Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spotted: Location-informed Reidentification of Hyenas and Leopards in Camera Trap Surveys

Halil Sina Kelebek

halil@robots.ox.ac.uk

University of Oxford

Julia Hindel

University of Freiburg

Kobus Hoffman

Bubye Valley Conservancy

Lauren Hoffman

Bubye Valley Conservancy

Andrew Loveridge

University of Oxford

Bob Mandinyenya

Gonarezhou National Park

Kudakwashe Ncube

Bubye Valley Conservancy

Justin Seymour-Smith

University of Oxford

Andrea Sibanda

University of Oxford

Abhinav Valada

University of Freiburg

Matthew Wijers

University of Oxford

Daniele De Martini

University of Oxford

Article

Keywords:

Posted Date: July 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9986953/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Spotted: Location-informed Reidentification of Hyenas and Leopards in Camera Trap Surveys

Halil Sina Kelebek[1,*], Julia Hindel[1,2], Kobus Hoffman[5], Lauren Hoffman[5], Andrew Loveridge[3], Bob Mandinyenya[6], Kudakwashe Ncube[5], Justin Seymour-Smith[3], Andrea Sibanda[3], Abhinav Valada[2], Matthew Wijers[3], and Daniele De Martini[1,4]

[1]Oxford Robotics Institute, University of Oxford, Oxford, UK
[2]Department of Computer Science, University of Freiburg, Germany
[3]Wildlife Conservation Research Unit, Department of Biology, University of Oxford, Oxford, UK
[4]Oxford e-Research Centre, University of Oxford, Oxford, UK
[5]Bubye Valley Conservancy, Zimbabwe
[6]Gonarezhou National Park, Chiredzi, Zimbabwe
[*]email: halil@robots.ox.ac.uk

ABSTRACT

Animal re-identification (ReID) in camera-trap surveys remains challenging due to low image quality, strong variation in illumination and viewpoint, and highly imbalanced numbers of observations per individual. As a result, current ReID performance is often insufficient for fully automated use, and practical workflows typically depend on expert review of algorithmically proposed candidate matches. Moreover, most existing approaches focus almost exclusively on visual cues and overlook auxiliary information routinely available in field studies, such as image timestamps and camera-trap locations. We introduce *Spotted*, a location-informed, human-in-the-loop animal ReID framework that integrates visual similarity with spatio-temporal feasibility priors derived from camera locations, thereby reducing the amount of required expert review. Our method (i) computes an image-model-agnostic feasibility score based on the minimum travel speed required for two detections to correspond to the same individual, (ii) uses these feasibility cues as pseudo-supervision to train a lightweight head on top of a frozen visual foundation model, and (iii) fuses adapted visual similarity with spatio-temporal feasibility to obtain a robust pairwise matching score. We additionally integrate an active pair sampling strategy to accelerate annotation by initially prioritizing uncertain predictions. We evaluate *Spotted* on three challenging camera-trap ReID datasets comprised of spotted hyenas and leopards, which we release as part of this work. Our model improves average top-5 identification accuracy by 9pp, 2pp and 9pp over the best baseline on our LeopardID102, SpottedHyenaID109 and SpottedHyenaID415 datasets, respectively. Further, we show that our human-in-the-loop strategy reduces the number of queried comparisons by up to 69pp while achieving equivalent positive matches.

Introduction

Camera traps are a central tool in wildlife ecology, enabling non-invasive monitoring of animal populations in the wild[1]. A challenge in camera-trap-based studies is associating repeated sightings of the same individual and differentiating between separate individuals within a given survey. This task, known as animal re-identification (ReID), remains a challenging computer vision problem, and current automated methods still produce unreliable predictions for many species and survey conditions. Trap-camera imagery is often noisy and exhibits high variation in illumination, animal scale, and viewpoint. Additionally, real-world wildlife monitoring datasets are typically highly imbalanced, with some individuals observed frequently while others are captured only once during the survey period[2,3]. This imbalance results in limited and biased training data, which drastically hinders the performance of purely vision-based models. However, camera-trap surveys possess a structural property: images are captured with a fixed network of camera traps with known spatial locations and timestamps, providing a rich supervisory signal that current state-of-the-art animal reID networks overlook. This metadata encodes implicit constraints on the physical feasibility of the imaged identity. For example, two images captured at distant camera traps within a short time interval imply an implausible travel speed and, consequently, are unlikely to be identity-consistent detections. More importantly, repeated captures at spatially close camera traps over extended time periods reflect stable movement patterns of individual animals. Our ecological motivation for spatial priors is further grounded in the concept of home range[4]: individuals of many species tend to remain within spatially bounded regions over extended periods, making spatial proximity an informative prior for identity. Consequently, these limitations motivate a tailored solution that leverages rich spatial priors beyond a simple binary feasibility

check.

Wildlife ecology demands flawless animal re-identification to reliably extract meaningful animal movement patterns. Fully automatic identification typically cannot meet this requirement and current workflows rely on human verification of algorithm-proposed image matches for the same or different individual. To avoid an exhaustive annotation effort over a large set of candidate matches[1], we motivate an active pair sampling strategy that initially prioritises uncertain candidate matches for human verification, reducing the total annotation burden while maximising the discovery of true identity matches. At an annotation rate of 5.78 queries per minute, as measured from an expert annotator in our study, the choice of pair selection strategy has a direct and measurable impact on total labeling time.

In this work, we propose *Spotted*, a location-informed, human-in-the-loop animal re-identification framework that integrates visual similarity with spatio-temporal feasibility priors derived from the camera location and timestamp metadata. Rather than treating location priors only as a hard constraint or post-hoc filter, we incorporate them as a continuous signal that shapes both representation learning and pairwise similarity scoring. Our approach consists of three components. First, we introduce a vision-agnostic, pairwise speed-based feasibility scoring that encodes the minimum plausible speed for two detections to correspond to the same individual. Second, we train a lightweight MLP head on top a frozen animal ReID backbone using location-derived pseudo-labels to suppress physically implausible image matches without requiring identity supervision. Third, we fuse the adapted visual similarities with the spatio-temporal feasibility scores and integrate an active pair-sampling strategy to improve retrieval accuracy while considerably reducing the number of pairwise comparisons required in a human-in-the-loop workflow. To evaluate our work, we release three camera-trap ReID datasets (LeopardID102, SpottedHyenaID109, and SpottedHyenaID415) with this work, comprising over $3,200$ camera images taken of 620 individual animals at 230 locations collected over a 2-month survey period. Our proposed method demonstrates consistent improvements over SOTA animal re-identification baselines.

Related Work

Animal Re-Identification

Seminal works in animal ReID such as HotSpotter[5] rely on handcrafted local feature descriptors to perform instance recognition by matching sparse SIFT-based descriptors[6] and ranking candidate matches based on geometric consistency. Despite its advantages in low data requirements and interpretability, this approach exhibits poor robustness to severe appearance and viewpoint variations. With the successes of Deep Learning techniques over the past decade, the field has shifted towards metric learning approaches that directly map images into an embedding space where intra-identity distances are minimized and inter-identity distances are maximized, resulting in embedding vectors corresponding to the same individual being mapped in proximity[7–9]. Although many previous works represent individuals with a single global feature vector extracted by passing an image in full through a backbone network[10,11], selected works apply local, part-aware approaches to improve robustness across various viewpoints by focusing on local regions or salient parts[8,12–14]. The hybrid method proposed by Shukla et al.[15] applies a finetuned DenseNet[16] to produce an initial set of pairwise matches, followed by a SIFT-based re-ranking approach to combat poor image quality and limited training data.

Another trend is the emergence of multispecies pretraining and foundation models for animal ReID, which aim to produce transferable embeddings across species and datasets. MegaDescriptor[11] trains a SWIN backbone[17] with an ArcFace loss[18] for animal ReID, providing a competitive baseline. Similarly, the multispecies identification model, MiewID[10] is trained with the ArcFace loss but rather uses the Convolutional Neural Network (CNN) based EfficientNetV2[19] architecture as their model backbone and makes use of a larger, community-curated dataset for model training.

Spatio-Temporal Auxiliary Supervision

The importance of meta-data for Animal ReID is increasingly explored. A recent animal ReID dataset introduces meta-features such as temperature, circadian rhythm, and orientation[20]. Alongside the dataset, the authors propose the Meta-Feature Adapter[20], which incorporates metadata into their ReID model as textual input via image and text encoders. Camera traps provide a fixed geometry for the sensor network with known relative positions and timestamps, imposing strong constraints on whether two detections of the same individual are physically plausible. A previous camera trap study with dense camera placement show that timestamps and camera identities can be used directly to reconstruct animal movement paths across a survey area[21], demonstrating that the spatial sensor structure contains rich information about individual movement. Spatio-temporal auxiliary data have been previously explored in tangential domains like fine grained classification problems such as species classification[22] A previous work has explored spatio-temporal priors in a closed set animal ReID setting[23], however requiring pre-labelled class-level training data for each individual due to the closed-set ReID setting, making it unsuitable for the task of unlabelled dataset annotation. While temporal meta-data is often available[3], many publicly available animal ReID datasets do not include fine-grained geo-spatial coordinates [3,24–26].

Active Pair Sampling

Fully automated animal ReID does not achieve flawless accuracy and is therefore often combined with human verification of algorithmically proposed candidate matches[27,28]. This workflow motivates approaches that minimize human annotation needs while optimizing correctly identified individuals. Active pair sampling has been explored in animal ReID[29] and species identification[30] as well as person ReID[31,32], however, with a focus on reducing the training set required for model training rather than to speed up complete dataset annotation[30–32]. Many active learning methods propose clustering approaches to discover informative samples for labeling, with specific strategies varying in how cluster structure is exploited: previous works have explored sampling strategies focusing on on cluster boundaries[31], distance to core points[33,34], or cluster mismatch between different clustering algorithms[29]. SPAL[31] proposes an uncertainty-driven strategy that clusters samples and queries pairs near cluster boundaries to efficiently improve representation quality with limited labels. DRAL[32] proposes a reinforcement learning based approach to propose uncertain pairs for labeling. Continual learning approaches have also been proposed to continuously update animal species classification models[35] and animal ReID models[29,36] using expert annotations for selected images, thereby enabling incremental performance improvements in model accuracy as new data is collected. Active selection strategies have also been studied more broadly in the animal ReID domain for population size estimation under limited supervision[37]. Across these works, however, the common objective is to reduce the number of labels needed to train or estimate a model or population statistic, rather than to minimise the annotation effort required to fully label a dataset. The problem of accelerating complete dataset annotation by reducing redundant pairwise comparisons for a human expert remains comparatively under explored.

Data

As part of this work, we release three challenging animal ReID datasets from camera trap surveys conducted in Zimbabwe: two contain spotted hyenas (SpottedHyenaID415 and SpottedHyenaID109), and one includes Leopards (LeopardID102). The images were obtained from a set of paired static camera traps distributed throughout survey sites. The images vary considerably in quality, time of day, viewpoint of the animal, and are highly unbalanced in terms of the number of images per individual. The timestamps and relative locations between the cameras are known. The SpottedHyenaID415 and LeopardID102 datasets originated from the same survey and contain all images of each species captured during the survey period. The SpottedHyenaID109 dataset is from a separate survey and comprises a subset of identity folders, as the full dataset contained more than 800 IDs with considerable uncertainty in label completeness (eg. potential missed matches). We therefore manually selected a subset of IDs for which we were confident in the individual classifications. The SpottedHyenaID109 dataset also contains hand-labelled binary orientation tags, indicating the visible side of the animal (left/right). All three datasets were collected directly by the authors of this work with permission from the landowners. Surveys were conducted by the property management organisations in partnership with the Wildlife Conservation Research Unit. No wildlife capture, handling, or invasive procedures were undertaken, and no permits or ethical approvals were required for the collection of these camera-trap images. The released location data was anonymized by converting GPS coordinates into relative camera distances and removing image regions containing location-identifying features such as text. These privacy measures ensure the safety of local wildlife and nearby communities. We show dataset examples and statistics in Fig. 2 and Table 1.

Table 1. Statistics of our animal ReID datasets. Labelled images refer to images with expert-identified animal id.

Dataset	Species	# Individuals	# Labeled Images	# of Cameras
LeopardID102	Leopard	102	717	105
SpottedHyenaID109	Spotted Hyena	109	704	125
SpottedHyenaID415	Spotted Hyena	415	1871	105

Methods

Preprocessing

We apply histogram equalization and histogram matching as preprocessing steps to reduce the major variations in image color and intensity. These image transforms minimize the impact of low-level lighting variations on the downstream ReID task. Thus, we visually select a reference image per dataset for histogram matching. Next, we use grounded SAM2[38] to mask and crop the individual animals as shown in Fig. 2.

Spatio-Temporal Model

The spatio-temporal feasibility score is based on the straight-line speed between two timestamped detections taken at known camera locations. This straight-line speed represents the minimum travel speed required for two detections to correspond to the

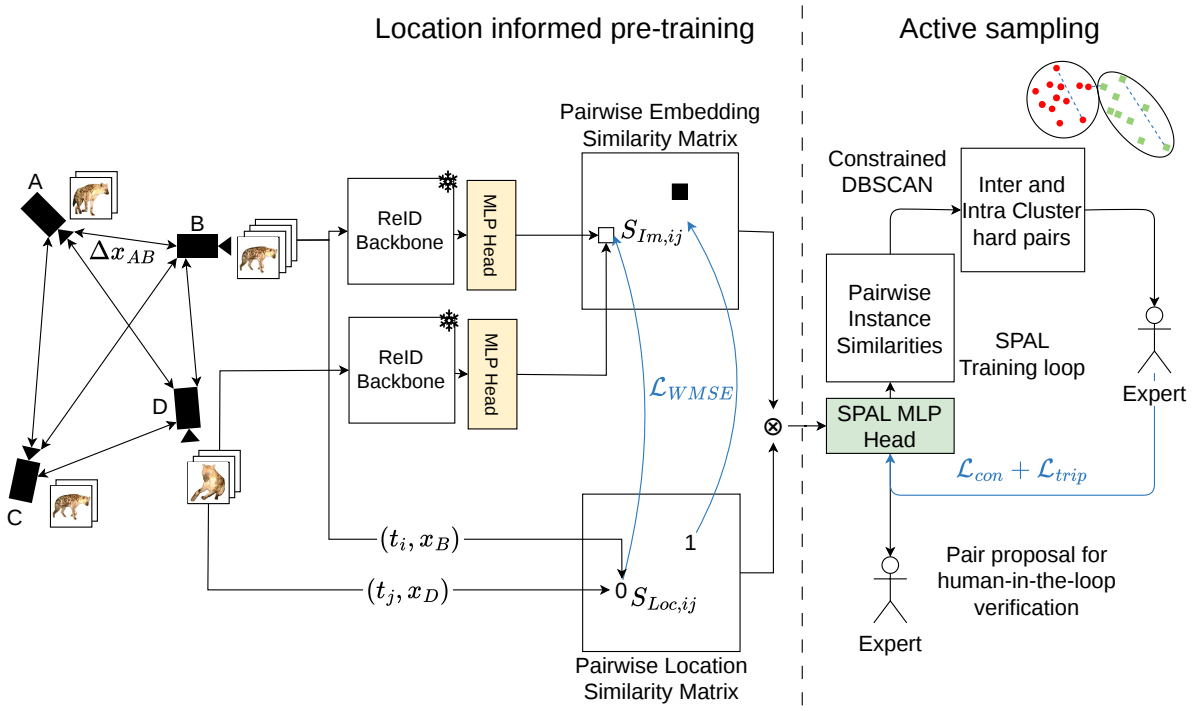


Figure 1. Overview of the proposed *Spotted* framework. Camera-trap images are encoded using a frozen ReID backbone and refined via a network head trained on location-derived pseudo-labels. The adapted visual similarities are fused with spatio-temporal feasibility scores and used within an active sampling loop for efficient human-in-the-loop labeling.

same individual. As such, an excessively high minimum speed indicates the spatio-temporal plausibility of the association. For each image pair, the euclidean distance between their respective camera trap location $\mathbf{x_i}$ and $\mathbf{x_j}$ and the time difference between their timestamps t_i and t_j, are computed and normalised to a range between 0 and 1 ($\Delta\tilde{x}$ and $\Delta\tilde{t}$) as detailed in equations (1) and (2). This normalization ensures scale invariance across survey sites and reduces the scoring function to a single tunable hyperparameter s with exponential decay applied to the normalized pairwise speeds (Eq. 3).

$$\Delta x_{ij} = \left\|\mathbf{x_i} - \mathbf{x_j}\right\|_2, \quad \Delta\tilde{x}_{ij} = \frac{\Delta x_{ij} - \Delta x_{min}}{\Delta x_{max} - \Delta x_{min}} \tag{1}$$

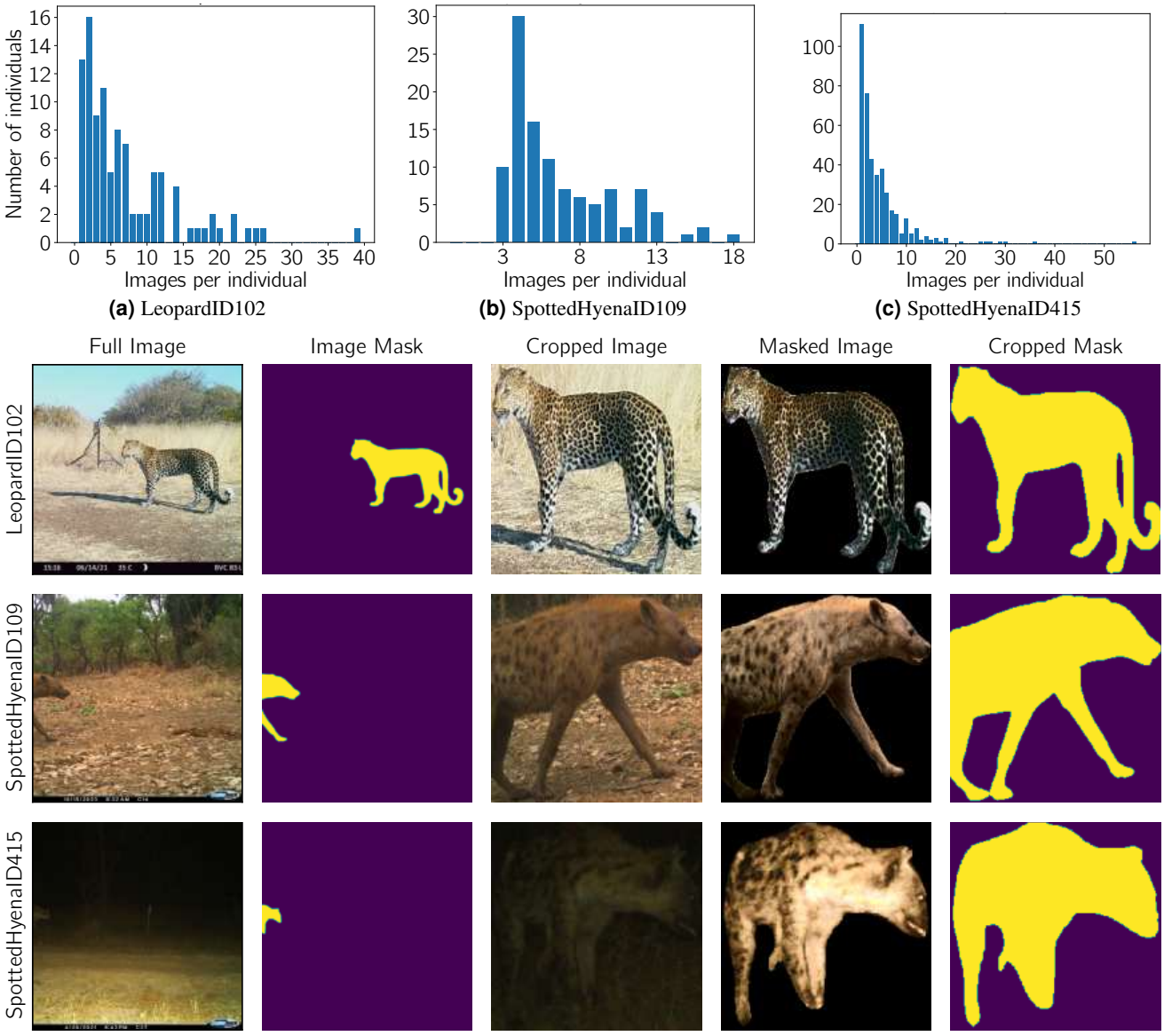
$$\Delta t_{ij} = |t_i - t_j|, \quad \Delta\tilde{t}_{ij} = \frac{\Delta t_{ij} - \Delta t_{min}}{\Delta t_{max} - \Delta t_{min}} \tag{2}$$

$$S_{loc}(i,j) = \exp\left(-s\frac{\Delta\tilde{x}_{ij}}{\Delta\tilde{t}_{ij}}\right) \tag{3}$$

This distribution biases towards higher match feasibility towards spatially close detections, even when the temporal separation $\Delta\tilde{t}$ is large, as seen in Fig. 3. This bias is desirable because it implicitly encodes the notion of home range which describes the phenomena that individual animals tend to remain within a limited spatial region even over extended periods[4]. The strength of this spatial bias is controlled by the parameter s, which governs the rate at which the score decays with increasing speed. The parameter s was set to 1 for the survey sites following a grid search and held fixed across experiments. A spatio-temporal score of zero indicates a strong negative constraint, corresponding to image pairs that are highly unlikely to depict the same individual, e.g., images are captured at similar times but at distant locations. These trends can be seen in Fig. 3 when $\Delta\tilde{t}$ is near zero. In contrast, matching scores close to one indicate proximity and act as a weak positive signal. We define these pairs as *weak* because intra-identity correspondence is not guaranteed. Such pairs may correspond to either the same individual or to different individuals within the same vicinity, including pairs within the same social pack.

Image Model

We make use of a frozen pre-trained animal Re-ID model as our visual backbone. To emphasise the image-model agnostic nature of our approach, we test with both the MiewID[10] and MegaDescriptor[11] multi-species ReID models in *Spotted*. On top



(a) LeopardID102 **(b)** SpottedHyenaID109 **(c)** SpottedHyenaID415

(d) Sample images from each dataset. The impact of the preprocessing steps is visualized along row. The final preprocessed image is provided in the right-most column.

Figure 2. Quantitative and qualitative overview of our three animal ReID datasets. (a–c) Sighting frequency distributions for LeopardID102, SpottedHyenaID109, and SpottedHyenaID415 datasets, highlighting the long-tailed imbalance in observations per individual. (d) Sample images illustrating our preprocessing pipeline for each dataset.

of this backbone, we train a lightweight MLP f_θ consisting of a down-projection to a lower-dimensional space followed by an up-projection back to the original embedding dimension. The head is implemented using two fully connected layers with a low-rank bottleneck and a residual connection. The residual connection adds the head's output to the original embedding, allowing the model to learn incremental adjustments to the pre-trained features. We select this lightweight design as it reduces the number of trainable parameters and helps mitigate overfitting in the presence of limited and highly imbalanced training data.

We derive the training supervision from the computed spatio-temporal feasibility score as explained in subsection *Spatio-Temporal Model*. The set of all image pairs with high feasibility scores ($S_{loc}(i,j) = 1$) are treated as weak positive samples ($\mathscr{P}^+$), while the set of all image pairs with low feasibility scores ($S_{loc}(i,j) = 0$) are treated as confident negative samples ($\mathscr{P}^-$). During training, we construct a pairwise similarity matrix $S_{im}(i,j)$ from the MLP head embeddings using cosine similarity and clip negative values. We apply a weighted L2 loss equation (4) to align the image embedding similarities with the spatio-temporal feasibility scores for the selected pairs (all image pairs with $S_{loc} = 1$ and $S_{loc} = 0$). We propose that low

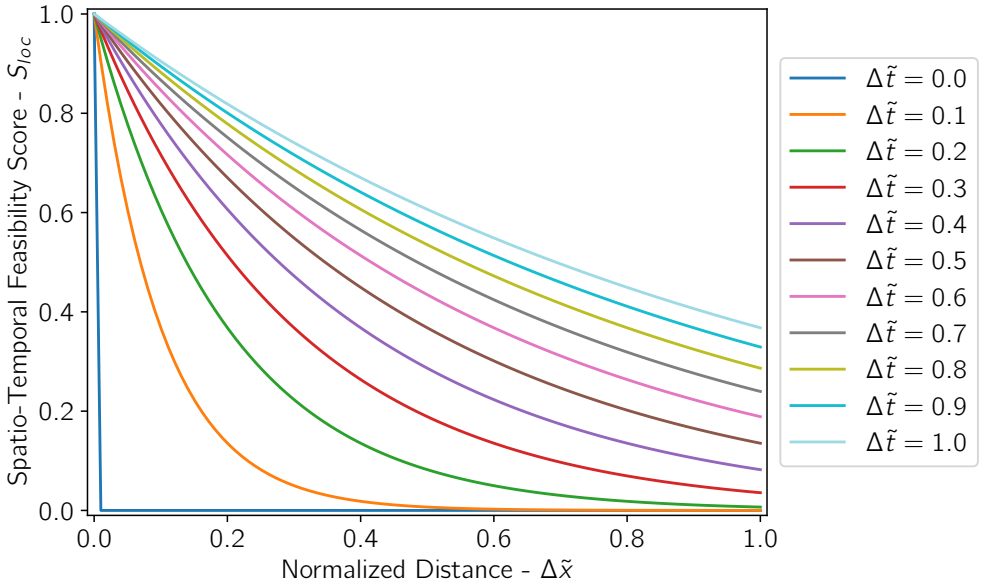


Figure 3. Spatio-temporal feasibility score as a function of normalised distance $\Delta\tilde{x}$. shown for multiple values of $\Delta\tilde{t}$. Larger $\Delta\tilde{t}$ results in slower decay with distance, reflecting increased tolerance to spatial displacement over longer time intervals.

feasibility pairs provide a stronger constraint than high feasibility pairs which do not guarantee identity correspondence.

$$\mathscr{L}_{WMSE} = \frac{\lambda_p}{|\mathscr{P}^+|} \sum_{(i,j)\in\mathscr{P}^+} (S_{im}(i,j) - 1)^2 + \frac{\lambda_n}{|\mathscr{P}^-|} \sum_{(i,j)\in\mathscr{P}^-} (S_{im}(i,j))^2 \tag{4}$$

$$\mathscr{P}^+ = \{(i,j) \mid S_{loc}(i,j) = 1\} \tag{5}$$

$$\mathscr{P}^- = \{(i,j) \mid S_{loc}(i,j) = 0\} \tag{6}$$

This objective encourages similar embeddings for spatio-temporally proximate samples while penalizing similarities for impossible pairs. Although high feasibility scores do not guarantee identity correspondence, they frequently capture co-occurrence within the same social group or pack, which gives a strong prior for the downstream task. Further, we include weakly positive pairs during training to prevent representational collapse, as training exclusively on negative pairs would encourage the model to trivially minimize similarity across all samples. As the goal of the system is to recommend candidate matches for human annotation rather than perform fully automatic identification, exploiting these weak positive cues improves the relevance of the proposed image pairs. Formulating the learning objective in a pairwise manner converts the problem into a binary classification task (intra-individual vs. inter-individual pairs), which increases the number of training signals per label, removes the need for prior knowledge of the number of identities, and partially alleviates class imbalance. Final pairwise similarities are computed through element-wise multiplication of spatio-temporal and image embedding similarities, yielding scores that leverage both modalities and improve robustness over image-only baselines.

Active Pair Sampling

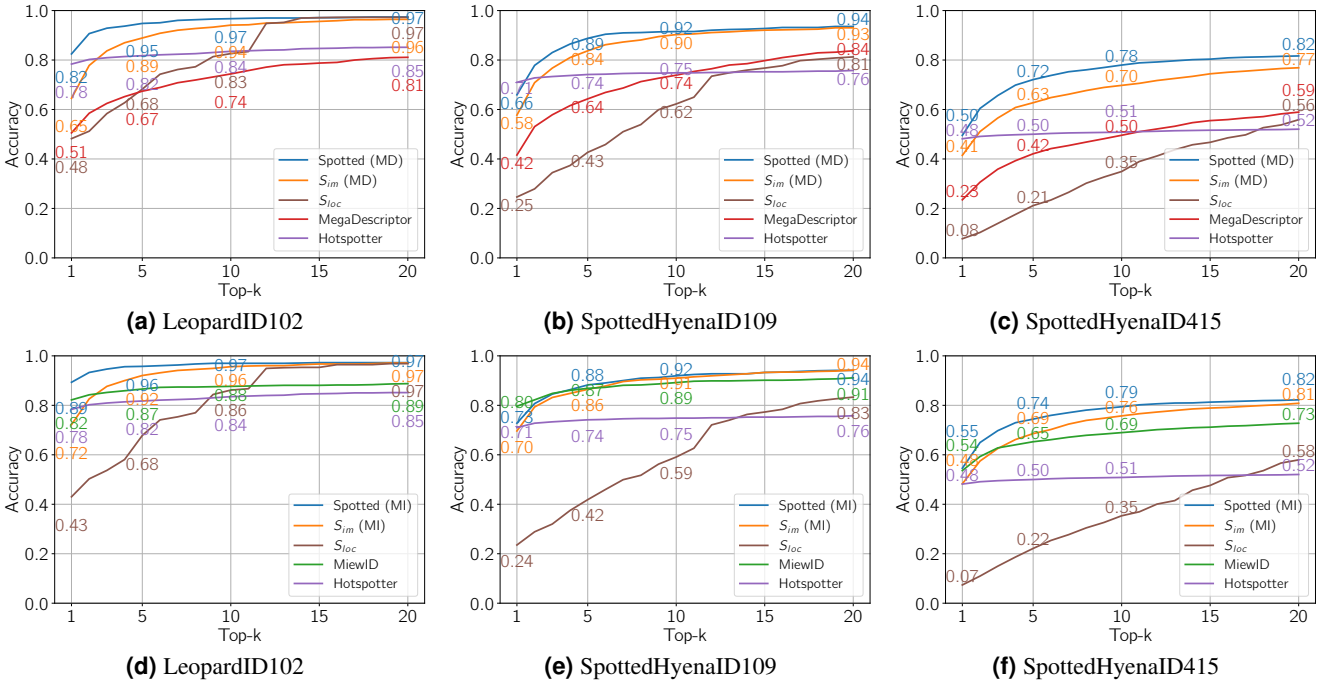
The objective of the annotation stage is to minimize the total number of required pairwise comparisons pursued by a human annotator. Expert annotated pairwise labels are encoded as constraints: confirmed matches are treated as *must-link* constraints, while confirmed non-matches between distinct individuals are treated as *cannot-link* constraints. At each annotation step, these pairwise constraints are propagated to infer all implicit relationships that can be derived from the new pairwise label obtained by the annotator. A simple greedy strategy, which always proposes the highest scoring image pair for labeling, is effective at identifying easy matches early but diminishes in performance with an increasing number of queries. As annotation progresses, only the ambiguous low-confidence set of matching image pairs remains for which the pairwise score quality is uninformative regarding their labels. Since we obtain ground-truth pairwise labels when the human annotator labels queried pairs and their labels are propagated, these annotations can be directly reused as a supervisory signal. To address this, we adapt the uncertainty-based pair sampling and training strategy introduced in the SPAL method[31]. We use our pairwise similarity scores to train an additional two-layer MLP, inspired by the training methodology presented in SPAL[31]. Uncertain pairs are

identified by clustering the data with constrained DBSCAN[31,34], selecting the most dissimilar pair of images within each cluster and the most similar pair of images between each cluster. These pairs are passed to the expert for labeling and the annotations serve as a training signal. In accordance with SPAL, we apply a combination of triplet and contrastive losses to push samples towards cluster centroids. SPAL aims to select informative pairs for annotation by sampling pairs from cluster boundaries. However, as clusters begin to stabilize, fewer new samples are selected as the cluster boundaries become static, at which point we propose to extend SPAL with a confidence-based sampling scheme to select the highest-scoring pair for labeling, which we refer to as ConfSPAL. We demonstrate that dynamic sampling strategies can effectively combine informativeness with coverage, outperforming prior approaches that prioritize informative samples exclusively in terms of queried human comparisons.

Experimental Results

Experimental Settings

During the location-informed pre-training stage, we use the Adam optimizer with a learning rate of $1e-3$ and train for 300 epochs with full-batch updates. The L2 loss weights, λ_n and λ_p, are empirically set to 2 and 1, respectively. Additionally, we train the active sampling head during the annotation stage separately using an Adam optimizer at a learning rate of $3e-4$ with full batch updates. For constrained DBSCAN clustering, the minimum number of samples was set to 2. The DBSCAN neighbourhood radius ε was selected through a grid search by qualitative inspection of the resulting distribution of cluster sizes, with the objective of limiting excessively large clusters, typically capped to the range of 10-50 samples, while reducing the number of single-element clusters. The selected values of ε for the LeopardID102, SpottedHyenaID109, and SpottedHyenaID415 were 0.44, 0.44 and 0.49 for *Spotted* (MD), 0.62, 0.60, and 0.57 for *Spotted* (MI), 0.21, 0.22, and 0.19 for MegaDescriptor, and 0.47, 0.45, and 0.41 for MiewID, respectively.



(a) LeopardID102 **(b)** SpottedHyenaID109 **(c)** SpottedHyenaID415

(d) LeopardID102 **(e)** SpottedHyenaID109 **(f)** SpottedHyenaID415

Figure 4. Top-k accuracy curves for our three datasets. The top-k accuracy is computed as the proportion of query images for which at-least one true match is retrieved within the top-k highest scoring image pairs. (a-c) MegaDescriptor is used as the frozen image backbone. (d-f) MiewID is used as the frozen image backbone.

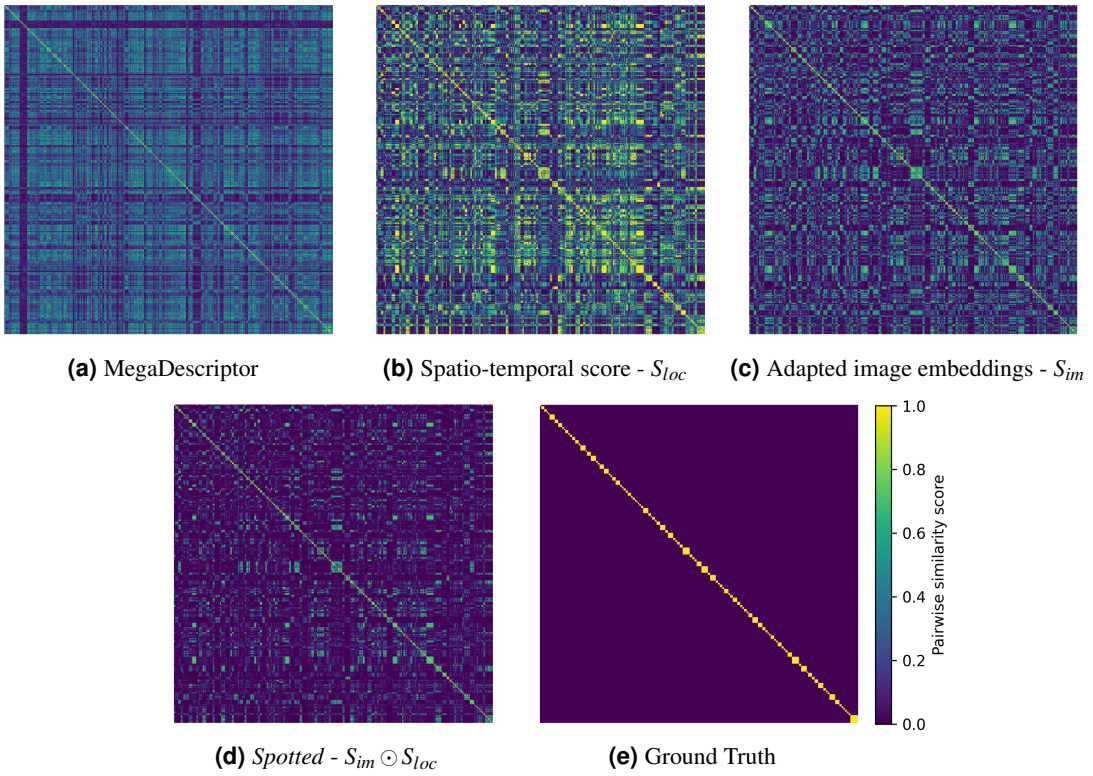
We evaluate our approach against three baselines, HotSpotter[5], MegaDescriptor[11], and MiewID[10]. MegaDescriptor and MiewID represent the current state-of-the-art in multi-species animal ReID and thus serve as strong baselines, while HotSpotter is still widely used in practical wildlife monitoring pipelines. We evaluate our model in two stages. First, we assess the quality of our pairwise similarity scores using top-k identification accuracy. Second, we simulate a human-in-the-loop dataset labeling scenario and measure annotation efficiency in terms of the number of queried image pairs until all positive pairwise matches are discovered.

Scoring Performance

As shown in Fig. 4, *Spotted* outperforms visual-only baselines across all datasets. In comparison to MegaDescriptor embeddings, our method achieves improvements of 28pp, 24pp, and 31pp over this backbone and 9pp, 2pp and 9pp over MiewID in top-5 accuracy on the LeopardID102, SpottedHyenaID109, and SpottedHyenaID415 datasets. While Hotspotter remains competitive in terms of top-1 accuracy, its performance stagnates for larger k, reflecting its tendency to identify a small number of strong matches and missing further plausible candidate matches.

To explore the relative impact of the spatio-temporal prior compared to the visual features, we additionally report the top-k accuracy curves for the image only (S_{im}) and location only (S_{loc}) scores. Although the spatio-temporal score alone shows limited accuracy at low k values, it achieves competitive top-k accuracy for high k values. In particular, we observe top-20 accuracies of 97%, 81%, and 56% for the LeopardID102, SpottedHyenaID109, and SpottedHyenaID415 datasets, respectively. This shows that spatial and temporal data capture highly informative priors, helping to constrain the search space for plausible matches. It is also important to note the impact of species and population size on the effectiveness of the spatio-temporal priors. For example, the SpottedHyenaID415 dataset, which contains 415 individuals, exhibits substantially lower top-1 accuracy on location-only scores (8%) compared to the SpottedHyenaID109 dataset (25%) which contains 109 individuals (Table 1). This is expected, as a larger population leads to more individuals per location, reducing the discriminative power of location-only priors.

These results also reflect some differences in species movement behaviour. Although the LeopardID102 and SpottedHyenaID109 datasets contain a similar number of individuals (102 - 109), the location-only model achieves substantially higher top-1 accuracy on LeopardID102 (48%) than on SpottedHyenaID109 (25%). This is in line with the movement patterns observed in the datasets, where leopards tend to remain in spatially consistent regions, whereas hyenas were more likely to travel longer distances between sightings, introducing greater uncertainty in location-based cues for hyenas. Consequently, we highlight the need to combine image and location-prior models to achieve the highest scores across all datasets and backbones.



(a) MegaDescriptor **(b)** Spatio-temporal score - S_{loc} **(c)** Adapted image embeddings - S_{im}

(d) *Spotted* - $S_{im} \odot S_{loc}$ **(e)** Ground Truth

Figure 5. Similarity matrices between all pairs of images within the SpottedHyenaID109 dataset. The data has been ordered by animal id.

Fig. 5 shows the pairwise similarity matrices for the SpottedHyenaID109 dataset at different stages of the proposed framework. The pairwise similarity matrix of the MegaDescriptor embeddings in Fig. 5a) visually exhibits limited discriminative structure, with the similarity scores distributed too uniformly across both matching and non-matching pairs to reliably distinguish between them. Adapting the embeddings using spatio-temporal pseudo-labels (Fig. 5b) suppresses a large fraction of these

spurious similarities, as can be seen from the lower-scoring regions throughout the matrix. Nonetheless, a subset of negative pairs remains highly scored. The location-based similarity matrix (Fig. 5c) assigns high scores to many image pairs, reflecting the coarse but inclusive nature of the spatial prior, while still ensuring that true matching pairs consistently receive high scores. Since spatio-temporal scores of zero correspond to guaranteed non-matches, whereas high-scoring pairs provide only a weak positive signal based on proximity, the training process tends to avoid false-negative matches. Because of this bias in the spatio-temporal scoring function, both the adapted image embeddings and the spatio-temporal scores tend to produce more false positives than false negatives. This is particularly useful in a human-in-the-loop setting because missed matches can lead to an exhaustive manual search for a match. Consequently, true matches generally receive high scores in both modalities. Element-wise fusion of the image and spatio-temporal similarity signals combines these signals by preserving pairs that are consistently strong in both representations while suppressing spurious matches arising from visual ambiguity or spatial proximity alone. This joint bias toward retaining potential matches allows the fused score to reduce erroneous high similarities while maintaining strong responses for genuine matches, producing a clearer block-diagonal structure that closely aligns with the ground-truth pairwise labels (Fig. 5e).

Human-in-the-Loop Performance

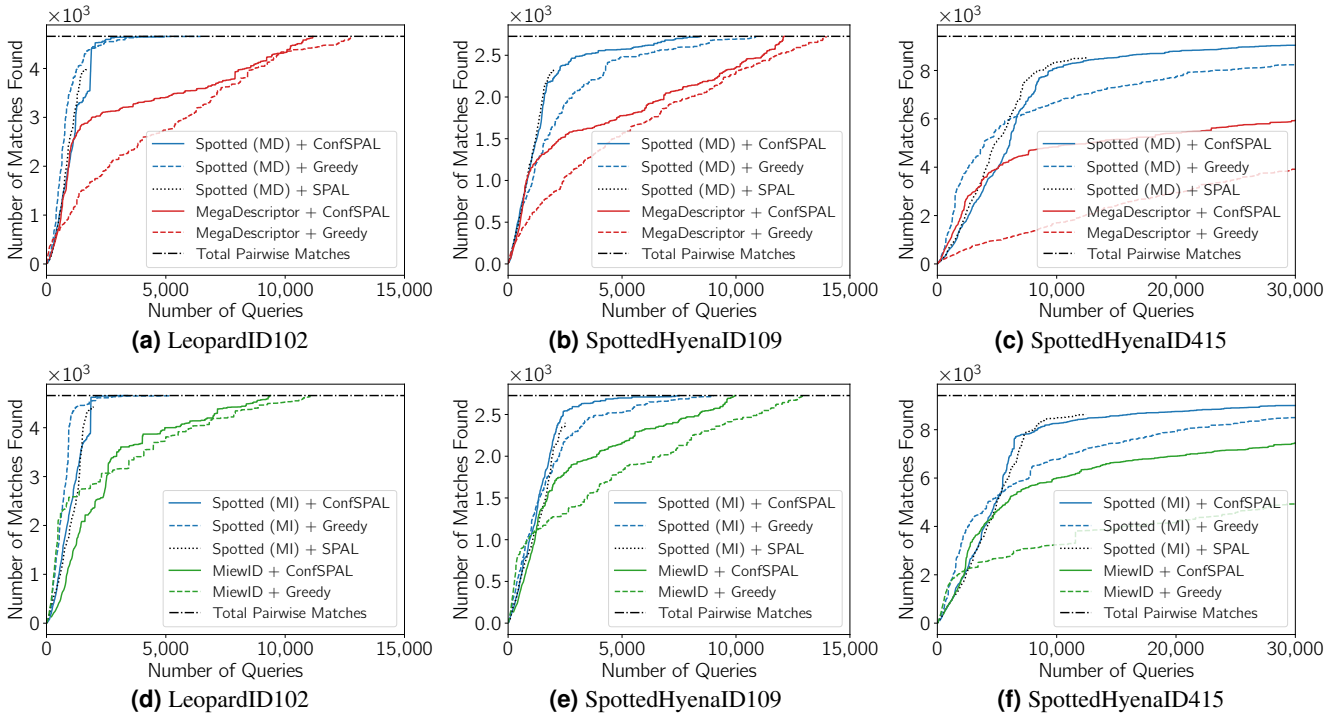


Figure 6. Cumulative number of confirmed positive matches discovered as a function of annotation queries for our three datasets. We compare greedy sampling against our SPAL-inspired active sampling with and without the confidence sampling step. MegaDescriptor, MiewID and our *Spotted* similarities are compared. (a-c) MegaDescriptor (MD) is used as the frozen image backbone. (d-f) MiewID (MI) is used as the frozen image backbone. The dashed black line indicates the total number of positive pairwise matches.

Figs. 6 and 7 compare the cumulative number of positive matches discovered and the reduction in unlabeled pairs as a function of annotation queries. Our model *Spotted* consistently outperforms both the MegaDescriptor and MiewID baselines, demonstrating the importance of spatio-temporal priors alongside visual similarity. *Spotted* (MI) + ConfSPAL reduces the number of human queries required to fully label the dataset by 69.2% on LeopardID102 and 42.0% on SpottedHyenaID109 compared to MiewID with greedy sampling, corresponding to a reduction from 31.8 to 9.8 and 37.4 to 21.7 labeling hours for the LeopardID102 and SpottedHyenaID109 datasets respectively, extrapolating a measured annotation rate of 5.78 queries per minute. For the larger SpottedHyenaID415 datasets *Spotted* (MI) + ConfSPAL discovers 4070 more positive matches than MiewID + Greedy within the same labeling budget of 86 hours, corresponding to 95.6% of all positive matches discovered compared to only 52.4% for MiewID + Greedy. Notably, even when the top-5 accuracy improvement on SpottedHyenaID109 is modest at 2pp, this translates to a substantial 42% reduction in required human queries in a simulated labeling scenario,

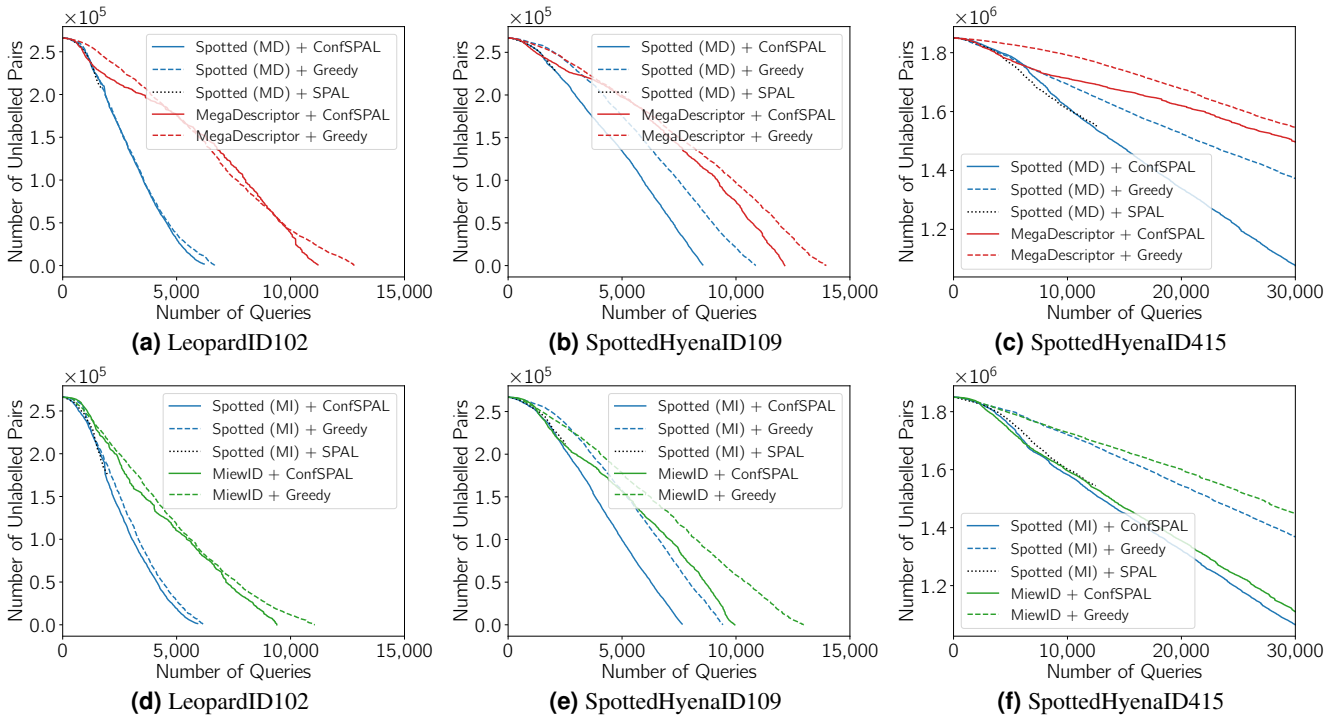


Figure 7. Number of remaining unlabelled image pairs as a function of annotation queries for our three datasets. (a-c) MegaDescriptor (MD) is used as the frozen image backbone. (d-f) MiewID (MI) is used as the frozen image backbone. Faster reduction indicates improved annotation efficiency.

highlighting that small improvements in ranking quality can lead to considerable gains in annotation efficiency. The impact of active sampling is especially noticeable for SpottedHyenaID415, the most challenging of our three datasets. Combining *Spotted* with our active sampling method yields 799 and 500 additional positive matches over *Spotted* with greedy sampling for the MegaDescriptor and MiewID backbones, respectively. The results also demonstrate the role of active sampling in the Re-ID pipeline. While our SPAL-based active sampling improves performance when combined with our fused embeddings, it alone is insufficient to compensate for weak underlying representations. In particular, while applying ConfSPAL to the MegaDescriptor or MiewID embeddings does consistently boost performance over greedy sampling, *Spotted* with greedy sampling still outperforms the baseline methods with ConfSPAL. This behavior can be attributed to the quality of the induced similarity structure. When the embedding space does not form sufficiently coherent clusters, boundary-based sampling strategies select pairs that lie near poorly defined cluster margins. Although these samples satisfy the heuristic criteria of being informative (e.g. dissimilar intra-cluster or similar inter-cluster pairs), they are not discriminative due to the weak cluster structure. This leads to inefficient querying and slower discovery of true matches. In contrast, the stronger separation induced by the adapted and fused embeddings yields more coherent clusters in the early stages of annotation. This allows our ConfSPAL to identify informative samples for labeling, thereby improving the tuning of the ConfSPAL MLP head. Across all three datasets, *Spotted* reduces the total number of queries required to reach an equivalent (and in the case of the SpottedHyenaID415 dataset, greater) number of confirmed matches. Additionally, the rate at which matches are discovered is consistently higher, indicating improved annotation efficiency. We additionally compare against the vanilla SPAL sampling scheme, which omits the confidence-based sampling stage. As shown in figs. 6 and 7, vanilla SPAL tracks closely to ConfSPAL, but fails to propose new pairs once the cluster boundaries stabilize. This causes vanilla SPAL to terminate pair discovery before all matches have been found, whereas ConfSPAL's reversion to confidence-based sampling continues to propose new pairs beyond this point.

Conclusion

We introduced *Spotted*, a location-informed, human-in-the-loop animal re-identification framework. Our method utilizes the structure of camera-trap deployments by combining image-based visual embeddings with a spatio-temporal feasibility score derived from camera geometry and timestamps. By using speed-based, continuous feasibility scores as matching cues and pseudo-labels for a lightweight head, our approach suppresses physically unlikely matches without requiring identity labels. Additionally, coupling our location-informed visual network with our active pair sampling strategy, reduces the number of

required expert reviews to discover positive matches. *Spotted* consistently improves top-k retrieval performance over strong visual baselines across three challenging camera-trap datasets that we publish with this work. We encourage future work to release camera-trap locations and timestamps alongside public datasets, and to explore additional auxiliary cues as well as transductive learning and uncertainty to further boost animal reID performance.

Data Availability

The three animal re-identification datasets generated and/or analysed during the current study are available in the Hugging Face repositories at https://doi.org/10.57967/hf/9288, https://doi.org/10.57967/hf/9290, and https://doi.org/10.57967/hf/9289. The camera locations are released in relative coordinates to protect the privacy of the survey sites.

References

1. Delisle, Z. J., Flaherty, E. A., Nobbe, M. R., Wzientek, C. M. & Swihart, R. K. Next-generation camera trapping: Systematic review of historic trends suggests keys to expanded research applications in ecology and conservation. *Front. Ecol. Evol.* **Volume 9 - 2021**, DOI: 10.3389/fevo.2021.617996 (2021).

2. Adam, L., Čermák, V., Papafitsoros, K. & Picek, L. Seaturtleid2022: A long-span dataset for reliable sea turtle re-identification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 7146–7156 (2024).

3. Adam, L., Čermák, V., Papafitsoros, K. & Picek, L. Wildlifereid-10k: Wildlife re-identification dataset with 10k individual animals. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2090–2100 (IEEE, 2025).

4. Powell, R. A. & Mitchell, M. S. What is a home range? *J. Mammal.* **93**, 948–958, DOI: 10.1644/11-MAMM-S-177.1 (2012). https://academic.oup.com/jmammal/article-pdf/93/4/948/2899018/93-4-948.pdf.

5. Crall, J. P., Stewart, C. V., Berger-Wolf, T. Y., Rubenstein, D. I. & Sundaresan, S. R. Hotspotter—patterned species instance recognition. In *2013 IEEE workshop on applications of computer vision (WACV)*, 230–237 (IEEE, 2013).

6. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110, DOI: 10.1023/B:VISI.0000029664.99615.94 (2004).

7. Guo, S. *et al.* Automatic identification of individual primates with deep learning techniques. *iScience* **23**, 101412, DOI: https://doi.org/10.1016/j.isci.2020.101412 (2020).

8. Yu, Y., Vidit, V., Davydov, A., Engilberge, M. & Fua, P. Addressing the elephant in the room: Robust animal re-identification with unsupervised part-based feature alignment. *arXiv preprint arXiv:2405.13781* (2024).

9. Hou, S. *et al.* Openanimals: Revisiting person re-identification for animals towards better generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14369–14379 (2025).

10. Otarashvili, L., Subramanian, T., Holmberg, J., Levenson, J. & Stewart, C. V. Multispecies animal re-id using a large community-curated dataset. *arXiv preprint arXiv:2412.05602* (2024).

11. 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 (WACV)*, 5953–5963 (2024).

12. Zhu, K. *et al.* Pass: Part-aware self-supervised pre-training for person re-identification. In *European conference on computer vision*, 198–214 (Springer, 2022).

13. Moskvyak, O., Maire, F., Dayoub, F. & Baktashmotlagh, M. Keypoint-aligned embeddings for image retrieval and re-identification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 676–685 (2021).

14. Moskvyak, O., Maire, F., Dayoub, F. & Baktashmotlagh, M. Learning landmark guided embeddings for animal re-identification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, 12–19 (2020).

15. Shukla, A. *et al.* A hybrid approach to tiger re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* (2019).

16. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

17. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022 (2021).

18. Deng, J., Guo, J., Xue, N. & Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

19. Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 10096–10106 (PMLR, 2021).

20. Li, Y. *et al.* Metawild: A multimodal dataset for animal re-identification with environmental metadata. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 13009–13015, DOI: 10.1145/3746027.3758249 (Association for Computing Machinery, New York, NY, USA, 2025).

21. Kays, R., Hody, A., Jachowski, D. S. & Parsons, A. W. Empirical evaluation of the spatial scale and detection process of camera trap surveys. *Mov. Ecol.* **9**, 41, DOI: 10.1186/s40462-021-00277-3 (2021).

22. Mac Aodha, O., Cole, E. & Perona, P. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/cvf international conference on computer vision*, 9596–9606 (2019).

23. Picek, L., Neumann, L. & Matas, J. Animal identification with independent foreground and background modeling. In Cremers, D. *et al.* (eds.) *Pattern Recognition*, 241–257 (Springer Nature Switzerland, Cham, 2025).

24. Rosa, M. J., Abu Baker, M., Tang, C. M. & Littlemore, J. Gcn-id: A benchmark dataset for great crested newt re-identification using ai foundation models. In *Proceedings of the BMVC* (2025).

25. Li, S., Li, J., Tang, H., Qian, R. & Lin, W. Atrw: A benchmark for amur tiger re-identification in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 2590–2598, DOI: 10.1145/3394171.3413569 (Association for Computing Machinery, New York, NY, USA, 2020).

26. Botswana Predator Conservation Trust. Panthera pardus csv custom export (2022). Retrieved from African Carnivore Wildbook on 2022-04-28.

27. Schneider, S., Taylor, G. W., Linquist, S. & Kremer, S. C. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods Ecol. Evol.* **10**, 461–470 (2019).

28. Osner, N. TrapTagger.

29. Sani, D., Khurana, M. & Anand, S. Active learning for animal re-identification with ambiguity-aware sampling (2025). 2511.06658.

30. Norouzzadeh, M. S. *et al.* A deep active learning system for species identification and counting in camera trap images. *Methods Ecol. Evol.* **12**, 150–161, DOI: https://doi.org/10.1111/2041-210X.13504 (2021). https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13504.

31. Jin, D. & Li, M. Towards fewer labels: Support pair active learning for person re-identification (2022). 2204.10008.

32. 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).

33. Mallapragada, P. K., Jin, R. & Jain, A. K. Active query selection for semi-supervised clustering. In *2008 19th International Conference on Pattern Recognition*, 1–4, DOI: 10.1109/ICPR.2008.4761792 (2008).

34. Zhao, W., He, Q., Ma, H. & Shi, Z. Effective semi-supervised document clustering via active learning with instance-level constraints. *Knowl. Inf. Syst.* **30**, 569–587, DOI: 10.1007/s10115-011-0389-1 (2012).

35. Miao, Z. *et al.* Iterative human and automated identification of wildlife images. *Nat. Mach. Intell.* **3**, 885–895, DOI: 10.1038/s42256-021-00393-0 (2021).

36. 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 Biol.* **102**, 875–897, DOI: 10.1007/s42991-022-00224-8 (2022).

37. Perez, G., Sheldon, D., Van Horn, G. & Maji, S. Human-in-the-loop visual re-id for population size estimation. In *European Conference on Computer Vision (ECCV)* (2024).

38. Ren, T. *et al.* Grounded sam: Assembling open-world models for diverse visual tasks (2024). 2401.14159.

Acknowledgements

All figures were prepared by H.S.K. with assistance from J.H.

Author Contributions

H.S.K. led methodology and software development, visualization, and wrote the original draft. J.H. contributed to the technical discussion and assisted with the figures. K.H., L.H., B.M., K.N., J.S.S. and A.S. contributed to data curation. A.L. led funding acquisition and data curation. A.V. provided institutional support. M.W. led conceptualization, funding acquisition, and data curation. D.D.M. contributed to conceptualization, technical discussion, and provided supervision. All authors contributed to writing-review and editing.

Funding

H.S.K., A.L., J.S.S, A.S, and M.W. were funded by Allen Family Philanthropies. J.H. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1597 – 499552394 and J.H. acknowledges travel support from ELISE (GA no. 951847) and ELSA - European Lighthouse on Secure and Safe AI funded by the European Union under GA no. 101070617. We are grateful to Allen Family Philanthropies for funding this work.

Competing Interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- LaTeXSupplementaryfile.bbl