



**HAL**  
open science

## Revisiting animal photo-identification using deep metric learning and network analysis

Vincent Miele, Gaspard Dussert, Bruno Spataro, Simon Chamailé-Jammes,  
Dominique Allainé, Christophe Bonenfant

### ► To cite this version:

Vincent Miele, Gaspard Dussert, Bruno Spataro, Simon Chamailé-Jammes, Dominique Allainé, et al.. Revisiting animal photo-identification using deep metric learning and network analysis. *Methods in Ecology and Evolution*, 2021, 12 (5), pp.863-873. 10.1111/2041-210X.13577 . hal-03425925

**HAL Id: hal-03425925**

**<https://hal.science/hal-03425925v1>**

Submitted on 11 Nov 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 Revisiting animal photo-identification using deep 2 metric learning and network analysis

**Vincent Miele<sup>1</sup>, Gaspard Dussert<sup>1</sup>, Bruno Spataro<sup>1</sup>, Simon  
Chamaillé-Jammes<sup>2,3,4</sup>, Dominique Allainé<sup>1,4</sup>, and Christophe  
Bonenfant<sup>1,4</sup>**

<sup>1</sup>Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de  
<sup>3</sup> Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France

<sup>2</sup>CEFE, Univ Montpellier, CNRS, EPHE, IRD, Univ Paul Valéry Montpellier 3, Montpellier,  
France

<sup>3</sup>Mammal Research Institute, Department of Zoology & Entomology, University of Pretoria,  
Pretoria, South Africa

<sup>4</sup>LTSER France, Zone Atelier "Hwange", Hwange National Park, Bag 62, Dete, Zimbabwe

## 4 Correspondence

5 V. Miele

6 Email: [vincent.miele@univ-lyon1.fr](mailto:vincent.miele@univ-lyon1.fr)

1 **Abstract**

2 1. An increasing number of ecological monitoring programs rely on photographic  
3 capture-recapture of individuals to study distribution, demography and abundance of  
4 species. Photo-identification of individuals can sometimes be done using idiosyncratic  
5 coat or skin patterns, instead of using tags or loggers. However, when performed  
6 manually, the task of going through photographs is tedious and rapidly becomes too  
7 time consuming as the number of pictures grows.

8 2. Computer vision techniques are an appealing and unavoidable help to tackle  
9 this apparently simple task in the big-data era. In this context, we propose to revisit  
10 animal re-identification using image similarity networks and metric learning with  
11 convolutional neural networks (CNNs), taking the giraffe as a working example.

12 3. We first developed an end-to-end pipeline to retrieve a comprehensive set of  
13 re-identified giraffes from about 4,000 raw photographs. To do so, we combined  
14 CNN-based object detection, SIFT pattern matching, and image similarity networks.  
15 We then quantified the performance of deep metric learning to retrieve the identity of  
16 known individuals, and to detect unknown individuals never seen in the previous years  
17 of monitoring.

18 4. After a data augmentation procedure, the re-identification performance of the  
19 CNN reached a Top-1 accuracy of about 90%, despite the very small number of  
20 images per individual in the training data set. While the complete pipeline succeeded  
21 in re-identifying known individuals, it slightly under-performed with unknown  
22 individuals.

23 5. Fully based on open-source software packages, our work paves the way for  
24 further attempts to build automatic pipelines for re-identification of individual animals,  
25 not only in giraffes but also in other species.

- 1 **Keywords:** individual identification, deep metric learning, image similarity networks,
- 2 open-source software

## 1 | Introduction

2 In many respects, population and behavioural ecology have immensely benefited from  
3 individual-based, long term monitoring of animals in wild populations (Clutton-Brock &  
4 Sheldon, 2010; Hayes & Schradin, 2017). At the heart of such monitoring is the ability  
5 to recognize individuals. Individual identification is often achieved by actively marking  
6 animals, such as deploying ear-tags or leg rings, cutting fingers or feathers, or  
7 scratching scales in reptiles (Silvy *et al.*, 2005). In some species, however, individuals  
8 display natural marks that make them uniquely identifiable. For instance, many large  
9 African mammals such as leopard (*Panthera pardus*), zebra (*Equus sp.*), kudu  
10 (*Tragelaphus strepsiceros*), wildebeest (*Connochaetes taurinus*) or giraffe (*Giraffa*  
11 *camelopardalis*), all present idiosyncratic fur and coat patterns particularly useful for  
12 non-invasive and reliable recognition of individuals. Individual identification in the wild  
13 has long been known to be feasible from comparisons of the distinctive coat patterns  
14 of individuals (Estes, 1991). As the number of individuals to identify increases,  
15 however, people-based visual comparisons of pictures can rapidly become  
16 overwhelming. With the recent move to digital technologies (namely digital cameras  
17 and camera traps), the problem becomes even more acute as the number of pictures  
18 to process can easily reach the thousands or ten of thousands.

19 Over the last decade, the use of computer vision rapidly spread into biological  
20 sciences to become a standard tool in animal ecology for many repetitive tasks  
21 (Weinstein, 2018). In a seminal publication, Bolger *et al.* (2012) first presented  
22 computer-aided photo-identification, initially for giraffes but more recently applied to  
23 dolphins (Renó *et al.*, 2019). The underlying computer technique is a feature matching  
24 algorithm, the Scale Invariant Feature Transform operator (SIFT; Lowe (2004)), where  
25 each image is associated to the  $k$ -nearest best matches. The current use of SIFT for  
26 ecologists requires human intervention to validate the proposed candidate images

1 within a graphical interface (Bolger *et al.*, 2011). In the same vein, other feature-based  
2 proposals were developed in the last decade to apply computer vision to different  
3 types of idiosyncrasies (Hartog & Reijns, 2014; Moya *et al.*, 2015). A drawback of the  
4 method frequently arises when two images are considered similar not because of  
5 similar skin or coat patterns of animals, but because of similarities in the backgrounds  
6 (presence of distinctive tree for instance), hence leading to false positive matches. For  
7 the best results with computer vision, all images should be cropped before, so that  
8 only the relevant part of the animal appears in the images to be analyzed and  
9 compared (for instance, excluding most of the neck, head, legs and background for  
10 large herbivores). Until now, this cropping operation was most often done manually  
11 (Halloran *et al.*, 2015), despite being a highly time-consuming task when processing  
12 thousands of images.

13 Meanwhile, the Deep Learning (DL) revolution was underway in computer vision,  
14 showing breakthrough performance improvements (Christin *et al.*, 2019). In particular,  
15 convolutional neural networks (CNNs) are now the front-line computer technique to  
16 deal with a large range of image processing questions in ecology and environmental  
17 sciences (Lamba *et al.*, 2019). Many recent studies tackle the general problem of  
18 re-identification using CNNs, which has been mostly developed and extensively used  
19 for humans (Wu *et al.*, 2019). Technically, re-identification consists in using a CNN to  
20 classify images of different individuals, some of them being not necessarily seen  
21 before, *i.e.* unknown individuals. However, despite the availability of proven and  
22 efficient techniques (Zheng *et al.*, 2016), and several successful attempts to apply the  
23 method to non-human species (Körschens *et al.*, 2018; Hansen *et al.*, 2018;  
24 Moskvayak *et al.*, 2019; Bouma *et al.*, 2019; Schofield *et al.*, 2019; He *et al.*, 2019;  
25 Bogucki *et al.*, 2019; Schneider *et al.*, 2020; Chen *et al.*, 2020; Ferreira *et al.*, 2020),  
26 re-identification remains a challenging task when applied to animals in the wild where  
27 re-observations are limited in number to train the model satisfactorily *sensu largo*

1 (Schneider *et al.*, 2019).

2 In practice, current CNN-based approaches have to be tailored to the needs of  
3 field ecologists interested in using these tools for individual recognition. For instance,  
4 batches of new images are regularly added to the reference database following yearly  
5 fieldwork sessions because of the recruitment of newborns or of immigrants if the  
6 study population is demographically open. Therefore, we expect the re-sighting of  
7 known individuals, as well as the observation of individuals never seen before. In other  
8 words, this standard sampling design implies to solve the re-identification in a mixture  
9 of known and unknown individuals. Chen *et al.* (2020) referred to this problem as the  
10 "open set" identification problem, and they proposed to identify images from unknown  
11 individuals and to assign them a single "unknown" label. Automatically identifying  
12 currently unknown individuals speeds up the picture sorting process, and facilitates  
13 adding them to the database of individuals whose life history is monitored.

14 A classical CNN classifier can re-identify already known individuals (usually with a  
15 *softmax* last layer) but will fail to identify new individuals because the number of  
16 predicted classes must match the number of known individuals. We therefore crucially  
17 need a CNN-based approach that can filter out individuals unknown at the time of the  
18 analysis. We propose to rely on deep metric learning (DML, see Hoffer & Ailon, 2015)  
19 as an ideal candidate to solve the "open set" identification problem. DML consists in  
20 training a CNN model to embed the input data (input images) into a multidimensional  
21 Euclidean space such that data from a common class (for instance, images of a given  
22 individual) are, in terms of Euclidean distance, much closer than with the rest of the  
23 data.

24 Here we addressed the problem of photo-identification with an updated,  
25 open-source, and end-to-end automatic pipeline applied to the case of the iconic,  
26 endangered giraffe (*Giraffa camelopardalis*). In a first step, we applied state-of-the art  
27 techniques for object detection with CNNs (Lin *et al.*, 2017) to automatically crop

1 giraffe flanks of about 4,000 raw photographs shot in the field at Hwange National  
2 Park, Zimbabwe. Indeed, the most recent CNN approaches clearly outperformed  
3 other approaches (Girshick *et al.*, 2014), including the Histogram of Oriented  
4 Gradients (HOG) approach that was recently used with giraffes too (Buehler *et al.*,  
5 2019). Second, following Bolger *et al.* (2012), we used the SIFT operator to calculate  
6 a numeric distance between all pairs of giraffe flanks. From the  $n \times n$  calculated  
7 distances, we followed the new framework of image similarity network (Wang *et al.*,  
8 2018) and applied unsupervised learning to retrieve different clusters of images  
9 coming from different individuals, hence removing any human intervention in the  
10 process of individual identification. Third, we manually validated a subset of our  
11 results to build a ground-truth data set of different individuals ( $n = 82$ ). Using this data  
12 set as a training set, we developed a supervised learning strategy using CNNs and  
13 evaluated its predictive accuracy with a cross-validation procedure.

## 14 **2 | Material and Methods**

### 15 **2.1 | Photograph database**

16 We carried out this study in the northeast of Hwange National Park (HNP), Zimbabwe.  
17 HNP park covers a 14,650 km<sup>2</sup> area (Chamailé-Jammes *et al.*, 2009). The giraffe  
18 sub-species currently present in HNP could be either *G. c. angolensis* or *G. c. giraffa*  
19 according to the IUCN (Muller *et al.*, 2018). Here we used data from a regular  
20 monitoring of individuals conducted between 2014 and 2018. Each year for at least  
21 three consecutive weeks, we drove the road network daily within <60km of the HNP  
22 Main Camp station, and took photographs of every giraffe encountered. Pictures were  
23 taken with 200mm to 300mm lenses mounted on Nikon DSRL cameras (sensor  
24 resolution ranged between 16 and 40 Mpx). When taking photographs in the field the  
25 camera burst mode is often set on producing sequences of very similar photographs in



1 the same second. For these sequences, we retained one single photograph per  
2 sequence yielding in total a set of  $n = 3,940$  photographs.

## 3 **2.2 | Image cropping with CNN and transfer learning using RetinaNet**

4 A range of CNN-based tools are now available for object detection and already used  
5 for animal detection (Parham *et al.*, 2018; Schneider *et al.*, 2018;  
6 Sadeh Norouzzadeh *et al.*, 2019). Among other options including YOLO (Redmon  
7 *et al.*, 2016; Bochkovskiy *et al.*, 2020) and Mask R-CNN (He *et al.*, 2017), RetinaNet  
8 (Lin *et al.*, 2017) is a CNN-based object detector able to detect a series of predefined  
9 object classes (*e.g.* different animal species) that returns the coordinates of a  
10 bounding box around these objects, and a confidence score as well. These two steps  
11 are performed at the same time with a single CNN, which makes RetinaNet a fast  
12 *one-stage* detector as opposed to two-stage detectors for which a first CNN searches  
13 for regions containing a potential object, and a second CNN classifies these regions  
14 (Redmon *et al.*, 2016). Moreover, RetinaNet allows for a better management of non  
15 informative objects' background (Lin *et al.*, 2017). Finally, it is known that the more  
16 heterogeneous the training data set is (various positions, backgrounds, scale or  
17 lighting), the most efficient a CNN is (Beery *et al.*, 2018), so we used data  
18 augmentation (flipping, rotation and color changes of photographs) to enhance our  
19 model performance.

20 For an efficient detection and classification of objects, a CNN has to be trained on  
21 a huge amount of images (usually  $>$  millions of images) to capture the most  
22 discriminant features associated with each class. Because of the limited number of  
23 photographs we have at hand, we relied on *transfer learning* (Shin *et al.*, 2016).  
24 Transfer learning is a specific method aiming at training a CNN on a small number of  
25 images that do not start CNN training "from scratch" with some random model  
26 parameters, but uses the parameters of a model previously trained on a large data set

1 and for similar tasks as the one of interest (Willi *et al.*, 2019). This approach works  
2 because the pre-trained model has already learnt a wide range of relevant and  
3 generic features.

4 We manually prepared our training data set by cropping bounding boxes around  
5 giraffe flanks, excluding most of the neck, head, legs and background, with the  
6 `labelImg` open source program for image annotation  
7 (<https://github.com/tzutalin/labelImg>). We obtained 469 bounding boxes  
8 associated to a subset of 400 photographs. We performed transfer learning with  
9 RetinaNet to detect a single object class, the giraffe flank, from a pre-trained model  
10 shipped with RetinaNet, that is a ResNet50 backbone trained on the COCO dataset  
11 (80 different classes of common objects including giraffes among a few other animal  
12 species; see Lin *et al.* (2014)). We trained the model with 30 epochs of 100 batches of  
13 size 2. Our pipeline was based on the Keras implementation of RetinaNet available at  
14 <https://github.com/fizyr/keras-retinanet>.

## 15 **2.3 | Identification of individuals using unsupervised learning**

### 16 **2.3.1 | Using the Scale Invariant Feature Transform operator**

17 We built on Bolger *et al.* (2012) to achieve pattern matching between giraffe flanks  
18 with the SIFT operator (Lowe, 2004), currently the most commonly used computer  
19 vision approach to identify individuals (Bellavia & Colombo, 2020). The SIFT algorithm  
20 extracts characteristic features in photographs called *key points* that are invariant with  
21 respect to scale and orientation. Comparing two photographs, pairs of matching key  
22 points (*i.e.* having similar characteristics) are retrieved and ranked by the respective  
23 Euclidean distance between their feature vectors. Here, we selected the 25 closest  
24 pairs of key points. However, for better results, we had to assess the extent to which  
25 matching key points were consistent in the two giraffe flanks, *i.e.* if their location  
26 matched on the giraffe body. To find out relevant cases where matching key points

1 were actual matches of coat patterns, we superimposed key points extracted from a  
2 pair of giraffe photographs with a geometrical transformation called *homography*. An  
3 homography is a perspective transformation between two planes, one for each image,  
4 that finds key points from the first image as close as possible from those of the second  
5 image. The homography preserves the relative positioning of key points but changes  
6 the perspective, *i.e.* the distance between points. Once retrieved from the 2 images,  
7 we superimposed the key points on a plane to compute the Euclidean distance  
8 between all pairs of key points in a pair of photographs, hence obtaining our  
9 SIFT-based distance. We used the implementation of SIFT and homography in the  
10 open source `openCV` library version 3.4 (Bradski, 2000).

## 11 **2.4 | Image similarity network, community detection and clusters of images**

12 Following the computation of distances between all pairs of giraffe flanks obtained with  
13 the SIFT operator approach, we searched for clusters of flank images that should  
14 come from one single individual giraffe. We first defined a network made of nodes and  
15 representing giraffe flank images, and of edges: we considered that two nodes were  
16 connected by an edge, *i.e.* two flanks were similar and came from the same giraffe  
17 individual if the SIFT-based distance between paired images felt below a given  
18 threshold (see below for more details). Therefore, the so-called *connected*  
19 *components* of this network should associate images from different individuals.

20 We estimated the distance threshold value by taking advantage of a property of  
21 complex networks called the *explosive percolation* (Achlioptas *et al.*, 2009). The  
22 explosive percolation predicts a phase transition of the network just above a threshold  
23 point. At this point, adding a small number of edges in the network, for example by  
24 slightly increasing the distance threshold (Hayasaka, 2016), leads to the sudden  
25 appearance of a *giant component* encompassing the majority of nodes. In other  
26 words, at some point, a small increase of the distance threshold leads to considering

1 that almost all images come from the same giraffe. We determined this threshold  
2 value graphically, selecting the transition point where the giant component starts to  
3 increase dramatically (Supp. Fig. fig:giant).

4 An additional issue arose when different nodes were erroneously connected  
5 (example in Figure S1), *i.e.* when two flanks were erroneously considered similar.  
6 Moreover, in some cases the body of two or more giraffes could overlap in one  
7 photograph. In this situation, two or more nodes might be linked by edges, when we  
8 actually should consider different giraffes. To solve this problem, we applied a network  
9 clustering algorithm called *community detection*, developed in network science  
10 (Fortunato, 2010), to split – only when relevant – any connected component into  
11 different groups of nodes that are significantly much more connected between  
12 themselves than with the others, a so-called *community*. Indeed, the presence of  
13 many edges inside a group of images suggested it was consistent and taken from the  
14 same individual, whereas the absence of many edges between two groups clearly  
15 informed about their inconsistency and heterogeneity (*i.e.* from two different  
16 individuals). We applied the community detection with the InfoMap algorithm (Rosvall  
17 & Bergstrom, 2008). The final product of the community detection algorithm was a set  
18 of *clusters* of images corresponding either to a connected component or to a  
19 community retrieved by InfoMap.

## 20 **2.5 | Re-identification of individuals, using supervised learning**

### 21 **2.5.1 | Deep metric learning and triplet loss with CNN**

22 The principle of deep metric learning is to find an optimal way to project images into  
23 an Euclidean space such that the Euclidean distance can be used for machine  
24 learning tasks. In this context, we trained a CNN model using the triplet loss (Hermans  
25 *et al.*, 2017), in line with recent studies on other species (Moskvyak *et al.*, 2019;  
26 Bouma *et al.*, 2019). The triplet loss principle relies on triplets of images composed by

1 a first image called *anchor* and another *positive* image of the same class (same giraffe  
2 here) and a third *negative* image of another class (any different giraffe) (see Bouma  
3 *et al.*, 2019, for details). The training step consists in optimizing the CNN model such  
4 that the Euclidean distance computed using the last CNN layer (hereafter called  
5 CNN-based distance) between any anchor and its positive image is minimal, while  
6 maximizing the distance between this anchor image with its negative counterpart. We  
7 used an improved algorithm called *semi-hard triplet loss* (Schroff *et al.*, 2015), that  
8 deals only with triplets where the positive and negative images are close (in other  
9 words, the "hard" cases), using the TripletSemiHardLoss function in TensorFlow  
10 Addons. After training completion, we computed the Euclidean distances between any  
11 pair of giraffe flank photographs, again using the vector composing the last layer of our  
12 CNN model.

### 13 **2.5.2 | Data augmentation, training and test data sets**

14 We derived the training and test data sets required for the CNN approach from the  
15 photograph clusters identified by the SIFT algorithm. We retained only those clusters  
16 fulfilling the following conditions: (i) the cluster contains a minimum of two sequences  
17 of images shot at least 1 hour apart; (ii) the cluster can be divided into a first set of  
18 sequences large enough to perform training (we imposed at least five images), and a  
19 second set of sequences; (iii) the cluster demonstrated a perfect and verified  
20 consistency. We used the first set of sequences for CNN training, and the second as  
21 an independent test data set to assess the model performance. The first condition  
22 ensured that we have complete independence between training and test data sets, *i.e.*  
23 giraffes being seen under different conditions (time, season or location). The third  
24 condition is of utmost importance because errors in the data set would lead to  
25 sub-optimal performances of the machine learning approach. We therefore carefully  
26 checked, manually, that the SIFT-based clusters we used in the CNN were perfectly

1 unambiguous. We achieved this high level of data quality by discarding all cases  
2 where two or more giraffes overlapped on the same frame, or when giraffes were  
3 indifferently oriented from the back to the front (orientation ambiguities).

4 We cropped all flank images to focus on the central part of the flank, keeping 80%  
5 of the original width and 60% of the height (in particular excluding the neck and its  
6 background). By doing so, we wanted to prevent our CNN model from capturing  
7 background noise. Additionally, we homogenized contrast of images by normalizing the  
8 three colour channels using the `Imagemagick` package (`normalize` option;  
9 <https://imagemagick.org>). In a final step, we resized all images to 224x224  
10 pixels.

11 We ended up with five flanks per individual at least, and a median of seven (Table  
12 1) in the training set. This particularly low number of images available to train the CNN  
13 led us to consider the few shot learning framework, a class of problems where only a  
14 few images are available for training. We implemented a 10-fold data augmentation  
15 procedure where we made extensive use of image augmentation using the `imgaug`  
16 Python library (<https://github.com/aleju/imgaug>). For each image in the  
17 training data set, we performed a random set of transformations such as modifying  
18 orientation and size, adding blur, performing edge detection, adding Gaussian noise  
19 and modifying colors or brightness (details in the available Python code). We finally  
20 used this set of eleven images per original image to train our CNN model, *i.e.* the  
21 original one and ten modified versions of this image.

## 22 **2.6 | Evaluation of CNN-based re-identification**

23 To quantify the overall predictive performance of our CNN deep metric learning, we  
24 replicated the following procedure ten times. We first randomly selected 25% of the  
25 individuals of the data set and, for the purpose of the evaluation here, considered  
26 these as *unknown* individuals. Then, for each of them, we randomly selected two

1 images, one in each of the sequences (see above). With this data set, we aimed to  
2 test the ability of the CNN model to detect unknown individuals. The remaining 75%  
3 individuals were considered known individuals. For these known individuals, we  
4 selected all photographs from the first sequence and used it to built a training data set  
5 for the CNN. We kept all images from the remaining sequences as the test data set for  
6 *known* individuals. This ensured a good independence between training and test data,  
7 mostly thanks to the one hour (at least) time lag between observations. Once the  
8 selection of individuals was completed, we performed transfer learning using the  
9 pre-trained model ResNetV2 readily available in Keras. We estimated the model  
10 parameters using the augmented training data set with 80 epochs with batches of size  
11 42. We used the stochastic gradient descent optimizer with a rate of 0.2. Our pipeline  
12 was implemented with Keras 2.3.0.

13 To mimic re-identification *per se*, literally re-seeing known individuals, we  
14 considered that we had a "reference book" with five *representative* images per known  
15 individuals: these images were randomly drawn out of the training data set. We then  
16 calculated the CNN-based distance between these representative images and each  
17 image from the test data set. In essence, we expected small distances between test  
18 images and representative ones when they came from the same known individual.  
19 Similarly, we calculated the CNN-based distance between representative images and  
20 images of the so-called unknown individuals. We also considered that two images can  
21 come from the same individual if their distance was below a given threshold. This  
22 distance threshold was a stringency condition that arbitrarily varied between 0 and 1.

23 We quantified the predictive performance of the trained CNN model on the range  
24 of distance threshold values. First, we computed *Top-1* accuracy for known individuals,  
25 consisting in checking for each query image if a representative image from the same  
26 individual was the one with smallest distance (*i.e.* the *Top-1* image) and with a  
27 distance below the threshold. In the following, *Top-1* accuracy was also called *true*

1 *positive* (TP) rate. Then, we computed the *false positive* rate (FP), checking cases  
2 where the Top-1 image was from a different individual. Finally, we quantified the CNN  
3 ability to sort out images from unknown individuals. Again, over the range of distance  
4 threshold values, we checked if Top-1 image of unknown individual images fell below  
5 the threshold. If not, we considered that we successfully detected an unknown  
6 individual, hence computing the *true negative* (TN) rate.

## 7 **3 | Results**

### 8 **3.1 | From thousands of photographs to thousands of images of giraffe** 9 **flank.**

10 We trained the object detection method with RetinaNet (Lin *et al.*, 2017) on a set of  
11 400 photographs for which the cropping of the giraffe flank has been previously done  
12 manually. Training took approximately 30 minutes on a Titan X card. When applying  
13 the automatic cropping procedure on our 3,940 photographs (see Figure 1a), we  
14 retrieved 5,019 images with associated bounding boxes, supposed to contain a single  
15 giraffe flank (see Figure 2a). The cropping failed for 186 photographs (failure rate:  
16 4.7%), mostly due to foreground vegetation and, unusual and difficult orientation of  
17 giraffes in the photograph (see examples on Figure 1b). In a few cases, a bounding  
18 box could contain the bodies of two overlapping giraffes, one being partially in front of  
19 the other (see Figure 2a). Similarly, in some rare instances giraffes were standing very  
20 close to each other on a photograph, a situation where RetinaNet could fail in  
21 retrieving the exact boundaries of each giraffe flank (see the worst case that we  
22 experienced, from a partially blurry photograph in Figure 2b).



## 1 **3.2 | From thousands of images down to hundreds of identified individ-** 2 **uals**

3 Running the SIFT algorithm (Lowe, 2004) to compare all pairs of flanks took about  
4 800 CPU hours of heterogeneous computing resources. We estimated the threshold  
5 value for the giant component (see Methods) at a distance of 340 (see Figure S2a),  
6 and obtained an image similarity network composed of 5,019 nodes and 11,249  
7 edges, yielding 1,417 connected components among which 781 were singletons of  
8 one image.

9 Our network-based approach, relying on community detection, retrieved consistent  
10 *clusters* of flank images (different colors in Figure 3). The cluster size distribution is by  
11 definition more concentrated after network clustering (see Figure S3) with a maximal  
12 size of 35 instead of 373. Indeed, this very large connected component was clearly an  
13 artifact due to a chain of giraffe overlaps, and has been successfully split by our  
14 procedure (see Figure S4). We detected 316 clusters with more than 5 images, and  
15 105 with more than 10 images. However, in rare cases, some images from the same  
16 individuals were found in different clusters (see Figure S4). Because these clusters  
17 arose from a single connected component, we could *a posteriori* check for  
18 consistencies by comparing clusters of the same component manually (such as  
19 performed for Figure S4).

## 20 **3.3 | From identified individuals to a deep learning approach for re-iden-** 21 **tification**

22 To perform a fair evaluation of the CNN performance, we saved 82 human-validated,  
23 unambiguous SIFT-based clusters that contained at least two different sequences of  
24 photographs shot at least with a one hour interval (see Material and Methods). Those  
25 82 clusters were made of 822 images of giraffe flanks from which we evaluated the

1 performance of our re-identification pipeline based on deep metric learning. Once  
2 trained using data augmentation, the CNN returned a Top-1 accuracy (TP rate) of  
3 about 85% on average (Figure 5) for images of known individuals. However, eleven  
4 images were found to be repeatedly impossible to classify because of bad orientation  
5 of the giraffe body on the photograph, or because of the presence of conspicuous and  
6 disturbing elements at the forefront (Supp. Figure S6). Without these problematic  
7 images, we achieved a Top-1 accuracy  $>90\%$ , on average. Interestingly, the  
8 associated false positive rate was close to 0 (Figure 5). In other words, when a Top-1  
9 image existed below a given threshold (here 1. at most), this Top-1 image was almost  
10 always from the correct known individual (Supp. Figure S5 a).

11 With our deep metric learning approach, images were projected into an Euclidean  
12 space. We expected images from the same known individual to be close in this space,  
13 whereas images from unknown individuals should be distant from those of known  
14 individuals. This prediction was partly supported only. If, for small distance threshold  
15 values ( $d \leq 0.1$ ) the true negative rate was  $TN > 95\%$ , TN decreased markedly with  
16 the distance threshold (Figure 5). At the same time, the positive rate started from  
17  $TP < 70\%$  for ( $d \leq 0.1$ ) but rapidly levelled off to 80% as the distance threshold  
18 increased (Figure 5). Hence, our CNN often predicted an unexpected small distance  
19 between a given image of unknown individual and another image of a known  
20 individual (Supp. Figure S5 b). Interestingly, a particular threshold value ( $d = 0.25$ ;  
21 crossing point in Figure 5) where both TP and TN rates reached 80% offered the best  
22 compromise.

## 23 **4 | Discussion**

24 We propose two complementary approaches to re-identify individual giraffes from a  
25 set of photographs taken in the field. Based on the new framework of image similarity  
26 networks, our unsupervised method goes one step further compared to previous

1 solutions from the literature since its end product is a comprehensive list of clusters of  
2 images, one cluster per identified individuals. Our supervised method, that relies on  
3 deep metric learning, achieves a very good re-identification of giraffes from a  
4 "reference book" of known individuals despite the rather small number of photographs  
5 per individuals available to train the model.

6       As a first step, we took advantage of the most recent computer vision techniques  
7 to perform object detection and crop the giraffe flanks before comparing coat patterns  
8 of giraffes. Image cropping proves to be particularly efficient when the body of several  
9 giraffes do not overlap in photographs. However, cascade of problems arises when  
10 overlapping occurs, including erroneous cropping and difficulties to assign a bounding  
11 box to a single individual because in this case, the coat patterns of two individuals are  
12 mixed. We show that a limited number of labeled photographs is needed to train  
13 RetinaNet (a few hundreds) with a very good performance on new photographs. To  
14 what extent our RetinaNet model parameters could be efficient in other study sites  
15 with different background vegetation (in "Terra Incognita", quoting Beery *et al.* (2018))  
16 remains an open question. Nevertheless fine tuning RetinaNet for a particular task  
17 and data set is within the reach of many researchers dealing with animal photographs  
18 thanks to the associated code we provide. Further perspectives now arise with  
19 contour segmentation methods (He *et al.*, 2017) than can extract contours of an object  
20 such as the whole body or any part of an animal by creating a so-called *segmentation*  
21 *mask* (Brodrick *et al.*, 2019). Giraffe body contouring could possibly help for the  
22 individual re-identification by removing background residual noise, but building a  
23 training set by manually contouring hundreds of animal bodies remains a huge effort.

24       We then recast the animal identification problem from photographs into a statistical  
25 one, namely a clustering problem in an image similarity network. In other words, given  
26 a network that we build using a distance between pairs of images, we can efficiently  
27 retrieve the image set of a given individual as a cluster in a network. We computed a

1 distance based on pattern matching between flanks with the well known SIFT operator  
2 (Bellavia & Colombo, 2020) as used by Bolger *et al.* (2012). The proposed  
3 network-based approach was particularly useful and efficient to deal with false positive  
4 matches. False positive matches are a recurrent issue occurring when two images  
5 have very similar background. This situation is often found when the same tree  
6 appears on two images (see nodes 3 and 4 in Figure 3), when giraffe orientation  
7 perfectly matches (see Figure S1), or when the bodies of two giraffes overlap on the  
8 same image, which is the most frequent configuration we faced (see node 2 in Figure  
9 3). In this latter case, this image linked two sets of images corresponding to the two  
10 overlapping individuals. Our network-based approach also handles false negative  
11 cases (*e.g.* two images of the same animal are declared different because of  
12 differences in lighting conditions or animal orientation) since community detection is  
13 robust to possibly missing edges: indeed, a missing edge can be compensated by the  
14 other edges inside a cluster. This step is fully reproducible and applicable to other  
15 animal species, as long as a feature matching algorithm can be used, be it SIFT or  
16 any other alternative methods such as Oriented FAST and rotated BRIEF (ORB  
17 Rublee *et al.*, 2011), or deep features (Dusmanu *et al.*, 2019; Ma *et al.*, 2020)).

18 We tackled the problem of animal re-identification, literally detecting and identifying  
19 previously seen animals, considering that we had a "reference book" with photographs  
20 of these known individuals. This fits the needs of field researchers that want to  
21 monitor the fate of animals by regularly adding new observations in time, for instance  
22 by collecting photographs with camera traps. To do so, we evaluated the possibility to  
23 use the rapidly developing convolutional neural networks in a supervised learning  
24 framework to achieve deep metric learning. Solving this problem was particularly  
25 challenging because of the size of our data set. Previous studies on animal  
26 re-identification with CNN indeed relied on a high number of photographs per  
27 individuals (Schneider *et al.*, 2020; Ferreira *et al.*, 2020). In our case, we had to train

1 the CNN with a few images per individuals only (see Snell *et al.*, 2017, on *few shot*  
2 *learning* methods) shot in the field with contrasting environmental and light conditions.  
3 This situation corresponds to many field studies, and particularly on large mammals  
4 (possibly with the exception of primates), for which population density and animal  
5 detection rate are low, limiting the expected number of photograph per individuals. To  
6 circumvent this problem, we developed a data augmentation strategy to increase  
7 artificially the variability of observation conditions encountered in the training data set,  
8 and improved the model performance substantially.

9 In terms of overall predictive performance, we reached about 90% Top-1 accuracy,  
10 which is comparable to the previously reported performance in animal re-identification  
11 of known individuals (see Schneider *et al.*, 2019, for a review) but usually achieved  
12 with a much higher number of photographs. The combination of recent deep learning  
13 algorithm and data augmentation appears very competitive and efficient, with possible  
14 application to difficult practical cases like when working on endangered or elusive  
15 species living at very low abundance such as leopard (*Panthera pardus*) or the Iberian  
16 lynx (*Lynx pardinus*). Compared to the more robust SIFT operator, we found that the  
17 performance of the CNN is affected by the orientation of giraffe body and noticeably by  
18 deviation from perfect side shot. In terms of computing requirements, training our CNN  
19 remained time-consuming because the number of images to process is increased  
20 dramatically by the data augmentation. This problem is partially counter-balanced by  
21 the more computationally efficient calculation of CNN-based distances that increases  
22 linearly with the number of photographs (computing one projection per image),  
23 compared to the SIFT-based approach for which the computing time is proportional to  
24 the square of the number of photographs (computing one matching per image pair).  
25 For instance, we got all distances in a minute with the CNN and about two hours with  
26 the SIFT operator when applied on the same test data set (see Table 2).

27 Our approach was also designed to deal with data sets where known and unknown

1 individuals were present. Dealing with unknown individuals is extremely challenging  
2 because no image of these new individuals are available in the training data set.  
3 Indeed, most classical CNN-based approaches solve classification problems where  
4 the number of classes, the number of individuals for us, was fixed. We showed here  
5 that it was possible to filter out unknown from known individuals, while re-identifying a  
6 large fraction of known individuals at the same time with a success of 80% (for both  
7 TP and TN). However, this trade-off came at the cost of a lower Top-1 accuracy, which  
8 we acknowledge is not fully satisfying and already experienced by other authors  
9 (Ferreira *et al.*, 2020). Still, in most cases, we could validate the proposed  
10 identification by examining the Top-1 for each query image (*i.e.* checking its closest  
11 image) for both known and unknown individuals. Despite not being fully automated,  
12 our CNN approach would require little human intervention.

13 To what extent the performance of our CNN-based pipeline could be improved with  
14 more data? Since it is suitable to any species, further data analysis on other species  
15 will help answer this question. However, additional strategies would help including the  
16 integration of contextual information (Beery *et al.*, 2019; Terry *et al.*, 2020) such as  
17 time, GPS positioning or animal social context. Using accurate segmentation of animal  
18 body (He *et al.*, 2017; Brodrick *et al.*, 2019) will undoubtedly be a solution against side  
19 effects of rectangular cropping. Moreover, this pipeline can be used in an active  
20 learning strategy where the machine learning model is assisted by human intervention  
21 on some specific cases (Norouzzadeh *et al.*, 2021). Indeed, using the proposed  
22 distance threshold in the Euclidean space, one can iteratively enrich the training data  
23 set after manual checking of the most confident Top-1 candidates (below a small  
24 distance threshold, to guarantee optimal TN rate) and re-run the estimation procedure.

25 Finally, this inter-disciplinary work provides guidelines about best practices to  
26 collect identification images in the field, if to be used later with an automated pipeline  
27 such as the one presented here. Better results can be achieved with simple framing

1 rules of animals with cameras. First the field operator should try to avoid as much as  
2 possible overlaying bodies of two or more individuals as this was the most acute issue  
3 in our giraffe experience. Note that several but well separated individuals in the same  
4 photograph is not a problem at all thanks to the CNN cropping performed at the  
5 preliminary stage. Another point to pay attention to is the background which, if too  
6 similar on the same images (*e.g.* photographs shot from the very same spot) with  
7 obvious structures (tree, pond, rocks. . . ) will likely mislead the computer vision  
8 algorithm, even on cropped images because cropping is rectangular and do not  
9 delineate the animal body. This situation often arises while photographing animals  
10 moving in line, as giraffes and many others often do. A last point is the heterogeneity  
11 of situations under which animals were observed. We did our best to improve the  
12 training data set with data augmentation, however, photographing animals in as many  
13 different conditions as possible could improve the results. This includes light  
14 conditions (dawn, dusk, noon), orientation of individual or background (open vs. more  
15 densely vegetated areas). More specific to CNN re-identification is the need to have a  
16 greater number of pictures of photographs per individuals ( $> 50$ ) than what is currently  
17 available, so a particular attention should be given, in the field under optimal shooting  
18 conditions, to the opportunity to take more photographs of each observed individual.

## 19 **Acknowledgments**

20 We thank Jeanne Duhayer for her considerable help in analysing our preliminary  
21 findings, and Laurent Jacob and Franck Picard for their insights on deep learning. This  
22 work was performed using the computing facilities of the CC LBBE/PRABI. Funding  
23 was provided by the French National Center for Scientific Research (CNRS) and the  
24 Statistical Ecology Research Group (EcoStat) of the CNRS. We are also grateful to  
25 Derek Lee for his kind advice in processing photographs, and for sharing with us his  
26 experience in the monitoring of giraffes. Finally, we acknowledge the director of the

1 Zimbabwe Parks and Wildlife Management Authority for authorizing this research, and  
2 support from the CNRS Zone Atelier / LTSER program for fieldwork and some of the  
3 photographs (collection by P.A. Seeber).

#### 4 **Authors' contribution**

5 V.M., D.A. and C.B. conceived the study with some inputs from S.C.J. V.M. and G.D.  
6 developed the approach and performed the analysis. V.M. and S.C.J. supervised G.D.  
7 D.A. and C.B. provided the photographs. B.S. set up the computing architecture. All  
8 authors contributed to the writing of the manuscript.

#### 9 **Data Availability**

10 The curated data set of re-identified giraffe individuals is freely available at  
11 `ftp://pbil.univ-lyon1.fr/pub/datasets/miele2020`. The code to  
12 reproduce the analysis is available at  
13 `https://plmlab.math.cnrs.fr/vmiele/animal-reid/` with explanations and  
14 test cases.

#### 15 **References**

- 16 Achlioptas, D., D'Souza, R.M. & Spencer, J. (2009) Explosive percolation in random networks.  
17 *Science*, **323**, 1453–1455.
- 18 Beery, S., Van Horn, G. & Perona, P. (2018) Recognition in terra incognita. *Proceedings of the*  
19 *European Conference on Computer Vision (ECCV)*, pp. 456–473.
- 20 Beery, S., Wu, G., Rathod, V., Votel, R. & Huang, J. (2019) Context r-cnn: Long term temporal  
21 context for per-camera object detection.
- 22 Bellavia, F. & Colombo, C. (2020) Is there anything new to say about sift matching?  
23 *International Journal of Computer Vision*, pp. 1–20.



- 1 Bochkovskiy, A., Wang, C.Y. & Liao, H.Y.M. (2020) Yolov4: Optimal speed and accuracy of  
2 object detection. *arXiv preprint arXiv:200410934*.
- 3 Bogucki, R., Cygan, M., Khan, C.B., Klimek, M., Milczek, J.K. & Mucha, M. (2019) Applying  
4 deep learning to right whale photo identification. *Conservation Biology*, **33**, 676–684.
- 5 Bolger, D.T., Morrison, T.A., Vance, B., Lee, D. & Farid, H. (2012) A computer-assisted system  
6 for photographic mark–recapture analysis. *Methods in Ecology and Evolution*, **3**, 813–822.
- 7 Bolger, D., Vance, B., Morrison, T. & Farid, H. (2011) Wild id user guide: pattern extraction and  
8 matching software for computer-assisted photographic mark.
- 9 Bouma, S., Pawley, M.D.M., Hupman, K. & Gilman, A. (2019) Individual common dolphin  
10 identification via metric embedding learning.
- 11 Bradski, G. (2000) The OpenCV Library. *Dr Dobb's Journal of Software Tools*.
- 12 Brodrick, P.G., Davies, A.B. & Asner, G.P. (2019) Uncovering ecological patterns with  
13 convolutional neural networks. *Trends in ecology & evolution*.
- 14 Buehler, P., Carroll, B., Bhatia, A., Gupta, V. & Lee, D.E. (2019) An automated program to find  
15 animals and crop photographs for individual recognition. *Ecological informatics*, **50**,  
16 191–196.
- 17 Chamaillé-Jammes, S., Valeix, M., Bourgarel, M., Murindagomo, F. & Fritz, H. (2009) Seasonal  
18 density estimates of common large herbivores in Hwange National Park, Zimbabwe. *African  
19 Journal of Ecology*, **47**, 804–808.
- 20 Chen, P., Swarup, P., Wojciech, M.M., Kong, A.W.K., Han, S., Zhang, Z. & Rong, H. (2020) A  
21 study on giant panda recognition based on images of a large proportion of captive pandas.  
22 *Ecology and Evolution*.
- 23 Christin, S., Hervet, E. & Lecomte, N. (2019) Applications for deep learning in ecology.  
24 *Methods in Ecology and Evolution*, **10**, 1632–1644.
- 25 Clutton-Brock, T. & Sheldon, B.C. (2010) Individuals and populations: the role of long-term,  
26 individual-based studies of animals in ecology and evolutionary biology. *Trends in ecology &  
27 evolution*, **25**, 562–573.
- 28 Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A. & Sattler, T. (2019) D2-net:  
29 A trainable cnn for joint description and detection of local features. *Proceedings of the IEEE  
30 Conference on Computer Vision and Pattern Recognition*, pp. 8092–8101.

- 1 Estes, R.D. (1991) The behavior guide to african mammals: including hoofed mammals,  
2 carnivores. *Primates*, pp. 509–519.
- 3 Ferreira, A.C., Silva, L.R., Renna, F., Brandl, H.B., Renoult, J.P., Farine, D.R., Covas, R. &  
4 Doutrelant, C. (2020) Deep learning-based methods for individual recognition in small birds.  
5 *Methods in Ecology and Evolution*, **11**, 1072–1085.
- 6 Fortunato, S. (2010) Community detection in graphs. *Physics reports*, **486**, 75–174.
- 7 Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014) Rich feature hierarchies for accurate  
8 object detection and semantic segmentation. *The IEEE Conference on Computer Vision  
9 and Pattern Recognition (CVPR)*.
- 10 Halloran, K.M., Murdoch, J.D. & Becker, M.S. (2015) Applying computer-aided  
11 photo-identification to messy datasets: a case study of t hornicroft's giraffe (g iraffa  
12 camelopardalis thornicrofti). *African Journal of Ecology*, **53**, 147–155.
- 13 Hansen, M.F., Smith, M.L., Smith, L.N., Salter, M.G., Baxter, E.M., Farish, M. & Grieve, B.  
14 (2018) Towards on-farm pig face recognition using convolutional neural networks.  
15 *Computers in Industry*, **98**, 145–152.
- 16 Hartog, J. & Reijns, R. (2014) Interactive individual identification system (i3s). *Boston, MA:*  
17 *Free Software Foundation Inc.*
- 18 Hayasaka, S. (2016) Explosive percolation in thresholded networks. *Physica A: Statistical  
19 Mechanics and its Applications*, **451**, 1–9.
- 20 Hayes, L.D. & Schradin, C. (2017) Long-term field studies of mammals: what the short-term  
21 study cannot tell us. *Journal of Mammalogy*, **98**, 600–602.
- 22 He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017) Mask r-cnn. *Proceedings of the IEEE  
23 international conference on computer vision*, pp. 2961–2969.
- 24 He, Q., Zhao, Q., Liu, N., Chen, P., Zhang, Z. & Hou, R. (2019) Distinguishing individual red  
25 pandas from their faces. *Chinese Conference on Pattern Recognition and Computer Vision  
26 (PRCV)*, pp. 714–724. Springer.
- 27 Hermans, A., Beyer, L. & Leibe, B. (2017) In defense of the triplet loss for person  
28 re-identification. *arXiv preprint arXiv:170307737*.
- 29 Hoffer, E. & Ailon, N. (2015) Deep metric learning using triplet network. *International  
30 Workshop on Similarity-Based Pattern Recognition*, pp. 84–92. Springer.

- 1 Körschens, M., Barz, B. & Denzler, J. (2018) Towards automatic identification of elephants in  
2 the wild. *arXiv preprint arXiv:181204418*.
- 3 Lamba, A., Cassey, P., Segaran, R.R. & Koh, L.P. (2019) Deep learning for environmental  
4 conservation. *Current Biology*, **29**, R977–R982.
- 5 Lin, T.Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017) Focal loss for dense object  
6 detection. *Proceedings of the IEEE international conference on computer vision*, pp.  
7 2980–2988.
- 8 Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C.L.  
9 (2014) Microsoft coco: Common objects in context. *European conference on computer  
10 vision*, pp. 740–755. Springer.
- 11 Lowe, D.G. (2004) Distinctive image features from scale-invariant keypoints. *International  
12 journal of computer vision*, **60**, 91–110.
- 13 Ma, J., Jiang, X., Fan, A., Jiang, J. & Yan, J. (2020) Image matching from handcrafted to deep  
14 features: A survey. *International Journal of Computer Vision*, pp. 1–57.
- 15 Moskvayak, O., Maire, F., Armstrong, A.O., Dayoub, F. & Baktashmotlagh, M. (2019) Robust  
16 re-identification of manta rays from natural markings by learning pose invariant embeddings.
- 17 Moya, Ó., Mansilla, P.L., Madrazo, S., Igual, J.M., Rotger, A., Romano, A. & Tavecchia, G.  
18 (2015) Aphis: a new software for photo-matching in ecological studies. *Ecological  
19 informatics*, **27**, 64–70.
- 20 Muller, Z., Bercovitch, F., Brand, R., Brown, D., Brown, M., Bolger, D., Carter, K., Deacon, F.,  
21 Doherty, J., Fennessy, J., Fennessy, S., Hussein, A., Lee, D., Marais, A., Strauss, M.,  
22 Tutchings, A. & Wube, T. (2018) *Giraffa camelopardalis* (amended version of 2016  
23 assessment). the IUCN Red List of threatened species 2018: e.t9194a136266699.
- 24 Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N. & Clune, J. (2021) A deep active  
25 learning system for species identification and counting in camera trap images. *Methods in  
26 Ecology and Evolution*, **12**, 150–161.
- 27 Parham, J., Stewart, C., Crall, J., Rubenstein, D., Holmberg, J. & Berger-Wolf, T. (2018) An  
28 animal detection pipeline for identification. *2018 IEEE Winter Conference on Applications of  
29 Computer Vision (WACV)*, pp. 1075–1083. IEEE.
- 30 Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016) You only look once: Unified,

- 1 real-time object detection. *Proceedings of the IEEE conference on computer vision and*  
2 *pattern recognition*, pp. 779–788.
- 3 Renó, V., Dimauro, G., Labate, G., Stella, E., Fanizza, C., Cipriano, G., Carlucci, R. &  
4 Maglietta, R. (2019) A sift-based software system for the photo-identification of the risso's  
5 dolphin. *Ecological informatics*, **50**, 95–101.
- 6 Rosvall, M. & Bergstrom, C.T. (2008) Maps of random walks on complex networks reveal  
7 community structure. *Proceedings of the National Academy of Sciences*, **105**, 1118–1123.
- 8 Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. (2011) Orb: An efficient alternative to sift or  
9 surf. *2011 International conference on computer vision*, pp. 2564–2571. Ieee.
- 10 Sadegh Norouzzadeh, M., Morris, D., Beery, S., Joshi, N., Jojic, N. & Clune, J. (2019) A deep  
11 active learning system for species identification and counting in camera trap images. *arXiv*  
12 *preprint arXiv:191009716*.
- 13 Schneider, S., Taylor, G.W. & Kremer, S. (2018) Deep learning object detection methods for  
14 ecological camera trap data. *2018 15th Conference on Computer and Robot Vision (CRV)*,  
15 pp. 321–328. IEEE.
- 16 Schneider, S., Taylor, G.W. & Kremer, S.C. (2020) Similarity learning networks for animal  
17 individual re-identification-beyond the capabilities of a human observer. *Proceedings of the*  
18 *IEEE Winter Conference on Applications of Computer Vision Workshops*, pp. 44–52.
- 19 Schneider, S., Taylor, G.W., Linqvist, S. & Kremer, S.C. (2019) Past, present and future  
20 approaches using computer vision for animal re-identification from camera trap data.  
21 *Methods in Ecology and Evolution*, **10**, 461–470.
- 22 Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D. & Carvalho, S.  
23 (2019) Chimpanzee face recognition from videos in the wild using deep learning. *Science*  
24 *advances*, **5**, eaaw0736.
- 25 Schroff, F., Kalenichenko, D. & Philbin, J. (2015) Facenet: A unified embedding for face  
26 recognition and clustering. *Proceedings of the IEEE conference on computer vision and*  
27 *pattern recognition*, pp. 815–823.
- 28 Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. & Summers,  
29 R.M. (2016) Deep convolutional neural networks for computer-aided detection: Cnn  
30 architectures, dataset characteristics and transfer learning. *IEEE transactions on medical*

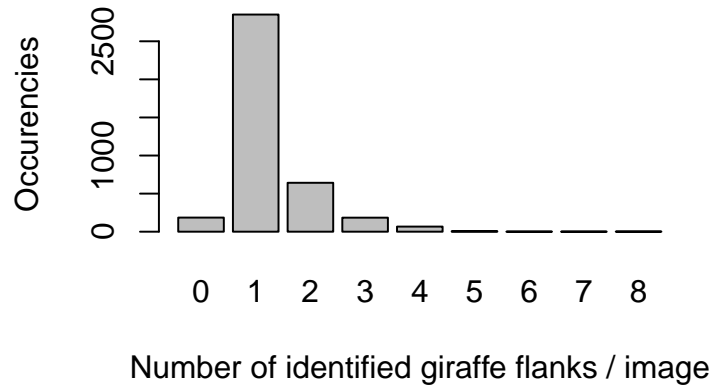
- 1 *imaging*, **35**, 1285–1298.
- 2 Silvy, N.J., Lopez, R.R. & Peterson, M.J. (2005) Wildlife marking techniques. *Techniques for*  
3 *wildlife investigations and management The Wildlife Society, Bethesda, MD*, pp. 339–376.
- 4 Snell, J., Swersky, K. & Zemel, R. (2017) Prototypical networks for few-shot learning.  
5 *Advances in neural information processing systems*, pp. 4077–4087.
- 6 Terry, J.C.D., Roy, H.E. & August, T.A. (2020) Thinking like a naturalist: Enhancing computer  
7 vision of citizen science images by harnessing contextual data. *Methods in Ecology and*  
8 *Evolution*, **11**, 303–315.
- 9 Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C.D., Batzoglu, S. & Leskovec, J.  
10 (2018) Network enhancement as a general method to denoise weighted biological networks.  
11 *Nature communications*, **9**, 1–8.
- 12 Weinstein, B.G. (2018) A computer vision for animal ecology. *Journal of Animal Ecology*, **87**,  
13 533–545.
- 14 Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldhuis, M. &  
15 Fortson, L. (2019) Identifying animal species in camera trap images using deep learning  
16 and citizen science. *Methods in Ecology and Evolution*, **10**, 80–91.
- 17 Wu, D., Zheng, S.J., Zhang, X.P., Yuan, C.A., Cheng, F., Zhao, Y., Lin, Y.J., Zhao, Z.Q., Jiang,  
18 Y.L. & Huang, D.S. (2019) Deep learning-based methods for person re-identification: A  
19 comprehensive review. *Neurocomputing*.
- 20 Zheng, L., Yang, Y. & Hauptmann, A.G. (2016) Person re-identification: Past, present and  
21 future. *arXiv preprint arXiv:161002984*.

**Table 1** Flank images were selected to ensure independence of observation, and then used for individual giraffe re-identification from coat patterns with a convolutional neural network. We tabulated the average number (and the associated range in squared brackets) of images and sequences (*i.e.* separated by at least one hour interval) per individual in the train, test and unknown data sets over 10 trials.

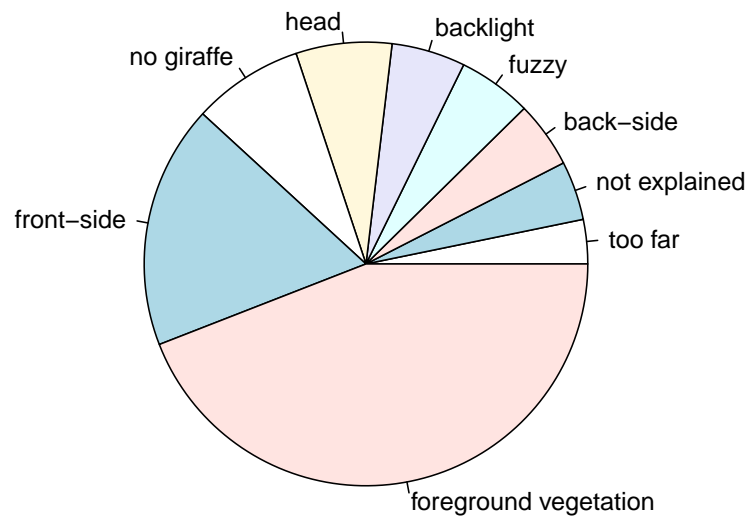
	Nb. images	Nb. indiv.	Nb. images per indiv.	Nb. sequences per indiv.
Train	503 [479-529]	62	7 [5-24]	2 [1-5]
Test	121 [118-126]	62	2 [1-5]	1 [1-4]
Unknown indiv.	40	20	2	2

**Table 2** Computing time needed to compare 310 representative images vs. 121 test images (CNN-training with about 5500 images) extracted from giraffe photographs shot at Hwange National Park, Zimbabwe, between 2014 and 2018. The hardware we used for these calculations was an Intel Xeon CPU E5-2650 v4 2.30GHz (CPU) and Nvidia Titan X card (GPU).

Task	Avg. computing time
SIFT-based distance	about 1 hour 45 minutes
CNN-based distance	about 1 minute
CNN training	about 3 hours 45 minutes (with GPU)



a)



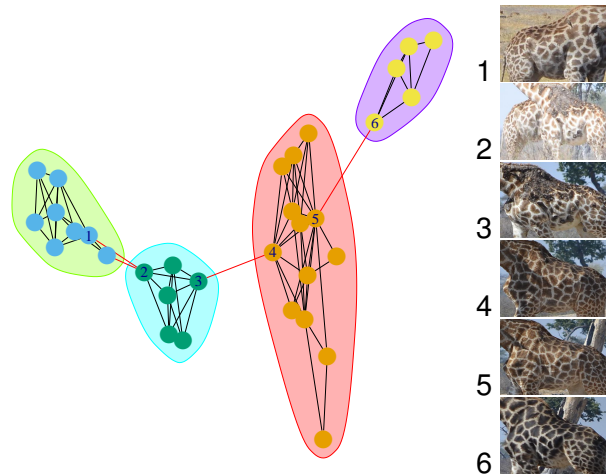
b)

**Fig. 1** Performance of RetinaNet flank detection of giraffes from a set of 3,940 photographs taken at Hwange National Park, Zimbabwe, between 2014 and 2018. In total, we could extract 5,019 images of giraffe flanks automatically. (a) Number of identified flanks per image; (b) Manual classification of cropping problems encountered in 186 images where Retinanet failed to identify a giraffe flank in the photographs.





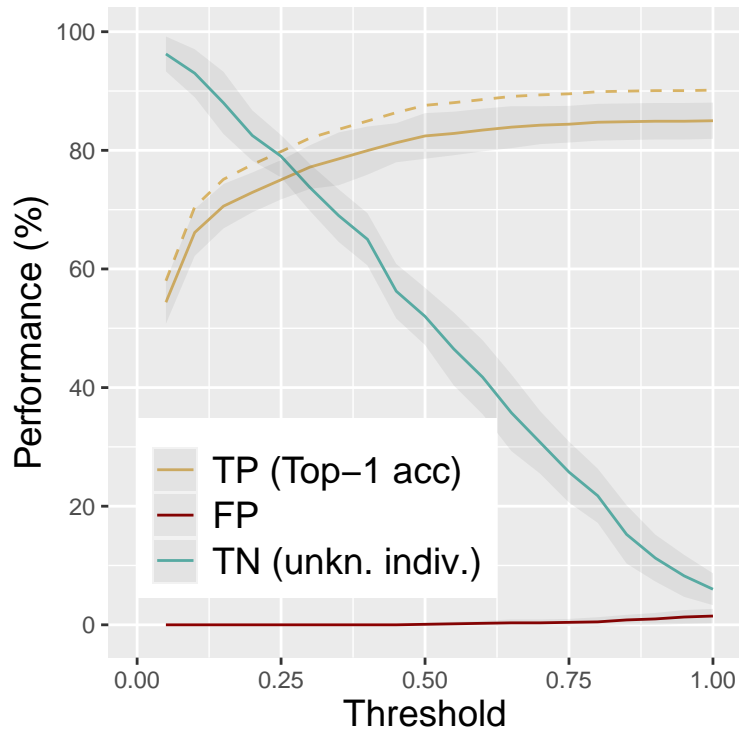
**Fig. 2** Examples of automatic cropping of giraffe photographs with RetinaNet to retrieve the flank of the animal body (red squares). Photographs were shot at Hwange National Park, Zimbabwe, between 2014 and 2018. In (a) the best-case scenario where all giraffes stand separately on the photograph, and RetinaNet successfully finds the flanks of the four individuals; (b) Worst-case, but rare, scenario where the body of the different individuals overlap, combined to a blur caused by the car window on the right-hand side of the photograph. In this case, RetinaNet missed two individuals, and cropped the body of two giraffes into one single image.



**Fig. 3** Example of a connected component split into four clusters using the InfoMap algorithm (see Methods) to assign images of giraffe flank to a given individual for re-identification. Each cluster, representing one individual giraffe, is delineated by an ellipse of different color. Node 2 is an image with two giraffes that we also have in images 1 and 3 respectively, accounting for why their two respective clusters (on the left) are connected. Clusters can sometimes be connected even if the flanks belong to two different giraffes. We illustrate this case with images 3 and 4, which are considered similar because of the presence of the same tree in the background. The same issue arises for images 5 and 6. We applied this method to re-identify giraffes from coat patterns on a collection of photographs taken at Hwange National Park, Zimbabwe, between 2014 and 2018.

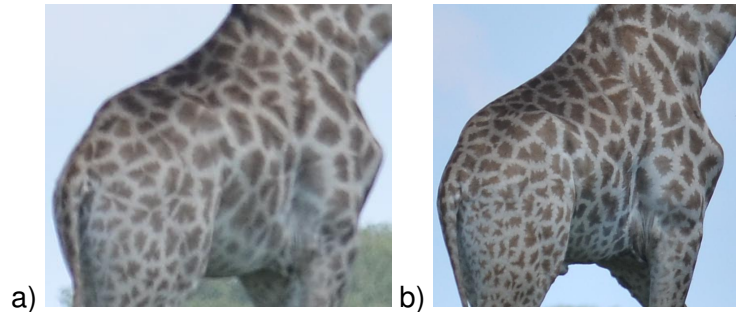


**Fig. 4** Training a convolutional neural network (CNN) requires a large and varied set of images (here giraffe flanks) to achieve reasonable performance when applied on new cases. In this study, we took giraffe photographs at Hwange National Park, Zimbabwe, between 2014 and 2018 but in the field, the opportunity to shoot pictures of the same giraffe in a variety of situations in terms of location or light condition is very limited. Therefore, we performed image data augmentation by randomly changing orientation and size, adding blur, performing edge detection, adding noise and modifying colors or brightness using the `imgaug` Python library (see Methods). Here we show an example of data augmentation, with the original image (left) and four different modified versions used to train our CNN for giraffe re-identification.

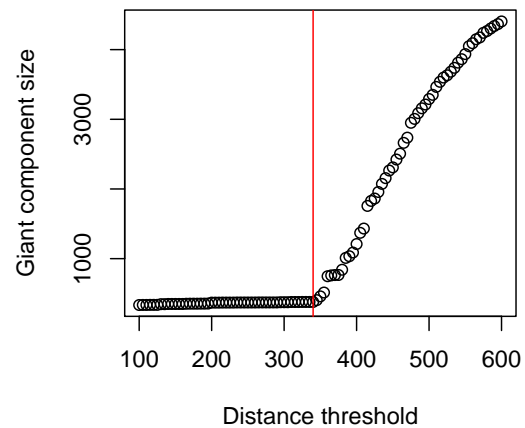


**Fig. 5** Performance of our convolutional neural network (CNN) pipeline for the re-identification of giraffes at Hwange National Park, Zimbabwe (between 2014 and 2018). We decided that two flank images came from the same giraffe using the Euclidean distance between the two images defined by our deep metric learning method. If the distance between the two images felt below a certain threshold distance, it was concluded they belonged to the same individual. Here we report on the true positive rate (TP), or Top-1 accuracy, as function of the distance threshold and calculated on images of know individuals in the test data set, with (plain) or without (dashed) 11 problematic images. Corresponding false positive rate (FP) or Top-1 error. True negative rate (TP) calculated on images of unknown individuals, displays the performance of the CNN model to detect new giraffes entering the data set, that is those individuals never seen before when training the CNN.

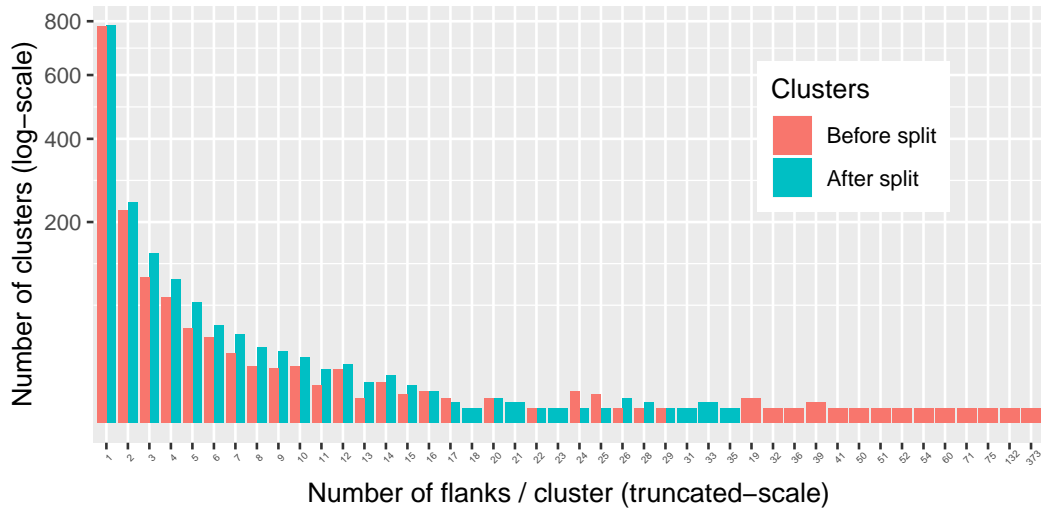
## 1 Supporting information



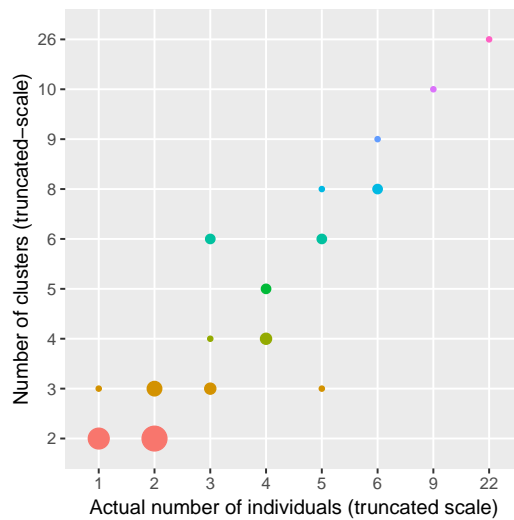
**Fig. S1** Rare SIFT false positive due to perfect shape and orientation matching. Two different giraffes have a similar pose in a) and b) and the SIFT-based distance between the two images is small and below the used threshold.



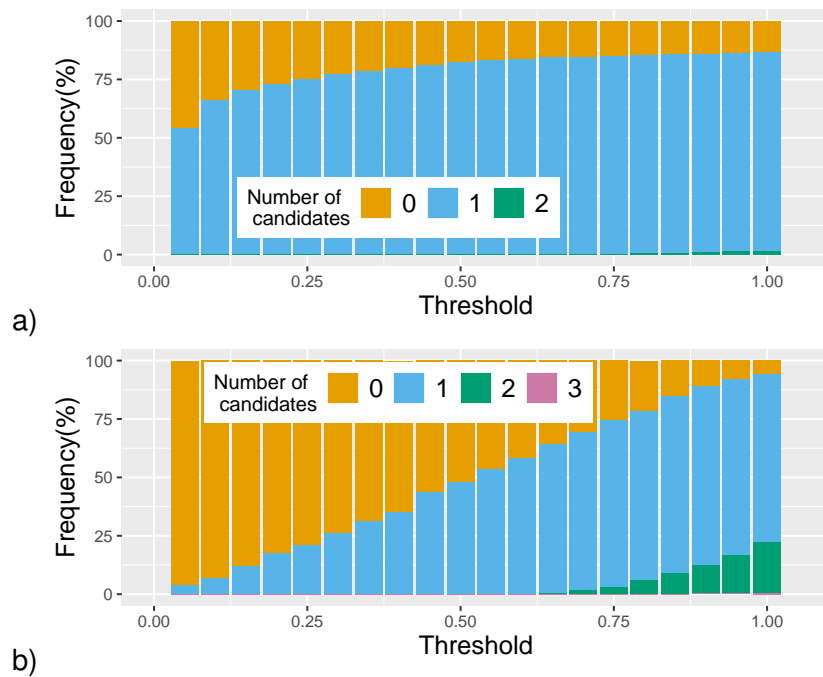
**Fig. S2** Giant component appearance. We manually estimated the threshold value (red line) used to build our image similarity network. The threshold is 340 when dealing with the SIFT-based distance.



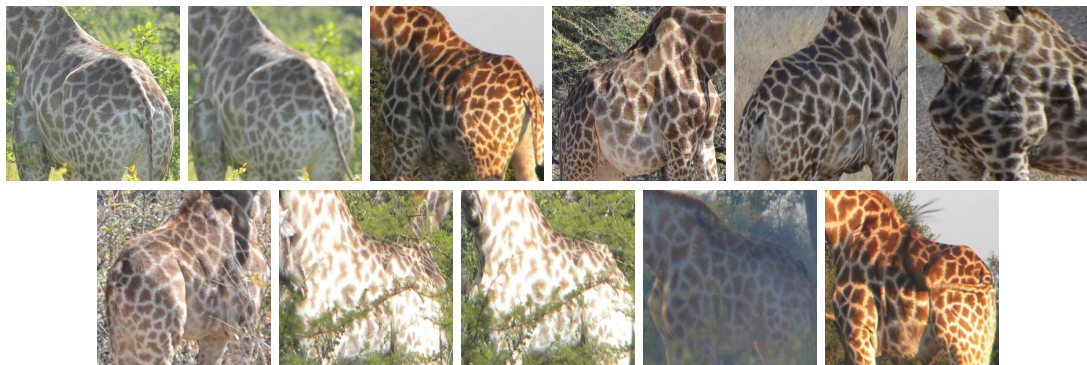
**Fig. S3** Re-identification from 5,019 giraffe flank images. Number of flank images retrieved by clusters, with the original clusters/connected component (red) or with the clusters retrieved using the InfoMap algorithm to split the connected components (blue; see Methods).



**Fig. S4** Agreement between the number of clusters (when at least two clusters were found out of a connected component) as returned by our machine-learning approach, and the human-based and manually-checked number of individuals. Circle size is proportional to the number of observations.



**Fig. S5** Number of giraffe individual candidates at different distance thresholds. a) Known individuals in the test data set. b) Unknown individuals.



**Fig. S6** 11 problematic images out of the test data set, decreasing Top-1 accuracy because of bad orientation (1st row) or element at the forefront (vegetation or giraffe queue; 2nd row).