. 224–228. IEEE (2020)

20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
21. Drechsler, A., Helling, T., Steinfartz, S.: Genetic fingerprinting proves cross-correlated automatic photo-identification of individuals as highly efficient in large capture–mark–recapture studies. Ecology and Evolution **5**(1), 141–151 (2015), https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.1340
22. Dunbar, S.G., Anger, E.C., Parham, J.R., Kingen, C., Wright, M.K., Hayes, C.T., Safi, S., Holmberg, J., Salinas, L., Baumbach, D.S.: Hotspotter: Using a computer-driven photo-id application to identify sea turtles. Journal of Experimental Marine Biology and Ecology **535**, 151490 (2021), https://www.sciencedirect.com/science/article/pii/S0022098120301738
23. Ferreira, A.C., Silva, L.R., Renna, F., Brandl, H.B., Renoult, J.P., Farine, D.R., Covas, R., Doutrelant, C.: Deep learning-based methods for individual recognition in small birds. Methods in Ecology and Evolution **11**(9), 1072–1085 (2020)
24. Freytag, A., Rodner, E., Simon, M., Loos, A., Kühl, H.S., Denzler, J.: Chimpanzee faces in the wild: Log-Euclidean CNNs for predicting identities and attributes of primates. In: German Conference on Pattern Recognition. pp. 51–63. Springer (2016)
25. Fritsch, F.N., Carlson, R.E.: Monotone piecewise cubic interpolation. SIAM Journal on Numerical Analysis **17**(2), 238–246 (1980)
26. Fu, L., He, G.: Cow dataset (2021), https://doi.org/10.6084/m9.figshare.16879780.v1
27. Gao, J., Burghardt, T., Andrew, W., Dowsey, A.W., Campbell, N.W.: Towards self-supervision for video identification of individual holstein-friesian cattle: The Cows2021 dataset. arXiv preprint arXiv:2105.01938 (2021), https://arxiv.org/abs/2105.01938
28. Gilman, A., Hupman, K., Stockin, K.A., Pawley, M.D.: Computer-assisted recognition of dolphin individuals using dorsal fin pigmentations. In: 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–6. IEEE (2016), https://ieeexplore.ieee.org/abstract/document/7804460
29. Holmberg, J., Norman, B., Arzoumanian, Z.: Estimating population size, structure, and residency time for whale sharks Rhincodon typus through collaborative photo-identification. Endangered Species Research **7**(1), 39–53 (2009), https://www.int-res.com/abstracts/esr/v7/n1/p39-53/
30. Kabuga, E.: Using neural networks to identify individual animals from photographs. Master's thesis, Faculty of Science (2019)
31. Kelly, M.J.: Computer-aided photograph matching in studies using individual identification: an example from Serengeti cheetahs. Journal of Mammalogy **82**(2), 440–449 (2001)
32. Korschens, M., Denzler, J.: ELPephants: A fine-grained dataset for elephant re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 263–270 (2019)
33. Lahiri, M., Tantipathananandh, C., Warungu, R., Rubenstein, D.I., Berger-Wolf, T.Y.: Biometric animal databases from field photographs: identification of individual zebra in the wild. In: Proceedings of the 1st ACM international conference on multimedia retrieval. pp. 1–8 (2011)
34. Li, S., Li, J., Tang, H., Qian, R., Lin, W.: ATRW: A Benchmark for Amur Tiger Re-identification in the Wild. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 2590–2598. Association for Computing Machinery (2020)

35. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018)
36. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. arXiv preprint arXiv:2306.13643 (2023)
37. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
38. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
39. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)
40. Miele, V., Dussert, G., Spataro, B., Chamaillé-Jammes, S., Allainé, D., Bonenfant, C.: Revisiting animal photo-identification using deep metric learning and network analysis. Methods in Ecology and Evolution **12**(5), 863–873 (2021)
41. Moreira, T.P., Perez, M.L., Werneck, R.d.O., Valle, E.: Where is my puppy? retrieving lost dogs by facial features. Multimedia Tools and Applications **76**(14), 15325–15340 (2017)
42. Moskvyak, O., Maire, F., Dayoub, F., Armstrong, A.O., Baktashmotlagh, M.: Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. In: 2021 Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8. IEEE (2021)
43. Nepovinnykh, E., Eerola, T., Biard, V., Mutka, P., Niemi, M., Kunnasranta, M., Kälviäinen, H.: Sealid: Saimaa ringed seal re-identification dataset. Sensors **22**(19), 7602 (2022)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
45. Palencia, P., Vada, R., Zanet, S., Calvini, M., De Giovanni, A., Gola, G., Ferroglio, E.: Not just pictures: utility of camera trapping in the context of african swine fever and wild boar management. Transboundary and Emerging Diseases **2023**, 1–9 (2023)
46. Papafitsoros, K., Panagopoulou, A., Schofield, G.: Social media reveals consistently disproportionate tourism pressure on a threatened marine vertebrate. Animal Conservation **24**(4), 568–579 (2021), https://doi.org/10.1111/acv.12656
47. Parham, J.R., Crall, J., Stewart, C., Berger-Wolf, T., Rubenstein, D.: Animal population censusing at scale with citizen science and photographic identification. In: 2017 AAAI Spring Symposium Series (2017), https://www.aaai.org/ocs/index.php/SSS/SSS17/paper/viewPaper/15245
48. Pedersen, M., Haurum, J.B., Moeslund, T.B., Nyegaard, M.: Re-identification of giant sunfish using keypoint matching. In: Proceedings of the Northern Lights Deep Learning Workshop. vol. 3 (2022)
49. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
50. Renò, V., Dimauro, G., Labate, G., Stella, E., Fanizza, C., Cipriano, G., Carlucci, R., Maglietta, R.: A SIFT-based software system for the photo-identification of the Risso's dolphin. Ecological informatics **50**, 95–101 (2019), https://www.sciencedirect.com/science/article/pii/S1574954118301377

51. Rowcliffe, J.M., Field, J., Turvey, S.T., Carbone, C.: Estimating animal density using camera traps without the need for individual recognition. Journal of Applied Ecology pp. 1228–1236 (2008)
52. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011)
53. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
54. Schneider, J., Murali, N., Taylor, G.W., Levine, J.D.: Can Drosophila melanogaster tell who's who? PloS one **13**(10), e0205043 (2018)
55. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823 (2015)
56. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., et al.: Bioclip: A vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19412–19424 (2024)
57. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021)
58. Trotter, C., Atkinson, G., Sharpe, M., Richardson, K., McGough, A.S., Wright, N., Burville, B., Berggren, P.: NDD20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. arXiv preprint arXiv:2005.13359 (2020), https://arxiv.org/abs/2005.13359
59. Trust, B.P.C.: Panthera pardus csv custom export (2022), https://africancarnivore.wildbook.org/
60. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. Advances in Neural Information Processing Systems **33** (2020)
61. Ueno, M., Kabata, R., Hayashi, H., Terada, K., Yamada, K.: Automatic individual recognition of Japanese macaques (Macaca fuscata) from sequential images. Ethology **128**(5), 461–470 (2022), https://onlinelibrary.wiley.com/doi/full/10.1111/eth.13277
62. Vidal, M., Wolf, N., Rosenberg, B., Harris, B.P., Mathis, A.: Perspectives on individual animal identification from biology and computer vision. Integrative and comparative biology **61**(3), 900–916 (2021)
63. Wahltinez, O., Wahltinez, S.J.: An open-source general purpose machine learning framework for individual animal re-identification using few-shot learning. Methods in Ecology and Evolution **15**(2), 373–387 (2024)
64. Wang, L., Ding, R., Zhai, Y., Zhang, Q., Tang, W., Zheng, N., Hua, G.: Giant panda identification. IEEE Transactions on Image Processing **30**, 2837–2849 (2021), https://ieeexplore.ieee.org/document/9347819
65. Weideman, H., Stewart, C., Parham, J., Holmberg, J., Flynn, K., Calambokidis, J., Paul, D.B., Bedetti, A., Henley, M., Pope, F., Lepirei, J.: Extracting identifying contours for African elephants and humpback whales using a learned appearance model. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1276–1285 (2020)
66. Weideman, H.J., Jablons, Z.M., Holmberg, J., Flynn, K., Calambokidis, J., Tyson, R.B., Allen, J.B., Wells, R.S., Hupman, K., Urian, K., et al.: Integral curvature representation and matching algorithms for identification of dolphins and whales.

In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 2831–2839 (2017)
67. Weissbrod, A., Shapiro, A., Vasserman, G., Edry, L., Dayan, M., Yitzhaky, A., Hertzberg, L., Feinerman, O., Kimchi, T.: Automated long-term tracking and social behavioural phenotyping of animal colonies within a semi-natural environment. Nature communications **4**(1), 2018 (2013)
68. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1492–1500. Honolulu (2017)
69. Yao, H., Zhang, S., Zhang, D., Zhang, Y., Li, J., Wang, Y., Tian, Q.: Large-scale person re-identification as retrieval. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 1440–1445. IEEE (2017)
70. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 694–699 (2002)
71. Zhao, X., Wu, X., Chen, W., Chen, P.C.Y., Xu, Q., Li, Z.: Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. IEEE Transactions on Instrumentation & Measurement **72**, 1–16 (2023). https://doi.org/10.1109/TIM.2023.3271000, https://arxiv.org/pdf/2304.03608.pdf