scientific reports

Check for updates

OPEN

An accurate, efficient, and accessible AI-powered solution for wildlife re-identification in conservation

Shahrzad Gholami[1]✉, Derek E. Lee[2], Caleb Robinson[1], Monica L. Bond[2,3], Rahul Dodhia[1] & Juan M. Lavista Ferres[1]✉

Accurate wildlife re-identification is critical for a wide range of ecological studies, including density estimation via capture–recapture, demographic analyses, and behavioral research. We pr esent GIRAFFE (Generalized Image-based Re-identification using AI for Fauna Feature Extraction), a system for automated re-identification of giraffes with extensibility to other species. Our approach uses local feature matching to identify known individuals and partition unknown individuals for label annotation at scale. Further, we develop a user interface that enables both technical and non-technical users to curate large datasets and analyze repeat survey data. In contrast with existing methods that require manual labeling to facilitate individual re-identification, GIRAFFE automates key steps in the re-identification pipeline, reducing manual effort while maintaining accuracy and interpretability. Validated on real-world giraffe datasets, the system achieves over 0.9 accuracy across nine standard metrics, with some metrics achieving near perfect score. It runs 120 times faster than baseline methods and delivers a 132-fold improvement in cost-effectiveness. This supports endangered species tracking, improves analysis of population dynamics and movement patterns, and, ultimately, allows for the implementation of data-driven conservation strategies.

Keywords Animal re-identification, Computer vision, Masai giraffes, Biodiversity monitoring, AI for conservation

Biodiversity loss is a major environmental problem[1] and has cascading adverse effects on ecosystems[2]. Monitoring wildlife populations is therefore critical for conservation efforts, with wildlife abundance serving as a key parameter for assessing population status and trends. Assessing wildlife abundance and understanding the factors influencing survival, reproduction, and movement are essential for reversing population declines and supporting the recovery of endangered species[3–5].

Various methods have been used to estimate wildlife abundance. Physical tagging for animal identification, such as radio frequency identification (RFID) tags, very high frequency (VHF) radio collars, and global positioning system (GPS) collars provide essential yet invasive tracking capabilities. GPS collars offer precise geospatial data for large-area wildlife monitoring, while RFID and VHF collars are used for close-range, long-term tracking[6,7]. In contrast, photographic individual animal identification based on unique natural markings offers a non-invasive and more feasible method for re-identification of individual animals with distinct physical features[8–10]. By conducting photographic surveys and creating a catalog with unique identification information, researchers can apply well-established capture–recapture methods to estimate abundance and track population trends over time[3–5].

In this study, we focus on Masai giraffes (Giraffa camelopardalis tippelskirchi) for our empirical analysis. The Masai giraffe, native to East Africa, is classified as endangered by the IUCN, with its population declining by 49% in recent decades due to poaching and habitat loss. Currently, an estimated 32,550 individuals remain in the wild[11]. The foundational work of Anne Innis Dagg, who in 1956 became the first scientist to conduct a field study of wild giraffes by manually tracking individuals based on their unique and unchanging spot patterns[12,13], underscores the enduring importance of accurate individual giraffe identification[10].

[1]AI for Good Research Lab, Microsoft, One Microsoft Way, Building 124, Redmond, WA 98052, USA. [2]Wild Nature Institute, Manchester, NH, USA. [3]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland. ✉email: sgholami@microsoft.com; jlavista@microsoft.com

Previous studies explored photo re-identification of various species using pattern recognition techniques, leveraging unique body markings as natural fingerprints. While the re-identification process based on photographic surveys is labor-intensive, artificial intelligence can significantly enhance efficiency by automating image analysis, re-identifying known individuals, and detecting previously unrecorded wildlife after each survey[14]. Besides giraffes[10–17], the AI-driven approaches have been applied to variety of species including killer whales[18], humpback whales[19], white sharks[20], manta ray[21,22], dolphins[23,24], salamanders[8], sea turtles[25], tigers[26], cattle[27], bears[28] and pandas[29], where individual-specific body markings enable reliable identification over time.

There are two primary approaches for encoding images in individual animal re-identification: local feature extraction and deep convolutional neural networks (DCNNs) for generating image embeddings. While computationally intensive, local feature descriptors have demonstrated high discriminative power for species characterized by distinctive textures and coat patterns, using keypoint-based representations for fine-grained matching[30]. Conversely, deep learning-based image embeddings, which encode global structural and semantic information, are well suited for species where contour morphology or facial feature similarity plays a dominant role in identification. Previous studies have explored DCNN-based multi-species models spanning diverse taxa[31,32]. However, further investigation is needed to determine whether grouping species based on shared visual characteristics impacts accuracy across species groups. Alternatively, some approaches partially leverage deep learning by using pre-trained DCNN embeddings to group images[33].

Annotated images are required to train re-identification algorithms; however, annotating datasets for individual re-identification remains a significant challenge in conservation science due to the time, expertise, and cost required. Active learning approaches based on deep learning framework enhances data annotation by selecting the most informative samples for labeling, thereby reducing the need for large labeled datasets[34]. Techniques such as uncertainty sampling[35] and query-by-committee[36] prioritize identifying ambiguous instances that can significantly improve model performance. While active learning strategies have been explored in person re-identification[37] and wildlife species identification tasks[38,39], similar approaches remain underexplored for animal individual re-identification tasks.

Methods that do not rely on deep learning, such as k-means and hierarchical clustering, typically use local features[40]. These methods often employ dimensionality reduction techniques to visualize and cluster images. Although generally less accurate than deep learning-based approaches, these alternatives provide interpretable and computationally efficient options for image clustering. In this study, we leverage a graph-based indexing structure for vector similarity search, along with optimized re-identification parameters, to group animal images with high accuracy and computational efficiency. By using local features, we enable fine-grained image similarity comparisons. The high retrieval efficiency allows for processing this detailed information at scale, thereby enhancing partitioning accuracy for dataset curation tasks.

Re-identification in wildlife datasets is particularly challenging due to highly skewed data distributions—where many individuals are represented by only a single image—and significant variability in image quality caused by factors such as illumination, camera distance, occlusion, and body orientation. Using local feature descriptors is a highly effective approach for Masai giraffes due to their uniquely patterned coat, which provides a rich set of distinctive, high-contrast local features that are well-suited for keypoint-based matching[41]. However, their high computational cost has limited accessibility for ecologists, preventing widespread adoption of solutions that achieve both high accuracy and efficiency in population monitoring efforts.

We present GIRAFFE (Generalized Image-based Re-identification using AI for Fauna Feature Extraction), the first unified, open-source (https://github.com/microsoft/giraffe-identification-ai-tool.) framework for accurate, scalable, and accessible wildlife individual re-identification. Our AI-driven system (Fig. 1), uniquely integrates deep learning–based computer vision models for image preprocessing, vector indexing and retrieval for efficient search, and matching algorithms within an interactive user interface. This end-to-end design

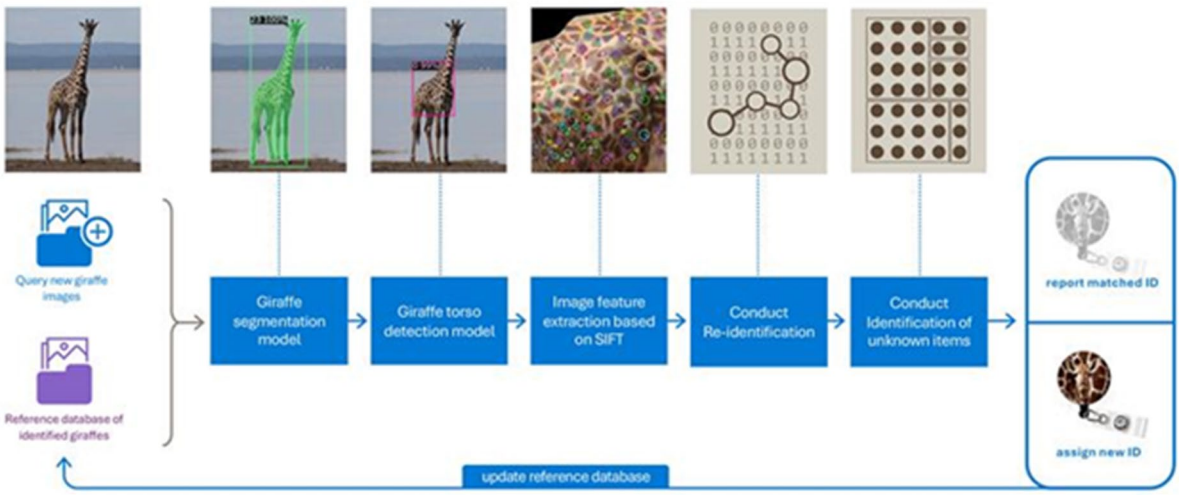

Fig. 1. GIRAFFE: End-to-end workflow for wildlife re-identification and identity cataloging.

supports large-scale visualization, expert-in-the-loop validation, and iterative refinement, enabling practical deployment in real-world ecological monitoring efforts. The hybrid human-in-the-loop approach utilizes the speed and scalability of AI while ensuring the accuracy and interpretability required for ecological research, ultimately facilitating large-scale population monitoring efforts[42,43].

This study presents the most comprehensive parameter optimization analysis conducted to date for giraffe re-identification on bulk survey images. Key contributions include the use of mode-based statistical strategies, the implementation of a distributed indexing and catalog data sharding framework for scalable and accurate automatic matching, and the integration of vector search libraries for efficient nearest-neighbor retrieval. We also introduce a data annotation method that leverages optimized re-identification parameters to improve image grouping and labeling accuracy.

The system consistently achieves over 90% accuracy across seven standard machine learning metrics for both known and unknown individuals, including recall, precision, and F1 score for each category, as well as overall accuracy computed across both groups. Notably, the re-identification accuracy reaches 99%, and the Adjusted Rand Index (ARI) is 0.98, indicating strong clustering agreement with the ground truth annotations for new individuals. Each query is processed in approximately two seconds, enabling a throughput of around 1,800 images per hour against a catalog containing thousands of images. Compared to the naive method, our approach is 120 times faster and yields a 132-fold increase in cost-effectiveness when fully executed on an NC6s v3 machine series, featuring an NVIDIA Tesla V100 GPU with 16 GB VRAM.

Results

We used an annotated dataset comprising 25,363 photographs of Masai giraffes, captured using active encounter surveys in Tanzania between 2012 and 2020 using consumer cameras, and individually identified using human-in-the-loop software that matches large datasets.

of giraffe images with low error rates[14]. This dataset represents 3,987 unique individuals, accounting for over 12% of the global Masai giraffe population in the wild. The dataset was divided into a reference catalog and a query set, ensuring that some individuals in the query set had no overlap with the reference catalog—i.e., no images of these individuals were present in the reference. To systematically evaluate the accuracy of the matching workflow, we designed multiple case studies assessing different retrieval scenarios and identification challenges.

Accuracy: re-identification of known individuals

To evaluate image matching model performance, we frame the task as a binary classification problem, where previously known individuals constitute the positive class and newly identified individuals belong to the negative class. We construct a confusion matrix to capture the model's classification outcomes and compute standard metrics in machine learning, including precision, recall, F1 score, and overall accuracy. We define true positives as cases where the model correctly re-identifies an individual already present in the reference catalog, while false negatives correspond to known individuals misclassified as unknown. Conversely, true negatives represent correctly identified new individuals, and false positives occur when the model incorrectly assigns a known identity to a previously unidentified giraffe.

Matching criteria and selecting optimal cutoff thresholds

We conducted an analysis of accuracy metrics across four dataset variations to assess how changes in image matching criteria and cutoff threshold impact performance as catalog size varies. We evaluated the performance of a fixed query set of 4,666 entries matched against reference catalogs of different sizes. These catalogs included a large set of 15,965 entries (Case Study 1), a mid-sized set of 7,000 entries (Case Study 2), a smaller set of 3,000 entries (Case Study 3), and a minimal set of 1,000 entries (Case Study 4). This approach allowed us to assess the impact of catalog size on the accuracy and trade-offs in matching criteria and thresholds. In this study, we define the cutoff thresholds and the required catalog size for optimal performance as system parameters, which govern the efficiency and effectiveness of the matching process. We analyzed two matching criteria—mode and mean distance—used in aggregating retrieved image descriptors to assess their impact on performance, as shown in Figs. 2 and 3, respectively. We demonstrate that the use of mode statistics for this problem yields over 90% accuracy across all metrics at the intersection point, substantially outperforming the previously explored mean-distance approach[16].

Our results show that applying stricter cutoff thresholds on the most frequently retrieved labels creates a trade-off between key performance metrics. While recall for known items decreases and precision for unknown items declines, recall for unknown items and precision for known items improve. This highlights the balancing act between minimizing errors for known giraffes and reducing false positives for unknown ones. The optimal threshold is determined by the intersection of these performance curves; however, this intersection point varies depending on the data distribution within the reference catalog. To mitigate this variability and enhance retrieval consistency in retrieval accuracy, we propose a distributed indexing and sharding strategy, where the catalog is partitioned into smaller segments and retrieved results are merged. This method ensures a more stable threshold application and improves consistency in performance accuracy for final ranked results. Additionally, our comparative analysis of matching criteria reveals that using mode statistics produces smoother, more consistent, and distinctive trends in trade-off curves, making it a more reliable choice for determining the optimal cutoff threshold.

Applying optimal cutoff thresholds

We present two sets of results. Table 1 evaluates a fixed query set of 4,666 entries matched against catalogs of 7,000, 15,965, and 20,000 entries with and without catalog sharding, while Table 2 examines varying query sizes from more recent surveys by field scientists against the entire annotated catalog as part of a pilot study. Both

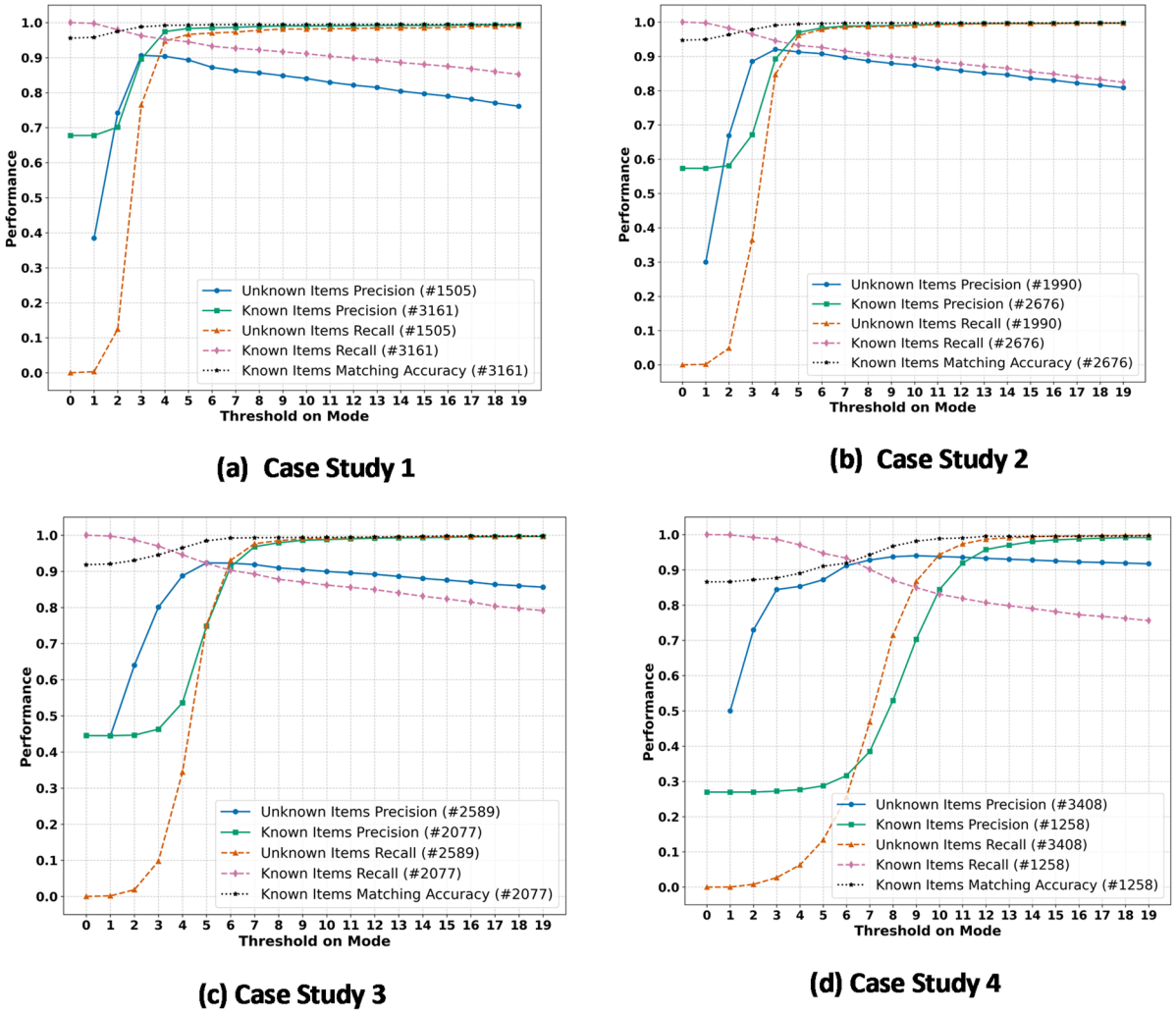

Fig. 2. Performance variation across all metrics for a query set of 4,666 entries using the optimal threshold on mode statistics in different case studies: (1) matching a large reference catalog of 15,965 entries; (2) a mid-sized reference catalog of 7,000 entries; (3) a smaller reference catalog of 3,000 entries; and (4) a minimal reference catalog of 1,000 entries. Numbers in parentheses indicate the number of items.

tables present the number of known and unknown items in each case study. We report accuracy results using an optimal cutoff threshold of 5 based on the mode matching criterion above. To enhance solution adaptability as the reference dataset evolves over time, we implemented a distributed indexing and sharding strategy, with shard sizes between 7,000 and 15,000 chosen based on trade-off curves in Fig. 2(a) and (b) for optimal performance.

As shown in Table 1, with or without sharding, our approach achieves a re-identification accuracy of 99%, demonstrating its effectiveness in correctly assigning identities to known individuals. With a sharding strategy, the overall accuracy of the system is 95%, with an F1 score of 95–96% for known items and 92–94% for unknown items. Recall for known items ranges from 93 to 95%, while recall for unknown items is between 93 and 98%. Precision is maintained at 97–98% for known items and 90–91% for unknown items, highlighting the effectiveness of our method in balancing retrieval performance across both known and unknown individuals.

Table 2 presents a breakdown of the smaller, more recent survey datasets and the corresponding performance results for matching with and without catalog sharding against a big catalog of 25,363 images. Due to the limited size of the query set, we report results in the form of a confusion matrix. Our analysis demonstrates that the proposed approach effectively identifies unknown items while some known items are occasionally missed. The sharding method consistently reduces misclassification errors, particularly by improving known individual re-identification (true positives) and reducing the number of known individuals incorrectly classified as unknown (false negatives), across all survey batches, improving overall retrieval performance. Although we did not have sufficient data to determine the optimal threshold for this catalog size, we applied the 7,000 to 15,000 range identified in our study for the data sharding strategy, which still yielded accurate results.

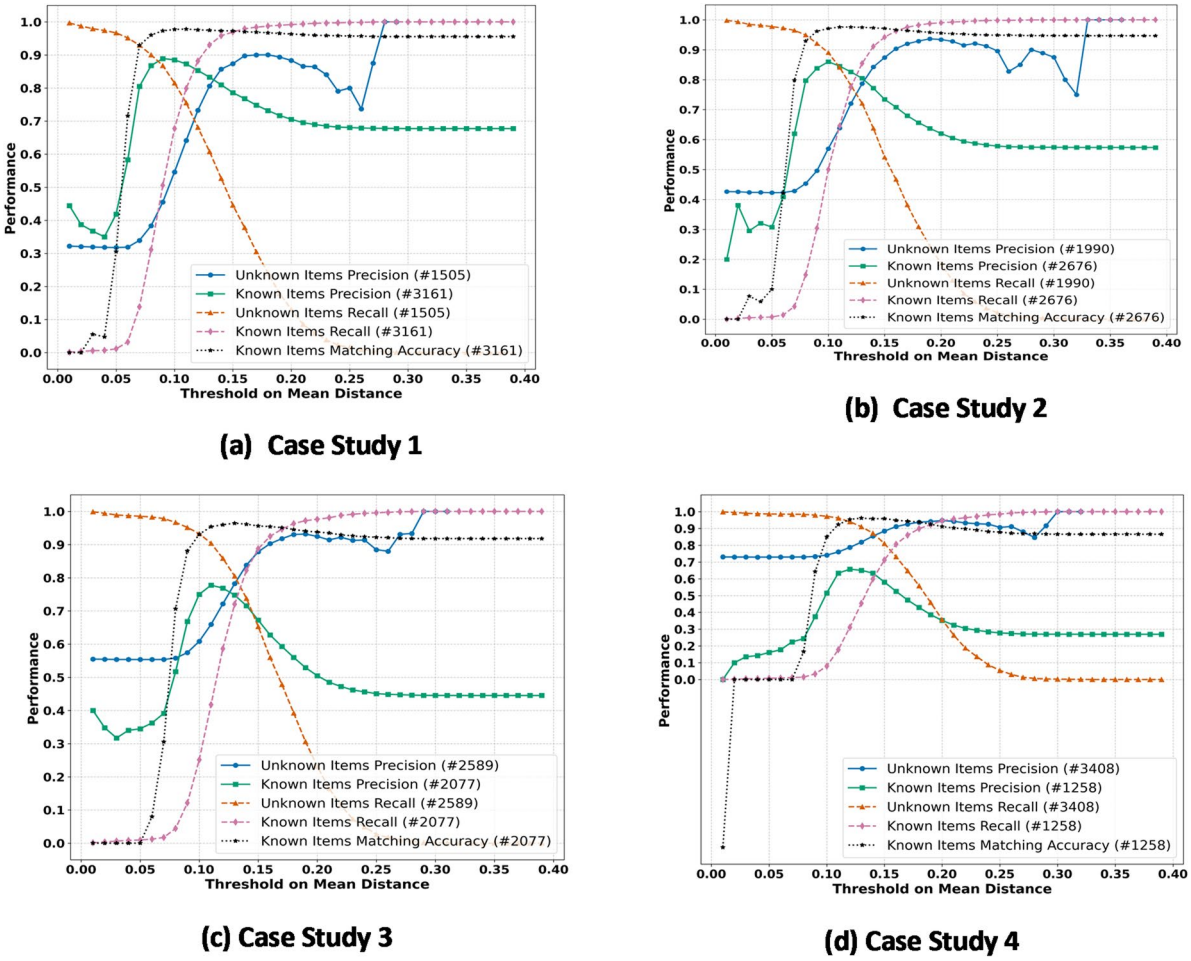

Fig. 3. Performance variation across all metrics for a query set of 4,666 entries using the optimal threshold on mean distance in different case studies: (1) matching a large reference catalog of 15,965 entries; (2) a mid-sized reference catalog of 7,000 entries; (3) a smaller reference catalog of 3,000 entries; and (4) a minimal reference catalog of 1,000 entries. Numbers in parentheses indicate the number of items.

Case	Data Split 1		Data Split 2		Data Split 3	
Reference catalog	20,687		15,965		7000	
Query set	4666		4666		4666	
Unknown items in query set	1470		1505		1990	
Known items in query set	3196		3161		2676	
Sharding	**No**	**Yes**	**No**	**Yes**	**No**	
Overall accuracy	95%	95%	94%	95%	95%	
Accuracy (re-identified items)	99%	99%	99%	99%	100%	
Recall (known)	0.94	0.95	0.93	0.95	0.93	
Precision (known)	0.98	0.97	0.98	0.97	0.98	
F1 score (known)	0.96	0.96	0.96	0.96	0.95	
Recall (unknown)	0.95	0.93	0.97	0.93	0.98	
Precision (unknown)	0.88	0.90	0.87	0.90	0.91	
F1 score (unknown)	0.92	0.92	0.92	0.92	0.94	

Table 1. Re-identification Accuracy: Data split size breakdown and performance results for matching with and without catalog sharding, based on an optimal cutoff threshold of 5 using the mode matching criterion.

Case	Survey 1	Survey 2	Survey 3			
Reference catalog	25,363	25,363	25,363			
Query set	374	36	270			
Unknown items in query set	38	3	16			
Known items in query set	336	33	254			

Sharding	No	Yes	No	Yes	No	Yes
Correctly identified as known	325	331	29	31	240	244
Correctly identified as unknown	38	38	3	3	16	16
Incorrectly identified as known	0	0	0	0	0	0
Incorrectly identified as unknown	11	5	4	2	14	10
Overall accuracy	97%	99%	89%	94%	95%	96%
Accuracy (re-identified items)	97%	99%	88%	94%	94%	96%
Recall (known)	0.97	0.99	0.88	0.94	0.95	0.96
Precision (known)	1.00	1.00	1.00	1.00	1.00	1.00
F1 score (known)	0.98	0.99	0.94	0.97	0.97	0.98
Recall (unknown)	1.00	1.00	1.00	1.00	1.00	1.00
Precision (unknown)	0.78	0.88	0.43	0.60	0.53	0.62
F1 score (unknown)	0.87	0.94	0.60	0.75	0.70	0.76

Table 2. Re-identification Accuracy: Smaller recent survey size breakdown and performance results for matching with and without catalog sharding, based on an optimal cutoff threshold of 6 using the mode matching criterion applied to the entire catalog for a pilot study.

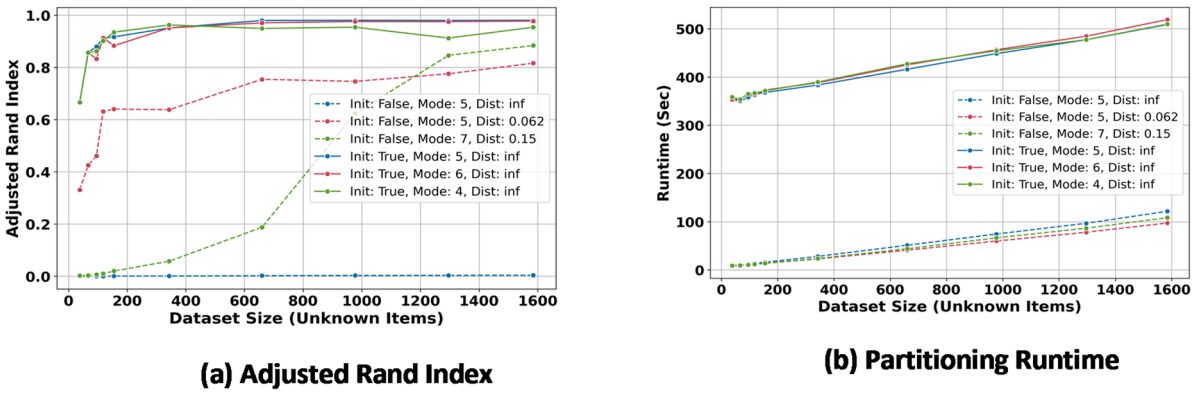

Fig. 4. Partitioning performance across different dataset sizes. Plots show (**a**) Adjusted Rand Index score for partitioning accuracy and (**b**) Partitioning runtime (seconds) as a function of dataset size (number of unknown items). Each line represents a different combination of index initialization choice, mode cutoff threshold, and distance cutoff threshold. In cases where the index was initialized, 7,000 images were used for initialization, and cutoff thresholds were determined based on re-identification.

Accuracy: partitioning of unknown individuals

Beyond classification performance, we assess the effectiveness of the partitioning process for newly identified individuals using the Adjusted Rand Index[44], which quantifies clustering quality by comparing predicted partitions against an ideal ground-truth assignment. The ARI quantifies the similarity between two data clusterings by comparing all pairs of samples and counting those assigned to the same or different clusters in both groupings. Unlike the basic Rand Index, ARI adjusts for the likelihood of random agreement, providing a more accurate measure of clustering similarity. It ranges from − 1 to 1, where 1 indicates perfect agreement, 0 corresponds to random labeling, and negative values suggest less agreement than expected by chance.

Figure 4 presents partitioning performance across different dataset sizes. The plots illustrate (a) ARI scores, reflecting partitioning accuracy, and (b) partitioning runtime (seconds) as a function of dataset size (number of unknown items). Each line corresponds to a different combination of index initialization choice, mode cutoff threshold, and distance cutoff threshold.

For cases where the index was initialized, we used 7,000 images from the reference catalog for initialization, with cutoff thresholds derived from the previous re-identification step. In contrast, for scenarios without index initialization, cutoff thresholds were determined empirically, based on the analysis shown in Supplementary Fig. X4. In this validation experiment, we evaluated the ARI on a dataset sample of 1,578 entries, varying the mode and distance cutoff thresholds. The highest ARI (0.62) was achieved with a distance cutoff of 0.15 and a mode

threshold of 7. These threshold values, along with two additional test cases—one with less strict cutoff thresholds and one based on the re-identification step—were used in our experiments.

Our findings shown in Table 3 suggest that initializing the index with a dataset aligned with the optimized cutoff thresholds consistently yields high ARI (0.98) and more stable ARI values, even for smaller datasets. Using this approach, we achieved an ARI of 0.86 even for a dataset with just 67 images of unknown items. In contrast, starting with an empty index may sometimes yield high ARI values for certain threshold combinations and larger datasets. However, this approach lacks performance consistency, as accuracy can vary significantly across different dataset sizes and distributions requiring more efforts from experts to refine outputs.

This finding is especially important for the annotation of unlabeled data, as it can significantly enhance the dataset curation efforts within the conservation community.

Efficiency: vector indexing and search

Prior to feature extraction, all images undergo preprocessing to retain the most relevant information for re-identification as shown in Fig. 5. We conducted all experiments on an NVIDIA Tesla V100 GPU (Microsoft Azure Virtual Machine NC6s v3 series) with 6 vCPUs, 112 GiB RAM, and 736 GiB of temporary storage. This includes segmentation and object detection, which takes 1.7 s per image, followed by key point and descriptor extraction using SIFT (Scale-Invariant Feature Transform), which requires 0.13 s per image. We configure the feature extraction pipeline to extract up to 1,500 keypoints per image using SIFT, along with their corresponding descriptors. This enables fine-grained representation of distinctive image regions for accurate retrieval and matching. The number of keypoints per image was selected based on empirical experiments conducted during the initial phase of the project, in which we evaluated a range of keypoint counts and assessed their impact on performance. This choice was therefore driven primarily by observed performance rather than by computational constraints.

We employ FAISS (Facebook AI Similarity Search)[45], a high-performance vector indexing and search library, for efficient similarity search and clustering of dense feature vectors. After training on the reference dataset, the index stores 22.5 million feature vectors for a catalog of 15,965 images, enabling rapid and accurate retrieval. While a traditional k-nearest neighbors (k-NN) search requires approximately three minutes to retrieve similar items for each query, using this indexing approach significantly enhances computational efficiency and reduces runtime to nearly two seconds per query from beginning to the end. Table 4 shows breakdown of runtime for all steps of the AI pipeline.

The index training process for the reference dataset is highly efficient, requiring approximately 0.026 s per image. For instance, indexing a catalog of 15,965 images takes approximately 7 min, while a larger dataset of 25,363 images requires around 11 min. However, once trained, the index can be stored and reused, eliminating the need for retraining unless the reference dataset is updated. After this initial overhead, the retrieval process for query images is highly efficient, with re-identification taking only 0.03 s per image.

For previously unknown individuals, a separate index is constructed, initialized with 7,000 reference images for image partitioning. For a dataset of 1,990 images, the partitioning step is completed in 0.28 s per image, enabling efficient processing of new data while maintaining computational scalability. Additionally, updating the reference catalog with newly identified individuals is highly efficient, requiring only 0.0075 s per image, allowing for seamless integration of new data with minimal computational overhead.

Since 2012, approximately 50,000 images of Masai giraffes have been collected through systematic field surveys led by co-authors on this study, providing a rich longitudinal dataset for this individual re-identification study. Traditional SIFT-based solutions with k-nearest neighbor similarity search can require up to four minutes to entirely process a single query image, resulting in substantial computational costs. Using the specified virtual machine configuration, the total cost of computation for such an approach is estimated to be approximately $33,000. In contrast, the AI-driven end-to-end solution proposed in this study significantly reduces computational expenses, requiring only an estimated $250 while maintaining high accuracy and efficiency. Thus, our approach is 120 times faster than the naive method, leading to a 132-fold improvement in cost-effectiveness when executed entirely on an NC6s v3 machine series.

Accessibility: expert-in-the-loop review and refinement via user interface

Our proposed solution is open source and allows technical users to run individual scripts independently, while also providing a comprehensive user interface (UI), implemented using Streamlit (v1.43.0), to ensure accessibility for non-technical users (Fig. 6). The UI enables interaction with each module separately, visualization of matching results, and validation and refinement of outputs prior to updating the database. It includes three distinct visualization checkpoints following the initial matching round for each image batch, enhancing explainability, improving user interaction, and supporting informed decision-making. Execution is simplified through button-based controls that run underlying scripts within a *tmux* session. tmux (Terminal Multiplexer) is a Unix-based tool that allows multiple terminal sessions to run concurrently within a single shell environment. It enables long-running processes to continue executing independently of an active user connection and provides persistent session management. In our implementation, using *tmux* ensures that tasks can run in the background, remain active even if the user disconnects, and automatically capture both standard output and error logs, thereby eliminating the need for direct or continuous terminal interaction.Functioning as a batch query management tool, the interface streamlines the processing of giraffe images following each survey and data collection phase, thereby supporting biodiversity research. This design ensures effective integration of automated processing with expert oversight for efficient and reliable catalog curation.

Case	Data Split 1	Data Split 2	Data Split 3
Reference catalog: used for initialization option	7000	7000	7000
Detected as unknown items in corresponding query set and used in partitioning	1586	1742	2146

Index Initialization	No	Yes	No	Yes	No	Yes
Adjusted Rand Index (partitioning)	0.82	0.98	0.81	0.98	0.83	0.98

Table 3. Breakdown of unknown data, utilized catalog sizes, and performance results for partitioning with and without index initialization. The results are based on an optimal cutoff threshold of 6, applying the mode matching criterion when using index initialization or a two-criteria approach when initialization is not used.

Fig. 5. Image preprocessing steps: the intersection of segmentation and detection outputs are computed, prioritizing the centrally positioned individual with an available identification label.

Review re-identified known individuals

The first checkpoint enables experts to compare each query image with its corresponding matched reference images, displaying the preprocessed image for contextual clarity. Multiple images of the same giraffe may exist in the reference dataset, and the expert can navigate through these images to verify the matches. The interface offers several action buttons, including accepting the re-identified matched label for individual query image results, or accepting all matching results for all queries. Experts can review the top three matching recommendations

Process in workflow	Time per query image (s)
Preprocess and save image using segmentation and object detection models (I)	1.7
Generate image key points and descriptors using SIFT (II)	0.13
Conduct query matching for re-identification using trained index based on 25,363 reference images (IV, V)	0.03
Conduct partitioning for unknown items by training a new index with initialization based on 7,000 reference images for 1,990 query images (VI)	0.28
Update database with existing items (VII)	0.0075

Table 4. Runtime results for the image processing workflow. Step III is not included in the table, as its computational cost is allocated separately as a one-time overhead for the index training process of the reference dataset.

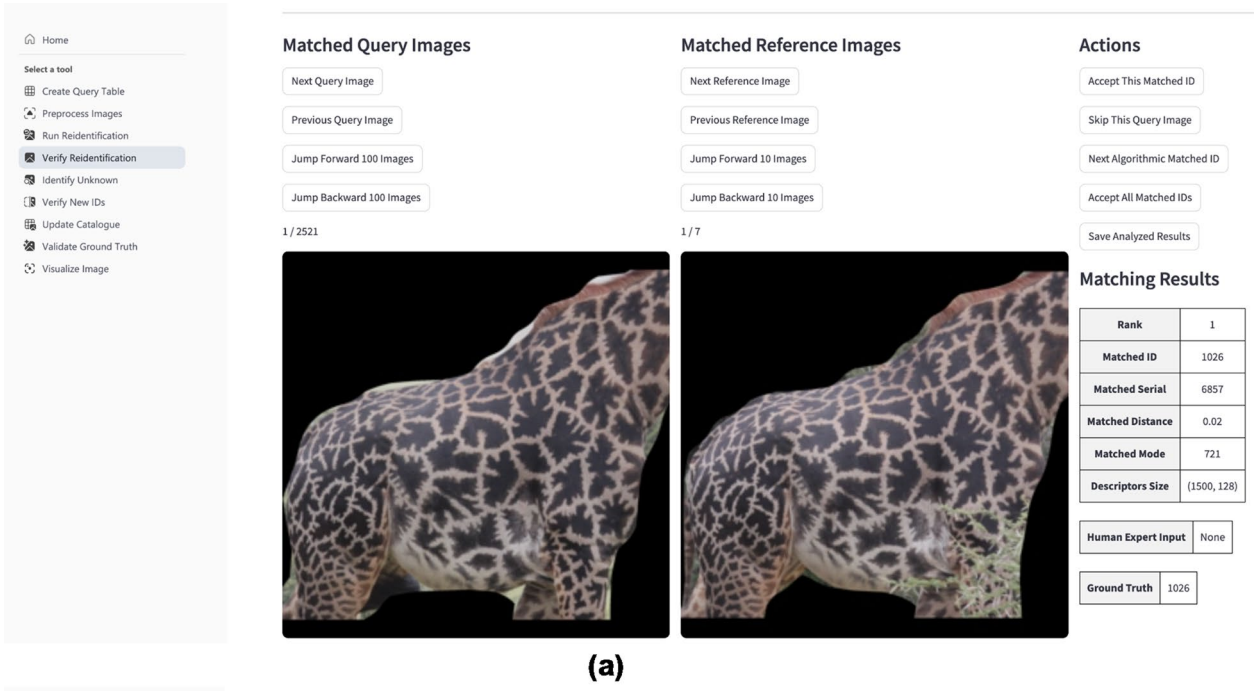

Fig. 6. User interface for expert review and refinement of matching results.

for each query, with detailed matching results displayed in a table for quick reference. Two dedicated buttons provide functionality for experts to skip low-quality images, streamlining the review process, or to reject the automated matching results generated by the AI solution, ensuring high-quality validation and refinement.

Review unknown individuals
At this checkpoint, the expert can review detected unknown individuals by viewing both the preprocessed image and the original query image. Several actions are available, including accepting the results and assigning a new ID, reversing an automatically rejected match by accepting the image retrieved by the AI solution (which may have occurred due to a suboptimal cutoff threshold), skipping the query (e.g., due to poor image quality), or automatically assigning a new identification label to all unmatched items without manual review. This functionality streamlines the labeling process for unknown individuals, enhancing efficiency and ensuring more accurate identifications.

Review partitioning of unknown individuals
After all unknown items are verified, and the partitioning algorithm over those individuals has completed, the UI enables the comparison of all clustered groups side by side. This step allows experts to verify the partitioning results, ensuring that each cluster represents a distinct group of individuals and facilitating accurate labeling.

Functionality of the user interface
The user interface is designed to be intuitive and accessible for non-technical users, enabling them to run code scripts at various stages of the process without needing to understand the underlying Python code. Additionally, the UI incorporates extensive logging capabilities, providing users with detailed records of actions taken and system responses. The interface can be deployed as a web-based application, providing seamless access through a dedicated link, enabling experts to efficiently integrate it into their review workflow.

We used GIRAFFE to match 16,326 new images collected between 2020 and 2025 that were not included in the training or test sets, identifying 421 previously unrecorded animals. Following GIRAFFE-based identification, we evaluated animal-specific accuracy across the full dataset using post hoc software with human operators[14]. This process involved reviewing all images associated with each identified animal to detect erroneous merges (i.e., two distinct animals incorrectly assigned the same identity) and rematching a single image per animal to identify erroneous splits (i.e., one animal assigned multiple identity numbers). The resulting animal-specific error rates were 0.010 for erroneous merges and 0.015 for erroneous splits.

Discussion

Our AI-driven re-identification system provides an accurate and efficient solution for wildlife monitoring, streamlining the process from data collection to re-identification and catalog update. By enabling both technical and non-technical users to independently operate the system, we aim to democratize access to advanced computational tools for biodiversity conservation.

While our system is tested and optimized for Masai giraffes, its methodology is broadly applicable to other species with individually unique body markings. The increasing demand for re-identification algorithms in conservation underscores the potential for expanding this framework to support diverse ecological studies and wildlife management initiatives. The adaptability of our approach makes it a valuable tool for researchers and conservationists working with species that require long-term individual tracking.

The system optimizes multiple parameters to semi-automate what is traditionally a highly manual and time-intensive process, with high accuracy and efficiency. Existing tools for ecologists primarily retrieve visually similar images, requiring users to confirm or reject matches individually. In contrast, our open-source algorithmic approach, coupled with the empirical optimization of system parameters and an integrated user interface, enhances accessibility by providing large-scale visualizations. This enables users to efficiently verify, validate, and refine AI-generated outputs, all without the need for expertise in machine learning or programming. This feature significantly reduces the workload for researchers and conservationists, making large-scale wildlife monitoring more feasible.

Our partitioning algorithms, originally developed for organizing unknown individuals in giraffe populations, can be extended to other species. This is particularly significant for curating datasets without predefined labels, as it enables automated grouping of individuals based on visual similarity. This functionality is crucial for studying populations of endangered species, where labeled data are scarce and costly. It enhances the structure and usability of long-term wildlife monitoring datasets, supporting more effective conservation efforts.

In this study, annotations were restricted to the right-side torso of the giraffe located nearest to the center of the image. This decision was primarily made to ensure consistency in pattern localization. We acknowledge that limiting the dataset to right-flank images may constrain generalizability in settings where only left-flank images are available. However, the proposed pipeline itself is not inherently restricted to right-side data. The key requirement is annotation consistency across the dataset. In principle, the system could be extended to left-flank images by training on consistently annotated left-side data.

An additional enhancement to the system could involve an image quality control module to filter out images with insufficiently visible torsos (or other body regions selected by the user) before entering the AI workflow. Since the visibility of an individual's torso, in the case of our giraffe example, is crucial for accurate feature extraction and matching, excluding low-quality images would improve overall system performance. Implementing such a quality control mechanism would further refine the dataset, ensuring more accurate identification and minimizing potential mismatches caused by incomplete or occluded visual information.

While deep learning models offer superior efficiency, they often lack a sufficient level of interpretability, making it difficult for users to understand how decisions are made. By using local feature extraction methods,

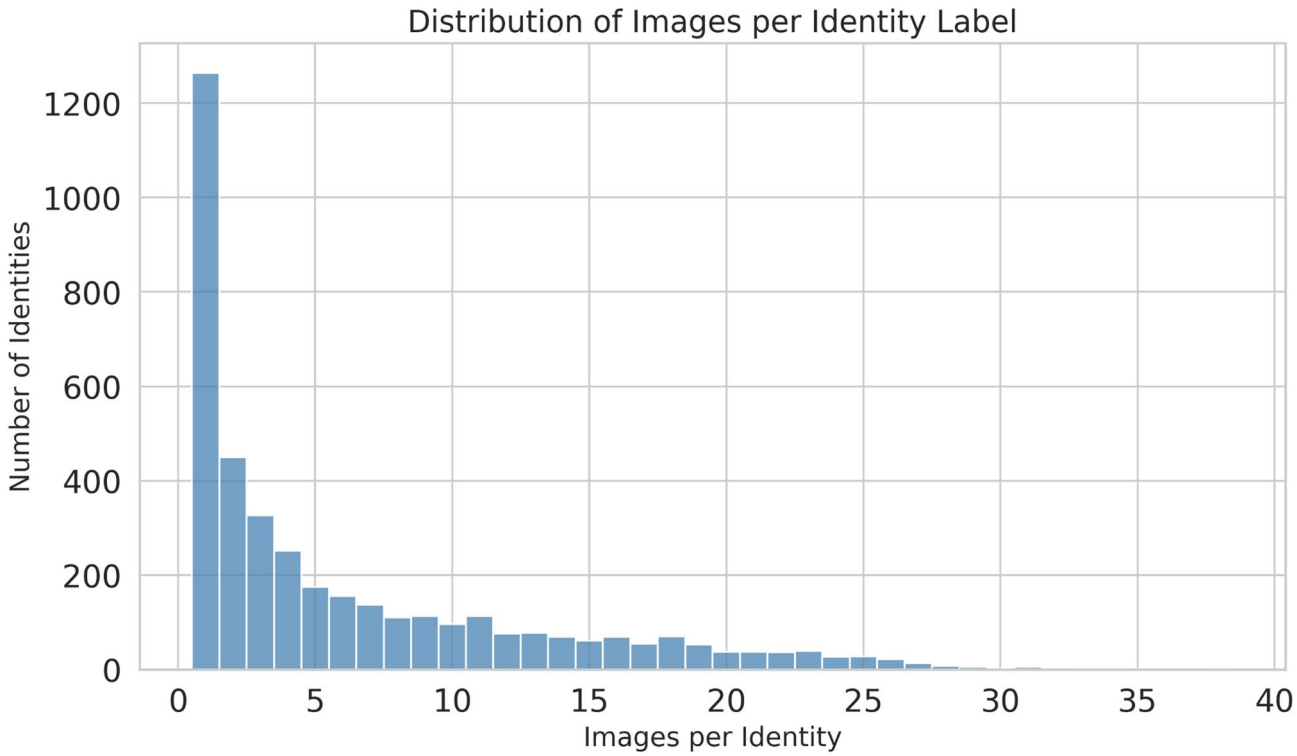

Fig. 7. Distribution of images per identity labels in the dataset.

our approach retains a higher degree of explainability, allowing users to trace how matches are determined. This transparency fosters trust in AI-assisted re-identification and partitioning, enhances user engagement, and supports scalable, long-term biodiversity tracking.

All experiments in this study were conducted on an NVIDIA Tesla V100 GPU equipped with 6 vCPUs, 112 GiB RAM, and 736 GiB of temporary storage. While this configuration enabled efficient training, large-scale inference, and the performance gains reported in this study, we acknowledge that the computational requirements of the proposed GIRAFFE pipeline may limit its accessibility for users without access to comparable high-performance computing resources. The observed runtime and scalability are therefore contingent on the availability of GPU acceleration and sufficient system memory. Although the full source code is publicly available to ensure reproducibility and transparency, users operating on less powerful hardware (e.g., CPU-only systems or GPUs with limited memory) may experience substantially longer processing times. Accordingly, performance metrics reported here should be interpreted within the context of the computational environment used, and expectations should be adjusted when deploying the pipeline on resource-constrained platforms.

Methods

Image dataset

We had access to a dataset of 25,363 Masai giraffe photographs collected in Tanzania between 2012 and 2020, representing 3,987 unique individuals. No filtering criteria were applied, ensuring a diverse range of images. The photographs depict giraffes at various distances and under different orientation and lighting conditions, sometimes surrounded by other species or conspecifics, and in some cases, partially obscured. Annotations were provided for the right-side torso of the giraffe located nearest to the center of the image. To standardize the data, we preprocessed the images to retain only the torso of the central giraffe for analysis. Figure 7 shows the distribution of images for unique giraffe identities. The number of images per identity ranged from 1 to 38. Nearly 31% of the individual giraffes were represented by only a single image.

Workflow overview

The identification approach follows a structured sequence of steps to match a batch of query images against a reference catalog as shown in Table 4. (I) All images, whether designated as query or reference, undergo preprocessing using computer vision algorithms to isolate the giraffe's torso while removing the background, thereby minimizing irrelevant information. (II) The processed images are transformed into feature descriptor vectors using a feature extraction method such as SIFT. (III) A similarity search index is trained on the reference dataset to optimize retrieval efficiency. (IV) Each query image is compared against the reference dataset using a similarity search algorithm to retrieve potential key point matches. (V) Retrieved candidates are assessed against predefined matching criteria and the corresponding cutoff to determine identity confidence, with images meeting the criteria classified as re-identified known individuals. (VI) Images that do not satisfy the matching criteria and the corresponding cutoff threshold are designated as unknown individuals and undergo

a partitioning process based on alternative criteria before being assigned appropriate labels. (VII) Finally, all expert-approved identified individuals, including both known and newly classified, are integrated into the catalog, thereby updating the reference dataset. This ensures continuous improvement and supports future surveys and population monitoring efforts.

Image preprocessing

All images, whether used as query or reference, undergo preprocessing using computer vision algorithms to isolate the giraffe's torso while removing background elements. This reduces irrelevant information and enhances feature extraction for re-identification. Notably, we did not apply any image pruning to filter high-quality images, though incorporating such quality control measures could further improve system performance.

For giraffe segmentation and background removal, we use the Instance Segmentation model, specifically Mask R-CNN with a ResNet50 and Feature Pyramid Network (FPN) backbone, trained on Common Objects in Context (COCO) dataset[46] using the 3x schedule from the Detectron2 library. This model demonstrated strong generalization and out-of-the-box performance for segmenting objects. To further refine the extracted regions and ensure precise isolation of the giraffe's torso, we fine-tune the COCO Detection model from the Detectron2 library, specifically Faster R-CNN with a ResNet101 and FPN backbone, also trained using the 3x schedule, using the Masai giraffe dataset available for this study. This model is optimized for object detection, allowing for more targeted localization of the giraffe's torso.

To ensure accurate identification of a single giraffe per image, we compute the intersection of the segmentation and detection outputs, prioritizing the individual positioned closest to the center of the image for which an identification label is available. This dual-model approach enhances precision in isolating relevant features for re-identification while mitigating background noise.

Feature extraction from images

To extract local features, i.e., key points and descriptors, we employ the Scale-Invariant Feature Transform, a widely used method known for its robustness and effectiveness in handling images captured in challenging field conditions. SIFT is particularly well-suited for wildlife imagery, where variations in lighting, pose, and occlusions are common, and where multiple high-quality images of the same individual may not always be available. By utilizing SIFT, we ensure more fine-grained feature extraction even in cases of limited or suboptimal image quality, facilitating accurate giraffe re-identification.

SIFT detects distinctive key points in an image and describes them using gradient-based local features that remain invariant to scaling, rotation, and moderate affine transformations. The algorithm works by constructing a Gaussian scale-space through progressively smoothing the image with different kernel sizes, identifying key points by detecting extrema in the Difference of Gaussians representation across multiple scales. These key points are then refined by discarding those with low contrast or that are sensitive to noise, such as points located along edges. To achieve rotational invariance, a dominant orientation is assigned to each key point based on the local gradient distribution, and a 128-dimensional vector is generated to summarize gradient magnitudes and orientations within a local neighborhood. This descriptor is robust to variations in illumination and minor geometric transformations, making it highly effective for matching giraffes in wildlife images. We extract up to 1,500 key points for each image.

Vector similarity search

For image similarity search, we employ Facebook AI Similarity Search, an optimized library for high-dimensional nearest neighbor retrieval, to efficiently index and search feature descriptors extracted from giraffe images. FAISS enables scalable storage and retrieval of descriptors, facilitating rapid comparisons between query images and an extensive reference dataset. Specifically, we utilize the IndexHNSWFlat index, which utilizes a hierarchical navigable small-world (HNSW) graph structure, known for its fast approximate nearest neighbor search and scalability[47]. We use the Euclidean distance (L2 norm) to quantify similarity, with lower distances indicating higher correspondence between images.

The FAISS index is structured to support efficient large-scale similarity searches, using data partitioning and optimized search algorithms to accelerate nearest-neighbor retrieval. This is particularly critical for datasets containing thousands of images, where a traditional brute-force search would be computationally prohibitive. Integrating such indexing and search libraries enhances the scalability of re-identification workflows, enabling both effective retrieval of known individuals and partitioning of previously unidentified ones based on feature similarity.

Image matching criteria for re-identification

We facilitate giraffe re-identification by using key points and their corresponding descriptors, comparing the key points of each query image against a reference dataset of identified individuals. Although annotations are provided at the image level, the similarity search is performed at the image descriptor level as demonstrated in Supplementary Figs. X1 and X2. The recognition process selects the nearest neighbor by identifying the descriptor from the reference dataset with the smallest distance to each key point in the query image. The number of key points varies per image, depending on the local features extracted using SIFT. Although various aggregation strategies (e.g., mean distance, mode) can be used for final image matching, our empirical analysis demonstrates that selecting the most frequently matched keypoint annotation—when it surpasses a predefined threshold—achieves the most consistent re-identification accuracy across catalog sizes, particularly in large-scale datasets with thousands of images.

Matching criteria cutoff threshold

We evaluate two statistical aggregation strategies—mean distance and mode—to determine an effective threshold for accepting or rejecting a match based on retrieved descriptor labels. To identify the optimal cutoff threshold, we analyze the trade-off between precision for known and unknown items across four case studies, each with a fixed query set but varying reference catalog sizes, assessing the impact of data distribution on accuracy. This analysis provides insights into how dataset size influences threshold selection, optimizing match acceptance criteria for improved accuracy.

Reference catalog sharding strategy

Our empirical analysis demonstrates that the optimal matching threshold is influenced by the distribution of data within the reference catalog, leading to variability in accuracy as the dataset expands over time. To mitigate this issue, we introduce a sharding strategy for distributed indexing, where the catalog is partitioned into shards with size constraints defined by minimum and maximum bounds. Each shard is queried independently, and the matching scores are aggregated to rank and determine final matches. This approach maintains an optimal threshold per shard, improving scalability and consistency in performance across datasets of varying sizes. The shard size bounds must be empirically optimized based on the characteristics of the dataset under study. Similar sharding-based strategies have been explored in other domains, such as natural language processing, where retrieval-augmented methods use partitioned datastores to improve efficiency and scalability[48].

Partitioning of unknown individuals

For the partitioning of unknown giraffes that were not re-identified in the initial round, we construct a new index and incorporate the unmatched items into it. We then compare key points from each query image of unknown individuals against all items in the index to group images based on similarity, using corresponding optimized parameters obtained during the re-identification process as demonstrated in Supplementary Fig. X3.

We conduct experiments under two different scenarios: (i) The new index is built exclusively from query images of unknown individuals, (ii) The new index is initialized with a subset of the reference catalog. For both cases, when analyzing query images, we retrieve the nearest descriptor from the index while ensuring that it does not correspond to the identical query image. This constraint prevents self-matching, ensuring that each image is compared only to distinct but visually similar items within the index, thereby enabling re-identification for unknown individuals.

In case (i), where the index is built solely from query images, we refine the matching process using a two-stage filtering approach: (1) initial filtering based on key point distances to narrow down potential matches, and (2) secondary filtering based on the frequency of labels for matched key points for all nearest neighbors. In case (ii), where the index is initialized with a subset of the catalog, we utilize the same cutoff thresholds and lower bounds on index size that were introduced and optimized during the initial re-identification step. This strategy enhances the accuracy of partitioning unknown individuals.

This proposed partitioning approach groups images based on similarity and can therefore also be used for data annotation for individual identification.

Data availability

The Masai giraffe dataset used in this study is provided by the Wild Nature Institute and is proprietary, thus restricted from public access. Access to the dataset may be granted upon reasonable request, subject to appropriate data-sharing agreements. For inquiries regarding dataset access, please contact Derek Lee at derek@wildnature-institute.org.

Received: 5 June 2025; Accepted: 18 May 2026
Published online: 30 May 2026

References

1. Ceballos, G., Ehrlich, P. R. & Dirzo, R. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proc. Natl. Acad. Sci.* **114**, E6089–E6096 (2017).
2. Ripple, W. J. et al. Collapse of the world's largest herbivores. *Sci. Adv.* https://doi.org/10.1126/sciadv.1400103 (2015).
3. Andrew Royle, J., Chandler, R. B., Sollmann, R. & Gardner, B. *Spatial Capture-Recapture* (Academic, 2013).
4. Morris, W. F. & Doak, D. F. *Quantitative Conservation Biology: Theory and Practice of Population Viability Analysis* (Sinauer Associates Incorporated, 2002).
5. Nichols, J. D., Capture-Recapture & Models Using marked animals to study population dynamics. *BioScience* **42**, 94–102 (1992).
6. Lahoz-Monfort, J. J. & Magrath, M. J. L. A Comprehensive Overview of Technologies for Species and Habitat Monitoring and Conservation. *Bioscience* **71**, 1038–1062 (2021).
7. Smith, B. P., Hampton, J. O., Waudby, H. P. & Alberthsen, C. *Wildlife Research in Australia: Practical and Applied Methods* (CSIRO PUBLISHING, 2022).
8. Gamble, L., Ravela, S. & McGarigal, K. Multi-scale features for identifying individuals in large biological databases: an application of pattern recognition technology to the marbled salamander Ambystoma opacum. *J. Appl. Ecol.* **45**, 170–180 (2008).
9. Tuia, D. et al. Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **13**, 1–15 (2022).
10. Lee, D. E., Lohay, G. G., Cavener, D. R. & Bond, M. L. Using spot pattern recognition to examine population biology, evolutionary ecology, sociality, and movements of giraffes: a 70-year retrospective. *Mammalian Biology*. **102**, 1055–1071 (2022).
11. IUCN et al. The IUCN Red List of Threatened Species 2018: e.T9194A136266699. *IUCN Red List. Threatened Species IUCN*. https://doi.org/10.2305/iucn.uk.2016-3.rlts.t9194a136266699.en (2016).
12. Innis, A. C. THE BEHAVIOUR OF THE GIRAFFE, GIRAFFA CAMELOPARDALIS, IN THE EASTERN TRANSVAAL. *Proc. Zoological Soc. Lond.* **131**, 245–278 (1958).
13. Foster, J. B. The giraffe of Nairobi National Park: Home range, sex ratios, the herd, and food. *Afr. J. Ecol.* **4**, 139–148 (1966).
14. Bolger, D. T., Morrison, T. A., Vance, B., Lee, D. & Farid, H. A computer-assisted system for photographic mark–recapture analysis: Software for photographic mark-recapture. *Methods Ecol. Evol.* **3**, 813–822 (2012).

15. Halloran, K. M., Murdoch, J. D. & Becker, M. S. Applying computer-aided photo-identification to messy datasets: a case study of Thornicroft's giraffe (Giraffa camelopardalis thornicrofti). *Afr. J. Ecol.* **53**, 147–155 (2015).
16. Miele, V. et al. Revisiting animal photo-identification using deep metric learning and network analysis. *Methods Ecol. Evol.* **12**, 863–873 (2021).
17. Buehler, P., Carroll, B., Bhatia, A., Gupta, V. & Lee, D. E. An automated program to find animals and crop photographs for individual recognition. *Ecol. Inf.* **50**, 191–196 (2019).
18. Bergler, C. et al. FIN-PRINT a fully-automated multi-stage deep-learning-based framework for the individual recognition of killer whales. *Sci. Rep.* **11**, 1–16 (2021).
19. Cheeseman, T. et al. Advanced image recognition: a fully automated, high-accuracy photo-identification matching system for humpback whales. *Mammalian Biology*. **102**, 915–929 (2021).
20. Maroen, M., Rowlinson, M., Matthee, C. & Andreotti, S. The effectiveness of the semi-automated Identifin software to accurately fast-track dorsal fin photographic identifications of sharks. *Aquat. Conservation: Mar. Freshw. Ecosyst.* **34**, e4159 (2024).
21. 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. https://ieeexplore.ieee.org/abstract/document/9647359
22. Kitchen-Wheeler, A. M. Visual identification of individual manta ray (Manta alfredi) in the Maldives Islands, Western Indian Ocean. *Mar. Biol. Res.* https://doi.org/10.1080/17451000903233763 (2010).
23. Bouma, S., Pawley, M. D. M., Hupman, K. & Gilman, A. Individual Common Dolphin Identification Via Metric Embedding Learning. https://ieeexplore.ieee.org/abstract/document/8634778/
24. Patton, P. T. et al. A deep learning approach to photo–identification demonstrates high performance on two dozen cetacean species. *Methods Ecol. Evol.* **14**, 2611–2625 (2023).
25. Papafitsoros, K., Hoh, D., Köhnk, S., Montagna, M. & Witzmann, J. Sea Turtle Photo Identification: A Practical Guide. (2024).
26. Xia, Y., Hu, Y., Liu, W. & Que, T. Research on Re-Identification of Amur Tigers and Bengal Tigers. https://ieeexplore.ieee.org/abstract/document/10649232
27. Qiao, Y. et al. Automated Individual Cattle Identification Using Video Data: A Unified Deep Learning Architecture Approach. *Front. Anim. Sci.* **2**, 759147 (2021).
28. Clapham, M., Miller, E., Nguyen, M. & Van Horn, R. C. Multispecies facial detection for individual identification of wildlife: a case study across ursids. *Mammalian Biology*. **102**, 943–955 (2022).
29. Identification of animal individuals. using deep learning: A case study of giant panda. *Biol. Conserv.* **242**, 108414 (2020).
30. HotSpotter. — Patterned species instance recognition. https://doi.org/10.1109/WACV.2013.6475023
31. 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* 5953–5963 (2024).
32. Otarashvili, L., Subramanian, T., Holmberg, J., Levenson, J. J. & Stewart, C. V. Multispecies Animal Re-ID Using a Large Community-Curated Dataset. (2024).
33. Bodesheim, P. et al. Pre-trained models are not enough: active and lifelong learning is important for long-term visual monitoring of mammals in biodiversity research—Individual identification and attribute prediction with image features from deep neural networks and decoupled decision models applied to elephants and great apes. *Mammalian Biology*. **102**, 875–897 (2022).
34. HuangPo-Yao A Survey of Deep Active Learning. *ACM Comput. Surv. (CSUR)*. https://doi.org/10.1145/3472291 (2021).
35. Ranganathan, H., Venkateswara, H., Chakraborty, S. & Panchanathan, S. Deep active learning for image classification. https://ieeexplore.ieee.org/abstract/document/8297020
36. Deep active learning for object detection. *Inf. Sci.* **579**, 418–433 (2021).
37. Liu, Z., Wang, J., Gong, S., Lu, H. & Tao, D. Deep Reinforcement Active Learning for Human-in-the-Loop Person Re-Identification. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 6122–6131 (2019).
38. Norouzzadeh, M. S. et al. A deep active learning system for species identification and counting in camera trap images. (2019).
39. Automated wildlife image classification. An active learning tool for ecological applications. *Ecol. Inf.* **77**, 102231 (2023).
40. Boom, B. J., Huang, P. X., He, J. & Fisher, R. B. Supporting ground-truth annotation of image datasets using clustering. https://ieeexplore.ieee.org/document/6460437/
41. Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*. **60**, 91–110 (2004).
42. Pollock, L. J. et al. Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nat. Reviews Biodivers.* **1**, 166–182 (2025).
43. ElephantBook: A Semi-Automated Human-in-the-Loop System for Elephant Re-Identification. https://dl.acm.org/doi/https://doi.org/10.1145/3460112.3471947 doi:10.1145/3460112.3471947.
44. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
45. Douze, M. et al. The Faiss library. (2024).
46. Lin, T. Y. et al. Microsoft COCO: Common Objects in Context. (2014).
47. Malkov, Y. A. & Yashunin, D. A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 824–836 (2020).
48. Shao, R. et al. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore. *Adv. Neural. Inf. Process. Syst.* **37**, 91260–91299 (2024).

Author contributions

S.G. conducted all analyses and experiments. C.B. contributed to the analysis focused on efficiency improvements. J.M.F., D.E.L., and S.G. contributed to the conceptualization of the study. M.L.B. and R.D. contributed to the manuscript writing. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-54115-w.

Correspondence and requests for materials should be addressed to S.G. or J.M.L.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026