Animal Re-Identification via Multiview Spatio-Temporal Track Clustering

Ankit K. Upadhyay[1], Ekaterina Nepovinnykh[1,2], S. M. Rayeed[1], Aidan Westphal[1], Lawrence Miao[1], Julian Bain[1], Jaeseok Kang[1] ,Tuomas Eerola[2], Heikki Kälviäinen[2,3], Charles V. Stewart[1]

upadha2, rayees, westpa, miaol2, bainj, kangj8, stewart@rpi.edu ekaterina.nepovinnykh, tuomas.eerola, heikki.kalviainen@lut.fi

[1]Rensselaer Polytechnic Institute, Troy, USA [2]LUT University, Lappeenranta, Finland

[3]Brno University of Technology (BUT), Brno, Czech Republic

Abstract

We introduce a novel framework for individual animal identification (re-id) from long image sequences or video that is robust to occlusions, viewpoint variations, and other challenges of in-the-wild imaging. We start by detecting and tracking animals across images in each sequence. Next, we rate detections according to how much distinguishing information they carry, and coarsely sample the most distinguishable, being careful to sample detections showing the left and the right sides of an animal when possible. For each of these samples we compute embeddings using the MiewID algorithm, and we cluster them in embedding space, which allows us to tentatively group sampled detections showing the same individual. Finally, we link clusters across different viewpoints when taken from the same track. This step improves identification accuracy by merging distinct clusters from complementary viewpoints, ensuring that otherwise disjoint detections are recognized as coming from the same animal. After this computation is complete, we flag genuinely new individuals, and we incorporate consistency constraints to correct errors in clustering through manual intervention. We show strong preliminary experimental results, demonstrating near perfect identification accuracy with very little manual verification.

1. Introduction

This paper addresses the problem of identifying individual animals from one or more long sequences of images or video, which could be captured by a drone, a GoPro or even from a hand-held camera following a herd of animals, assuming they could be tracked from one frame to the next. This variation on the animal re-id problem is important for working in inaccessible areas, for counting animals when no single frame captures all animals of interest, and for behavior monitoring over extended periods of time.

The problem exhibits several distinctive challenges that we address here. First, animals may move in and out of the camera's field of view, requiring previously seen animals to be re-identified and previously unseen animals to be identified as new. Second, an animal in view might still be unidentifiable for significant time intervals if it is too distant from the camera, partially hidden, or oriented in a way that obscures its distinguishable markings. Third, non-overlapping views of an animal may show markings that are completely unrelated to each other, making it difficult to determine when two views indeed show the same animal. Fourth, current animal identification algorithms are imperfect, even for identifiable views, necessitating human-in-the-loop decision-making. Given the volume and velocity of the data, the need for human decisions must be reduced as much as possible.

Our method (as depicted in Figure 1) begins with animal detections and tracks derived from video footage. From each track, we first select a representative set of frames that are that are recognizable and represent the different distinguishable viewpoints. These selected frames are then embedded using a standard re-id algorithm. We initially cluster these embeddings separately for each viewpoint, creating preliminary groups of annotations that likely represent the same animal from a similar viewpoint. The core of our approach then lies in systematically linking these viewpoint-specific clusters across different views by leveraging the spatio-temporal continuity inherent in the tracking data; this step is key to tying together disparate appearances of the same individual. Finally, a series of consistency checks, supplemented by minimal human oversight, are applied to resolve any ambiguities and assign robust individual identities across all observations. We describe algorithms for each step of this computation, some novel and some adapted from other contexts.

Preliminary experiments with a newly curated subset using drone videos from the KABR Grevy's zebra dataset [7] give an initial indication of the efficacy of each algorithmic component and the overall approach. Wherever possi-

ble, we explore alternative approaches to each algorithmic step. Although the results are limited and will be continually augmented, the current experiments show near perfect identification.

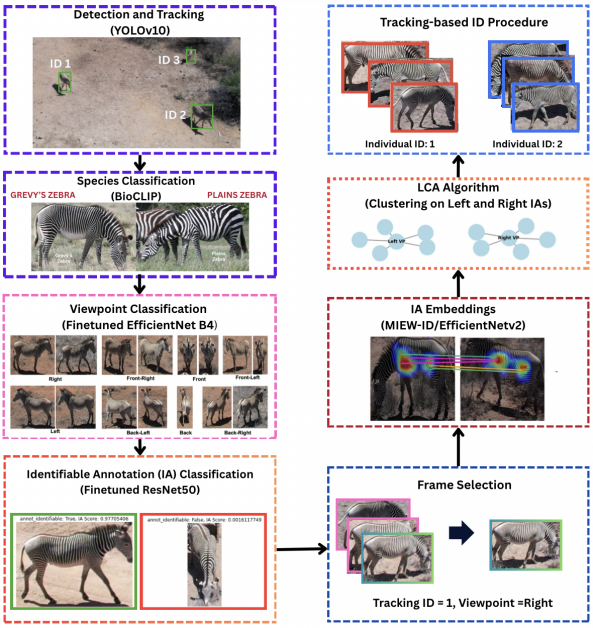

Figure 1. Animal Re-Identification via Multiview Spatio-Temporal Track Clustering Pipeline.

2. Background on Animal Re-ID

Most work on animal re-identification has focused on matching animals detected in a single query image against a database of images and their associated id labels. Early work was based on hand-crafted features [5, 11], but these have been supplanted by learned feature methods, and end-to-end algorithms. Methods have been developed for many species, ranging from salamanders [6, 16] and manta rays [9] up through whales [2] and elephants [8]. Recently, multispecies foundation models have been introduced and made publicly available [4, 12]. Our work here employs MiewID [12], an EfficientNetV2 [18] model trained using ArcFace loss on over 50 species including Grevy's zebra, due to its demonstrated performance. It is important to note that despite algorithmic improvements, human decisions to confirm identities are still needed.

Work on video-based re-id is less widespread but growing. A self-supervised system for identifying Friesian cattle that is almost fully automated is introduced in [1] and [20], applied first for overhead views from drones [1] and then for images taken by three cameras statically-placed in a barn [20]. The distinctive markings on the cattle and along with the controlled farm setting open the door to this high-level of automation. Nepovinnykh et al. [10] showed that re-identification accuracy of Saimaa ringed seals could be improved substantially by aggregating features across image sequences captured by automatic camera traps. Similar ideas for integrating information directly from video clips have been proposed. Work includes published datasets and benchmarks for polar bear re-id from short video clips in zoo environments [21] and for meerkat re-id in wildlife videos [13]. In baseline experiments, these are shown to outperform single frame id algorithms. Methods like these require substantial training data, but if available could be used as a drop-in replacement for MiewID. Our focus here is not on these algorithms per se, but on using them as a component of our system which emphasizes view selection and disparate view combination across longer sequences.

3. Proposed Pipeline

The inputs to the pipeline (Figure 1) are one or more drone video sequences each up to 3 minutes long with about 6K frames. Each can show multiple animals, and these may move in and out of the field of view of the drone's camera. Figure 2 displays example frames from a drone video with increasing timestamp (a-c) showing cases of occlusion, introduction, and disappearance of animals under environmental and viewpoint variations. Our goal is to correctly assign the same id to each instance of an individual in each image in which it appears. Our long-term goal for behavior monitoring is to do this for continuous streaming videos.

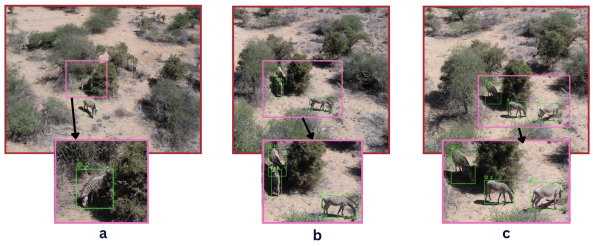

Figure 2. Example frames from drone dataset with increasing timestamp (a-c) showing occlusion (a), introduction of new animal (b), and viewpoint variation (c) of animals.

3.1. Detection and Tracking

We begin by extracting frames from drone videos at 8 fps (downsampled from 30 fps) to reduce redundancy and support effective tracking. Frame-by-frame detection and tracking of target species is performed using YOLOv10 [19]. The tracker produces bounding boxes, object IDs, and tracking IDs across frames. We retain only detections with a confidence score of 60% or higher, leading to slightly over-segmented tracks. The next step is to assign ID labels to each track, with the option to split tracks when necessary.

3.2. Species Classification

For species classification, we use BioClip [14] which is a CLIP-variant vision foundation model specifically designed

for fine-grained biological taxonomic classification tasks. Because our study focuses solely on Grevy's zebra (Equus grevyi), we restrict the model's predictions to that species via its custom classifier.

3.3. Viewpoint Classification

Animal viewpoints are discretized to indicate what aspects of an animal are visible. We start with five cardinal directions — left, right, top, front, and back — and develop a multi-label classifier that is a fine-tuned EfficientNet B4 architecture [17] trained using manually labeled annotations taken from both camera and drone images. Post-processing eliminates rare cases of contradictory labels (e.g. front and back cannot both be assigned to a single annotation), but allows all other combinations. Our classifier achieves an overall accuracy of 93.3% across all viewpoint classes as shown in Figure 9 in the supplementary material.

3.4. Identifiable Annotation (IA) Classification

Among tracked annotations, only some reveal body patterns that reliably identify individual animals. For Grevy's zebras, this identifiable information is found on the left and right sides—not the front, back, or top. Poor image quality or occlusions can prevent identification. To reduce computation, we focus only on annotations deemed "identifiable"—those showing both the hip and chevron on either side of the animal [15].

While self-supervised methods could eventually learn what makes an annotation identifiable, this work uses a supervised approach. We trained a classifier on a manually labeled dataset. Since our classifier was trained on right-side annotations, we horizontally flip left-side annotations before classification. After testing several architectures, ResNet50 yielded the best performance with 90.7% accuracy as depicted in Figure 11 in the supplementary material.

3.5. Frame Selection

Frame selection chooses representative IAs from each track for subsequent matching. This process is principally governed by three parameters: a minimum time interval, T_{seconds}, separating selected frames, and two IA score thresholds, τ_1 and τ_2.

The selection proceeds as follows: Each track is first split into maximal sequences of consecutive IAs. Within every such subsequence, an initial filtering occurs. We retain the IA with the highest score (let this be designated as max_IA for the current subsequence). Additionally, any further local maxima within this subsequence are kept if their score is at least $\tau_1 \cdot$ max_IA and they are at least T_{seconds} frames apart from other selected IAs in that subsequence. This procedure is run independently for each viewpoint (left, right, or both) and functions like non-maximum suppression.

In a second pass, for each *(viewpoint, track)* group, considering the highest score among its currently surviving annotations as max_IA for this step, we keep only those annotations whose score exceeds $\tau_2 \cdot$ max_IA. Tracks that do not contain any valid annotations after these two passes are skipped in the process.

3.6. Identifiable Annotation Embeddings

All filtered IAs from the previous stage are then input to the MiewID [12] embedding algorithm. Annotations of an individual from a specific viewpoint will have embeddings that are close to one another in the embedding space and far apart from embeddings of other individuals. Annotations from the same individual but from different viewpoints will have distinct embedding vectors that are far apart.

3.7. Local Clusters and Alternatives (LCA) Algorithm

We have tracks with IAs and embeddings. Some show both left and right sides; others show only one. Our goal is to cluster tracks by individual, using dual-side tracks to link left-only and right-only tracks. The process has two stages: 1. Cluster left-side and right-side IAs separately, ignoring tracking IDs. 2. Merge clusters using dual-side tracks to link viewpoints and correct inconsistencies. Next, we cluster embeddings so each group represents a unique individual-viewpoint pair. We construct an identification graph where nodes are embeddings and edges reflect similarity. The goal is to form tight, well-separated clusters. A clustering score guides the selection of optimal local groupings.

3.8. Tracking-based ID Procedure

The output clusters from the previous stage are processed using the steps outlined in the procedure below and the final output is the IDs assigned for each track. A detailed description of the procedure can be referred to in Section 11 in the supplementary material.

1. **Consistency Check**
   a) For each track and cluster set (left and right), ensure the track is associated with only *one* cluster per viewpoint and that each cluster set is equivalent to at most one cluster set from the opposite viewpoint.
   b) *If inconsistent, ask human for verification, choosing pairs with highest IA scores.*
2. **Handling Remaining Inconsistencies**
   a) If a track still appears in more than one cluster for a given viewpoint, treat those tracking IDs as *distinct individuals* (i.e. split them).
   b) If one cluster set incorrectly links to multiple cluster sets in the opposite viewpoint, treat that as a *track-linking mistake* and split accordingly.
3. **Time-Overlap Verification (per viewpoint)**
   a) If *no* time overlap exists between cluster pairs, prompt human for pairwise verification. *If they truly belong to the same individual, merge them; otherwise, keep them separate.*

4. **Final ID Assignment**
   a) Across viewpoints, link equivalently matched clusters to the same individual ID. For all other clusters, assign different IDs.

4. Experiments

The experiments were conducted on seven drone videos from **KABR Dataset** [7], each approximately 194 seconds (3.2 minutes) long, with frames extracted at 8 fps from their original 30 fps. For analysis, these videos were grouped into three distinct sets, further detailed in Table 1: **D1** (recorded 10:56-11:03, comprising two videos with 3 zebras), **D2** (13:02-13:08, two videos capturing 5 zebras), and **D3** (11:48-11:58, three videos featuring 3 zebras). This experimental design ensured diverse daylight conditions, with softer morning light for two sets and brighter overhead illumination for the noon set, providing natural variations in lighting and shadows. Figures 6, 7, and 8 in the supplementary material illustrate the viewpoint distribution across annotations.

Table 1. Number of Tracks and IA Statistics for Drone Video Datasets.

Dataset	Number of Tracks			Number of IAs		Avg. IAs per Track	
	Left	Right	Both	Left	Right	Left	Right
D1	12	8	3	46	61	3.83	7.63
D2	11	22	7	24	158	2.18	7.18
D3	13	9	5	55	12	4.23	1.33

4.1. Optimization of Frame Selection Parameters

We conducted a three-dimensional grid search over the minimum time interval T_{seconds} and IA score thresholds τ_1 and τ_2, averaging mean mAP and mean annotation count across three datasets. For each configuration, we applied the frame selection procedure described in Section 3.5; the configuration minimizing annotations while retaining strong re-identification performance is $T_{\text{seconds}} = 15$, $\tau_1 = 0.7$, $\tau_2 = 0.9$, yielding $\text{Avg. \#Annots} \approx 118.7$ and $\text{Avg. mAP} \approx 0.9548$ (Table 2).

Table 2. Optimal values for frame selection parameters.

Objective	T_{seconds} (s)	τ_1	τ_2	Avg. mAP	Avg. # Annots
Maximize mAP	5	0.8	0.7	0.9780	295.7
Balanced trade-off	5	0.7	0.9	0.9552	120.0
Minimize annotations	**15**	**0.7**	**0.9**	**0.9548**	**118.7**

4.2. Evaluation of Tracking-based ID Procedure

We conduct experiments on our three datasets utilizing procedure outlined in Section 3.8 to analyze the correlation between re-identification accuracy and manual decisions. We utilize a dual-threshold approach with lower (0.1) and upper (0.9) bounds, incrementing/decrementing by 0.1 steps. For each pair of clusters, we compute the embedding distance (cosine metric). If the distance is below the lower threshold, clusters auto-merge and if above the upper threshold, they remain separate. Distances between thresholds trigger a manual decision: clusters merge if ground truth IDs match, otherwise they stay distinct. This process iterates until thresholds converge at 0.5, balancing automation and manual intervention for cluster assignment. Accuracy is assessed against ground truth IDs, as shown in Figure 3, demonstrating effective animal re-identification with as little manual intervention as possible.

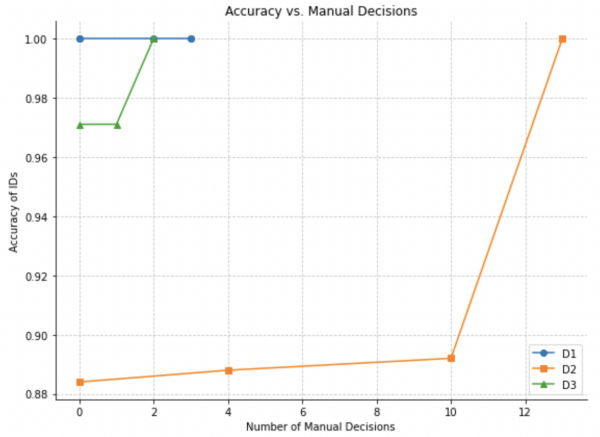

Figure 3. Accuracy of individual IDs vs. manual decisions over all datasets.

5. Ablation Study

5.1. Effect of Viewpoint and IA Classifier

We tested four different settings for `use_viewpoint` *and* `use_IA_score` on each of the three datasets as detailed in Table 3. It can be observed that activating `use_viewpoint` *and* `use_IA_score` (bold rows) is the only setting that matches ground-truth identity cardinality and achieves perfect ID accuracy, precision, and recall, while limiting the workload to $\leq$182 frames and $<$15 manual decisions per dataset. Dropping viewpoint grouping fragments identities (ID-acc.,$\downarrow$ to 0.77–0.97, recall,$\downarrow$ to 0.33–0.92), and omitting IA filtering multiplies the frame budget 4–6× for negligible mAP gains. Hence, both components are jointly necessary and sufficient for a precise and computationally efficient re-identification pipeline.

5.2. Effect of LCA

To evaluate the efficacy of our LCA algorithm, we conducted a controlled comparison with a widely used density-based clustering method, HDBSCAN [3]. Both methods were applied to the same viewpoint-separated, IA-filtered annotation embeddings across datasets D1, D2, and D3. Results are summarized in Table 4.

In terms of overall re-identification accuracy, LCA achieved perfect scores on all metrics—ID accuracy, global precision, recall, and F1—across all datasets. HDBSCAN, while competitive, consistently underperformed relative to LCA: on D1, it produced lower global precision (0.891 vs.

Table 3. Ablation of viewpoint grouping and IA-score filtering. Bold rows mark the only operating mode that (i) preserves correct identity cardinality, (ii) attains perfect frame–level precision, (iii) attains perfect recall, and (iv) minimises annotation load and manual intervention.

Dataset	use_viewpoint	use_IA_score	Frames	$mAP_{@5}$	$\lvert\mathcal{I}_{GT}\rvert$	$\lvert\mathcal{I}_{pred}\rvert$	ID-acc. ↓	Precision	Recall	#Manual
D1	FALSE	TRUE	98	0.949	3	5	0.771	**1.00**	0.332	1
	TRUE	**TRUE**	**107**	**0.953**	3	**3**	**1.00**	**1.00**	**1.00**	**3**
	FALSE	FALSE	516	0.973	3	–	–	–	–	–
	TRUE	FALSE	516	0.971	3	–	–	–	–	–
D2	FALSE	TRUE	161	0.947	5	8	0.930	**1.00**	0.706	18
	TRUE	**TRUE**	**182**	**0.944**	5	**5**	**1.00**	**1.00**	**1.00**	**13**
	FALSE	FALSE	694	0.973	5	–	–	–	–	–
	TRUE	FALSE	693	0.975	5	–	–	–	–	–
D3	FALSE	TRUE	64	0.966	3	4	0.969	**1.00**	0.922	3
	TRUE	**TRUE**	**67**	**0.967**	3	**3**	**1.00**	**1.00**	**1.00**	**2**
	FALSE	FALSE	399	0.982	3	–	–	–	–	–
	TRUE	FALSE	399	0.983	3	–	–	–	–	–

Table 4. Comparison of **HDBSCAN** versus **LCA** clustering on the viewpoint + IA configuration. Global metrics (P_{glob}, R_{glob}, $F1_{glob}$) are computed over the final IDs. LCA achieves perfect identity cardinality and precision/recall on every dataset, whereas HDBSCAN leaves duplicates on D2 and incurs moderate precision loss on D1–D3. M stands for manual decision counts.

Dataset	Clustering	M ↓	$\lvert\mathcal{I}_{GT}\rvert$	$\lvert\mathcal{I}_{pred}\rvert$	ID-acc.	P_{glob}	R_{glob}	$F1_{glob}$
D1	HDBSCAN	6	3	3	0.935	0.891	0.919	0.905
	LCA	3	3	3	**1.000**	**1.000**	**1.000**	**1.000**
D2	HDBSCAN	13	5	8	0.963	0.984	0.854	0.914
	LCA	13	5	5	**1.000**	**1.000**	**1.000**	**1.000**
D3	HDBSCAN	2	3	3	0.955	0.925	0.970	0.947
	LCA	2	3	3	**1.000**	**1.000**	**1.000**	**1.000**

1.000) and F1 (0.905 vs. 1.000); similar gaps were observed for D2 and D3. Manual decision counts were comparable between the methods, suggesting that LCA achieves higher fidelity without additional human effort.

6. Limitations

Track ID Mistakes: As noted earlier, unpredictable occlusions and sudden camera movements (especially in drone footage) can break individual tracks and cause the tracker to assign the same ID to a different animal in close proximity. In our datasets, we observed cases where one tracking ID spanned two distinct individuals (Figure 4). Fortunately, in the final ID assignment stage using procedure in Section 3.8, we automatically detect split clusters (where one ID appears in multiple clusters) and invoke a manual verification step, after which we can reliably identify and then either merge or split clusters to restore correct IDs.

Multiple Animals in One Cluster: Although our pipeline is designed to produce clusters belonging to a single animal, frames containing overlapping individuals can yield ambiguous embeddings that lie midway between two true centers (Figure 5). In such cases, the LCA algorithm may group

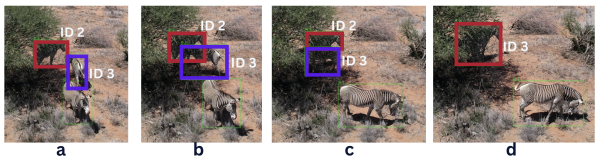

Figure 4. Split cluster case showcasing how tracking ID 2 in frames a-c is depicted as tracking ID 3 in d due to occlusion and overlap.

annotations from different animals into the same cluster, degrading identification accuracy. While we do have consistency checks in place, this overlapping-driven clustering error represents one of the primary limitations in uncontrolled, multi-animal scenes. To mitigate overlapping-animal clusters, we plan to integrate a Census Annotation Region regression step immediately after IA classification to eliminate spurious embeddings and further reduce the need for manual corrections.

Figure 5. Example depicting multiple individual animal in a single cluster due to overlapping individuals in one of the frames.

7. Future Work and Conclusion

We presented a novel framework for animal re-identification via multiview spatio-temporal track clustering that addresses challenges like occlusion and viewpoint variation. By integrating viewpoint-specific clustering, cross-viewpoint linking, and timestamp no-overlap verification, our method achieves near-perfect accuracy with minimal manual intervention. Future work will extend validation to diverse species (Plains Zebra, African Wid Dogs and Giraffes) and environments, exploring the pipeline's scalability and adaptability for wildlife monitoring.

8. Acknowledgments

This work is supported by the Imageomics Institute, the ICICLE Institute, and the Finnish Cultural Foundation. The Imageomics Institute is funded by the US National Science Foundation's Harnessing the Data Revolution (HDR) program under Award # 2118240 (Imageomics: A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning). The ICICLE Institute is a U.S. National Science Foundation funded AI Institute (Intelligent Cyberinfrastructure with Computational Learning in the Environment (ICICLE)) under Award # 2112606.

9. Ethics Statement

The data used in this work was collected in accordance with Kenya's National Commission for Science, Technology & Innovation research license NACOSTI/P/22/18214 and Princeton University's Institutional Animal Care and Use Committee (IACUC) 1835F.

References

[1] William Andrew, Colin Greatwood, and Tilo Burghardt. Aerial animal biometrics: Individual friesian cattle recovery and visual identification via an autonomous uav with onboard deep inference. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 237–243. IEEE, 2019. 2

[2] Drew Blount, Shane Gero, Jon Van Oast, Jason Parham, Colin Kingen, Ben Scheiner, Tanya Stere, Mark Fisher, Gianna Minton, Christin Khan, et al. Flukebook: an open-source ai platform for cetacean photo identification. *Mammalian Biology*, 102(3):1005–1023, 2022. 2

[3] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer, 2013. 4, 1

[4] Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. 2024. 2

[5] J.P. Crall, C.V. Stewart, T.Y. Berger-Wolf, D.I. Rubenstein, and S.R. Sundaresan. Hotspotter - patterned species instance recognition. 2013. 2

[6] Axel Drechsler, Tobias Helling, and Sebastian Steinfartz. Genetic fingerprinting proves cross-correlated automatic photo-identification of individuals as highly efficient in large capture–mark–recapture studies. *Ecology and Evolution*, 5 (1):141–151, 2015. 2

[7] Maksim Kholiavchenko, Jenna Kline, Michelle Ramirez, Sam Stevens, Alec Sheets, Reshma Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, Jackson Miliko, Eduardo Bessa, Isla Duporge, Tanya Berger-Wolf, Daniel Rubenstein, and Charles Stewart. Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 31–40, 2024. 1, 4

[8] Peter Kulits, Jake Wall, Anka Bedetti, Michelle Henley, and Sara Beery. Elephantbook: A semi-automated human-in-the-loop system for elephant re-identification. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 88–98, 2021. 2

[9] Olga Moskvyak, Frederic Maire, Feras Dayoub, Asia O. Armstrong, and Mahsa Baktashmotlagh. Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2021. 2

[10] Ekaterina Nepovinnykh, Antti Vilkman, Tuomas Eerola, and Heikki Kälviäinen. Re-identification of saimaa ringed seals from image sequences. In *Proceedings of Scandinavian Conference on Image Analysis*, pages 111–125. Springer, 2023. 2

[11] Ekaterina Nepovinnykh, Tuomas Eerola, Heikki Kälviäinen, and Ilia Chelak. Norppa: Novel ringed seal re-identification by pelage pattern aggregation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1–10, 2024. 2

[12] Lasha Otarashvili, Tamilselvan Subramanian, Jason Holmberg, JJ Levenson, and Charles V Stewart. Multispecies animal re-id using a large community-curated dataset. *arXiv preprint arXiv:2412.05602*, 2024. 2, 3

[13] Mitchell Rogers, Kobe Knowles, Gaël Gendron, Shahrokh Heidari, David Arturo Soriano Valdez, Mihailo Azhar, Padriac O'Leary, Simon Eyre, Michael Witbrock, and Patrice Delmas. Recurrence over video frames (rovf) for the re-identification of meerkats. *CoRR*, 2024. 2

[14] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 2

[15] Avirath Sundaresan, Jason R Parham, Jonathan Crall, Rosemary Warungu, Timothy Muthami, Margaret Mwangi, Jackson Miliko, Jason Holmberg, Tanya Y Berger-Wolf, Daniel Rubenstein, et al. Adapting the re-id challenge for static sensors. *arXiv preprint arXiv:2412.00290*, 2024. 3

[16] Kosuke Takaya, Yuki Taguchi, and Takeshi Ise. Individual identification of endangered amphibians using deep learning and smartphone images: case study of the japanese giant salamander (andrias japonicus). *Scientific Reports*, 13 (1):16212, 2023. 2

[17] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3

[18] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *Proceedings of International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021. 2

[19] Ao Wang, Hui Chen, Lihao Liu, Kai CHEN, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOv10: Real-time end-to-end object detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[20] Phoenix Yu, Tilo Burghardt, Andrew W Dowsey, and Neill W Campbell. Multicamcows2024–a multi-view image dataset for ai-driven holstein-friesian cattle re-identification on a working farm. *arXiv preprint arXiv:2410.12695*, 2024. 2

[21] Matthias Zuerl, Richard Dirauf, Franz Koeferl, Nils Steinlein, Jonas Sueskind, Dario Zanca, Ingrid Brehm, Lorenzo von Fersen, and Bjoern Eskofier. Polarbearvidid: A video-based re-identification benchmark dataset for polar bears. *Animals*, 13(5):801, 2023. 2