A human-in-the-loop solution for individual leopard identification in unlabeled camera trap images

Cheng Guo, *Student Member, IEEE*, Agnieszka Miguel, *Senior Member, IEEE*, Anthony A. Maciejewski, *Fellow, IEEE*

***Abstract*—This article introduces a human-in-the-loop (HITL) algorithm designed to address a real-world individual animal identification problem: identifying distinct animals from a set of unlabeled camera-trap images when the total number of individuals is unknown. A key contribution of this article is its explicit balance of two critical factors in evaluating identification algorithms: identification accuracy and the cost of expert human annotation, i.e., the number of manual decisions required to determine whether an image pair belongs to the same individual. Our HITL strategy uses human involvement only in the most difficult cases, employing autonomous identification for easily distinguishable cases, to obtain high overall identification accuracy with minimal human effort. The proposed framework consists of three components: HITL-based core clustering generation, HITL-based clustering verification, and HITL-based clustering growth. Experimental validation was performed on an African leopard dataset provided by *Panthera*, with the algorithm achieving identification accuracy comparable to a human baseline method, where each image is manually compared to its top-2 most similar images. The proposed approach reduces human involvement by 77.3%, requiring only 0.05% of all pairs to be manually labeled as belonging to the same or different individuals.**

***Note to Practitioners*—This paper addresses the real-world challenge of individual animal identification in an unlabeled image dataset. We propose a human-in-the-loop algorithm that achieves high identification accuracy with minimal human involvement and without prior knowledge. The method is particularly effective for unlabeled camera-trap image datasets of species with distinctive identifiable markings and a high individual-to-image ratio, especially when many animals appear in only a single image. By incorporating human confirmation, the algorithm can differentiate image pairs of the same individual that appear visually similar from those of different animals, due to significant variations in animal poses and other factors. Importantly, the algorithm significantly reduces the human effort required to determine whether image pairs belong to the same identity. This approach is particularly valuable for researchers studying new habitats or working with species for which deep learning techniques are not viable due to the lack of labeled datasets.**

***Index Terms*—computer vision for automation; robotics and automation in life sciences; object detection, segmentation and categorization; human-in-the-loop; automated animal identification; camera-trap images; clustering**

I. Introduction

INDIVIDUAL animal identification is increasingly important in wildlife studies, such as monitoring animal behaviors, tracking migration patterns, and estimating species-level populations [1]. Progress in this area is largely driven by camera traps, which offer a low-cost and noninvasive method for collecting a large volume of wildlife images. Camera traps are automated cameras equipped with motion or infrared sensors that are deployed across various locations throughout a target habitat during observation periods. However, because camera traps are triggered for many reasons, the captured images will include not only the target species but other wildlife, swaying vegetation, and sometimes only backgrounds. While modern deep learning models can reliably classify these images by species [2], identifying individual animals within a species from a new set of camera trap images remains a major challenge, particularly for elusive and seldom-seen wild animals. Even when a large set of images is available for a target species, the number of images per individual is often limited. This makes it difficult to distinguish unique biometric features, particularly under substantial variation in viewpoint, exposure, occlusion, and image quality [3].

Most studies on individual animal identification [4] treat the problem as a retrieval-based identification (Re-ID) task, extending concepts from face recognition. These methods typically assume that all individuals in the dataset are already known, i.e., the dataset is a closed set. In this setting, human annotators have already manually confirmed the identities for all existing image pairs, while the algorithm is only required to identify a new image by matching it to the most similar images with known IDs. However, earlier approaches did not account for the possibility that new images could belong to previously unknown individuals, making the dataset an open set. Stewart et al. [5] made initial progress in addressing this issue. Still, their method did not solve the "real-world problem", i.e., datasets with no labels at all, where an unknown number of individuals appear in a set of camera trap images [6]. Specifically, let $\mathcal{I}$ be an unlabeled image set with an unknown number of unique identities, denoted N. The goal is to partition the image set $\mathcal{I}$ into subsets, denoted C_i, each representing a distinct individual, forming a clustering, denoted $\mathcal{C}$. The number of individuals, denoted K, is then estimated such that $\bigcup_{i=1}^{K} C_i = \mathcal{I}$. This real-world problem is especially challenging compared to traditional Re-ID tasks because it requires evaluating identification algorithms in terms of both identification accuracy and the cost of human expertise annota-

Cheng Guo and Anthony A. Maciejewski are with the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523 USA (e-mails: Cheng.Guo@colostate.edu and aam@colostate.edu)

Agnieszka Miguel is with the Department of Electrical and Computer Engineering, Seattle University, Seattle, WA 98122 USA (e-mail: amiguel@seattleu.edu)

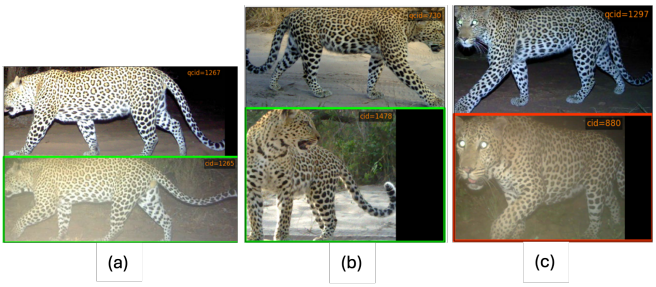

Fig. 1. This figure presents representative examples of image pairs from African leopard camera trap images. In (a), the image pair is easily identified as belonging to the same individual due to the clearly visible and distinctive spot patterns. In (b), the image pair is identified as belonging to the same individual primarily based on the head spot pattern. In (c), although the two leopards exhibit similar posture and eye-region appearance, the image pair corresponds to different individuals, as evidenced by the distinct spot patterns on their main bodies.

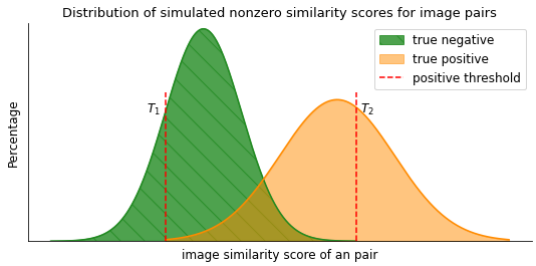

Fig. 2. This figure illustrates simulated probability density functions of nonzero image similarity scores for true positive and true negative pairs in a typical dataset. It reveals a noticeable overlap between true positives with small similarity scores and true negatives with relatively large similarity scores, primarily due to significant variations in animal poses. One key to developing an accurate and efficient individual animal identification algorithm is determining an appropriate positive threshold that maximizes correct automated identification of true positives while minimizing misidentification of true negatives.

tion, i.e., the number of human decisions to determine whether an image pair belongs to the same identity. This decision can be very difficult and time consuming due to the wide variation in leopard viewpoints. There has been previous work [7] proposed to fully automate individual identification operating directly on unlabeled data without any prior knowledge of individual identities. It achieves high identification accuracy when the images exhibit clearly distinct spot patterns, as shown in Fig. 1(a). However, due to significant variations in leopard poses and other factors, some image pairs from the same individual (referred to as *true positives*) often appear indistinguishable from those of different leopards (referred to as *true negatives*), leading to misidentification. This challenge is illustrated in Fig. 2, which shows a considerable overlap between true positive pairs with small similarity and true negatives with relatively large similarity. Fig. 1(b) and (c) show a true positive pair with small similarity and a true negative pair with relatively large similarity, respectively, from African leopard camera trap images. This overlap makes it difficult for a fully automated algorithm to set an appropriate threshold for identifying images of the same identity. For example, the threshold T_1 that captures all true positives will inevitably misidentify some true negatives, leading to false positives. Conversely, the threshold T_2, which avoids false positives, may cause the algorithm to miss some true positives, splitting images of the same individual across multiple clusters.

Therefore, in this paper, we propose an algorithm to address the real-world challenge of identifying individuals within an unlabeled camera trap image set of a given species, African leopards, provided by *Panthera* [8], and estimating the number of identities. This problem is particularly challenging because many individuals are represented by only a single image, and visual similarities make it difficult to distinguish between images of the same animal with low similarity and those of different individuals with relatively large similarity. The main contribution of this paper is the design of a novel algorithm that efficiently incorporates human participation through a human-in-the-loop (HITL) strategy for individual animal identification. Compared to the fully automated algorithm, our approach achieves higher identification accuracy with minimal human involvement. Specifically, we (1) design a novel HITL-based core clustering generation algorithm that enables a maximal number of confident autonomous identifications with limited human inputs by determining an appropriate positive threshold, as illustrated by T_2 in Fig. 2; (2) develop a novel HITL-based clustering verification mechanism, where a small number of human validations are periodically used to verify algorithmic identifications; (3) propose a new HITL-based clustering growth algorithm, where human involvement is used to determine whether image pairs selected by the algorithm as potential positives belong to the same identity. This is particularly important for pairs that lie in the overlap between true positives with small similarity and true negatives with relatively large similarity, as illustrated between T_1 and T_2 in Fig. 2. Together, these strategies ensure that human effort is focused primarily on cases that are inherently challenging for fully automated identification algorithms, while maximizing autonomous identification on distinguishable cases. In addition, to more accurately estimate clustering performance, we (4) extend the definition of the silhouette score. The remainder of this paper is organized as follows. Section II reviews related work on individual animal identification and provides background on the techniques leveraged by our algorithm. Section III presents our proposed HITL-based algorithm for identifying individual African leopards in unlabeled images. The evaluation of the algorithm is detailed in Section IV. Finally, Section V concludes the paper.

II. Background

A. Related Literature on Animal ID

The individual animal identification problem is typically decomposed into several subproblems. First, to reduce background interference, image segmentation techniques are employed to isolate the animal from its surroundings, often using deep learning-based methods, such as Mask R-CNN [9]. Next, feature extraction approaches are applied to segmented images to capture unique biometric characteristics. The choice of feature extraction method typically depends on the identifiable traits of a target species. For animals with distinctive

identifiable markings, such as leopards, zebras, and tigers, local region feature detectors can be effective. Examples of this approach include Hotspotter [10], Superpoint [11], and ALFRE-ID [12]. For animals without unique identifiable markings, such as bears, wolves, and sheep, deep learning-based methods are generally employed [13], [14]. Finally, to compare the similarity between image pairs, a matching score mechanism is established [10], [15], where the score magnitude reflects the confidence that the pair belongs to the same identity.

As mentioned, most recent research [16]–[18] continues to treat the individual animal identification problem as either a retrieval task or a classification problem, i.e., the algorithm determines the identity of a new image by retrieving the most similar images from a set of labeled images. However, these approaches assume that human annotators have already identified the existing images by manually confirming whether each pair belongs to the same identity. As dataset size increases, the demand for human input grows significantly, resulting in a substantial labor burden. In addition, individual animal identification is far more difficult than species-level identification. It requires human annotators with specialized expertise to distinguish distinct spot patterns across different images. At present, the limited number of human expert annotators and the substantial time required to handle visually challenging cases result in only small labeled image datasets for individual identification. Fortunately, recent studies have begun to acknowledge this challenge [15], [19], incorporating human labeling costs into algorithm evaluations, and proposing HITL-based mechanisms. However, these works focus solely on estimating the number of individuals in images and do not address the more complex task of identifying specific individuals. Furthermore, Sani et al. [20] applied active learning for animal re-identification, but their work does not address the real-world animal identification problem, namely identification without any prior information about individual identities. To address the real-world animal ID problem, Crall et al. [21] reduce manual effort by having human annotators confirm whether each image shares the same identity with its top-m most similar images, based on a similarity score. Their approach focuses human effort on the most promising image pairs, avoiding exhaustive annotation of all possible pairs. However, selecting an appropriate value for m requires prior knowledge. A large m improves identification accuracy but entails a substantial labor burden, while a small m reduces human effort at the cost of accuracy. Moreover, human effort is significantly affected by the dataset size, leading to heavy involvement for large-scale datasets.

We propose a novel HITL-based individual animal identification algorithm that requires human involvement only to confirm a limited number of image pairs as belonging to the same identity. Most of these pairs are identified by the algorithm as potential positives, while a small subset is randomly sampled from the set of automatically identified positive pairs. It is important to note that due to significant variations in animal poses, these image pairs present challenges not only for algorithmic identification but also for human annotators. In this paper, we evaluate human effort solely based on the number of annotations, without considering annotation difficulty. Additionally, our algorithm leverages existing image segmentation, feature extraction, and similarity score methods as preprocessing steps. It starts with the similarity scores computed for all image pairs in the dataset.

B. Background on Hotspotter

We selected Hotspotter [10] to compute similarities between leopard images. First, Hotspotter identifies all keypoints (typically corner or edge locations) in an image using a variant of SIFT (scale invariance feature transform) [22]. Then, for each keypoint, Hotspotter calculates the histogram of oriented gradients in its neighborhood as its descriptor [23].

To compare the similarity of an image to another image, Hotspotter first measures similarities between their descriptors. Assume two images, I_x and I_y, are from an image dataset $\mathcal{I}$. For each descriptor, denoted $\boldsymbol{d}_{x_k}$, in the first image I_x, Hotspotter computes the Euclidean distance to find its two nearest neighbor descriptors, denoted $\boldsymbol{d}_{\text{nn}}$ and $\boldsymbol{d}_{\text{nn2}}$, from all other images in the dataset. That is, given a descriptor $\boldsymbol{d}_{x_k}$, its matching score, denoted $\delta(\boldsymbol{d}_{x_k})$, to the two nearest neighbor descriptors is

$$\delta(\boldsymbol{d}_{x_k}) = \|\boldsymbol{d}_{\text{nn2}} - \boldsymbol{d}_{x_k}\|^2 - \|\boldsymbol{d}_{\text{nn}} - \boldsymbol{d}_{x_k}\|^2, \tag{1}$$

where $\boldsymbol{d}_{\text{nn}}, \boldsymbol{d}_{\text{nn2}} \in (\mathcal{D}_{\text{all}} - \mathcal{D}_x)$. The variables $\mathcal{D}_{\text{all}}$ and $\mathcal{D}_x$ are sets of descriptors from all images in the dataset $\mathcal{I}$ and the image I_x, respectively. Now, the similarity score, denoted $\Delta(I_x, I_y)$, of an image I_x to another image I_y is calculated by summing the matching scores for all qualified descriptors to their two nearest neighbor descriptors if their nearest neighbor descriptor is from the image I_y. That is,

$$\Delta(I_x, I_y) = \sum_{\boldsymbol{d}_{x_k} \in \mathcal{D}_x} \delta(\boldsymbol{d}_{x_k}), \text{ only if } \boldsymbol{d}_{\text{nn}} \in \mathcal{D}_y, \tag{2}$$

where $\mathcal{D}_y$ is a set of descriptors from the image I_y. The value of $\Delta(I_x, I_y)$ is set to 0 when no nearest neighbor descriptor from image I_y is found. Hotspotter finally computes an $N \times N$ pairwise similarity score matrix for N images within a given African leopard image dataset, which served as the input to our algorithm.

C. Background on Clustering

Our algorithm, presented in Section III, uses a graph partitioning algorithm [24] to generate a clustering. Specifically, at first, all images are regarded as clusters that only include themselves, and then image pairs recognized as the same identities are assigned to the same clusters to build a clustering.

In addition, to make our HITL-based individual leopard identification algorithm more efficient, we build a minimum spanning tree (MST) [25] for each cluster in the clustering. To construct an MST for a cluster, each image within the cluster is treated as a node in a graph. Edges are formed between pairs of images, inversely weighted by their similarity scores. For a graph with n nodes, the MST connects all nodes using $n-1$ edges, while minimizing the total edge weight. In our case, the minimum total edge weight corresponds to the maximum total similarity scores across the selected image pairs.

D. Background on Silhouette Score

We employ a silhouette score [26] to evaluate the quality of a given clustering. The silhouette score measures how well an image fits its own cluster compared to its nearest neighbor cluster. To calculate the silhouette score of an image, its nearest neighbor cluster, i.e., the cluster with the largest mean similarity score to the image except for its own cluster, is determined first. The mean similarity score, denoted $\bar{\Delta}_i$, for an image I_x with a cluster C_i is given by

$$\bar{\Delta}_i = \frac{1}{|C_i|} \sum_{I_y \in C_i} \Delta(I_x, I_y)^2. \tag{3}$$

The mean similarity score of an image with its own cluster is calculated using (3) except for removing I_x from the calculation. Recall that the silhouette score estimates the degree to which an image belongs to its cluster by comparing the difference from its own cluster to its nearest neighbor cluster. That is, a silhouette score, denoted s_x, of an image I_x is given by

$$s_x = \frac{\bar{\Delta}_o - \bar{\Delta}_{\text{nn}}}{\max\{\bar{\Delta}_o, \bar{\Delta}_{\text{nn}}\} + \epsilon}, \tag{4}$$

where $\bar{\Delta}_o$ and $\bar{\Delta}_{\text{nn}}$ are the mean similarity score of the image I_x with its own cluster and its nearest neighbor cluster, respectively, and ϵ is needed to avoid division by zero. According to (4), the range of a silhouette score is from -1 to 1. A value of 1 indicates that an image belongs to its cluster with maximum confidence, while -1 means that it does not with maximum confidence. Fig. 3(a) shows an example of an image that belongs to its cluster with high confidence. This own cluster containing multiple images is referred to as *nonisolated cluster* and is denoted C_{niso}.

It is important to note that the standard silhouette score is not defined for clusters containing a single image because (3) is not defined if $|C_i| = 1$. These clusters are referred to as *isolated clusters* and are denoted C_{iso}. Unfortunately, single image clusters are very common in wild leopard image datasets. In addition, the similarity scores of image pairs vary greatly due to differences in leopard poses and variations in the surrounding environment, which affects the accuracy of the silhouette score. To address these issues, we propose a novel expanded definition of the silhouette score in Section III.

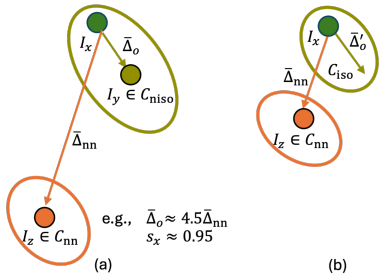

Fig. 3. This figure illustrates the relationships between an image and its own cluster and its nearest neighbor cluster. Images I are represented by solid circles, and clusters C are enclosed by colored ellipses. The length of an arrow is proportional to the reciprocal of the mean similarity score $\bar{\Delta}$ of an image with a cluster. In (a), an example is shown where an image I_x belongs to a nonisolated cluster C_{niso} with high confidence, i.e., the silhouette score s_x is large. This indicates that the image I_x is strongly similar to another image I_y within its own cluster and highly dissimilar to an image I_z in its nearest neighbor cluster C_{nn}. In (b), an example is displayed where an image I_x belongs to an isolated cluster C_{iso} with low confidence. Because Δ_o is undefined, we define a suitable value for the mean similarity score of an isolated image, denoted $\bar{\Delta}'_o$, to calculate its silhouette score.

III. HUMAN-IN-THE-LOOP ON INDIVIDUAL LEOPARD ID

A. Overview of HITL-based Identification Algorithm

In this article, we propose a human-in-the-loop (HITL) based individual animal identification algorithm that is applied to an unlabeled camera trap image dataset of African leopards. Due to significant variations in leopard poses and environment, distinguishing between images of the same leopard and different leopards presents a major challenge. Therefore, our goal is to correctly label the identities of all images and determine the number of individuals in the dataset with limited human involvement. The algorithm only asks animal experts to determine whether two images belong to the same leopard (true positive) or different leopards (true negative). Our algorithm consists of four stages: 1) initialization, where an initial positive threshold is determined using random human input, allowing image pairs with large similarity scores to be automatically identified; 2) core clustering generation, where a clustering is built based on positive image pairs using an adaptive positive threshold; 3) clustering verification, where the algorithm asks a human to check randomly selected positive image pairs to verify the correctness of the autonomous decisions; and 4) clustering growth, where the generated clustering is continuously updated with newly identified true positive image pairs. A flowchart of the process is shown in Fig. 4. More specifically, before our algorithm, we use a pretrained mask R-CNN [9] model to segment leopard objects from the background in images and employ Hotspotter [10] to compute a similarity score matrix for the dataset, which serves as input to our algorithm. To achieve algorithmic decisions for image pairs with large similarity scores, we propose an initial positive threshold based on a randomly selected small subset of image pairs that are identified as true positives by a human. A core clustering is then generated from positive image pairs whose similarity score exceeds this initial threshold. To validate the correctness of these positive image pairs, we introduce clustering verification, where the algorithm randomly selects a small subset of positive pairs and requests human verification. If the validation is successful, the positive threshold is reduced based on the similarity scores of positive pairs in the current clustering. This process iterates until the validation fails or the positive threshold cannot be reduced. The next phase is to grow the core clustering by identifying potential positive image pairs, where an image with a low silhouette score is paired with an image from its nearest neighbor cluster. These potential image pairs are confirmed by a human, and the clustering is updated accordingly. To ensure continuous validation of unchecked positive image pairs, the clustering verification procedure runs simultaneously. The process continues until the clustering no longer changes. The pseudocode for our algorithms is provided in Algorithms 1, 2, and 3. It is important to note that each image in the dataset

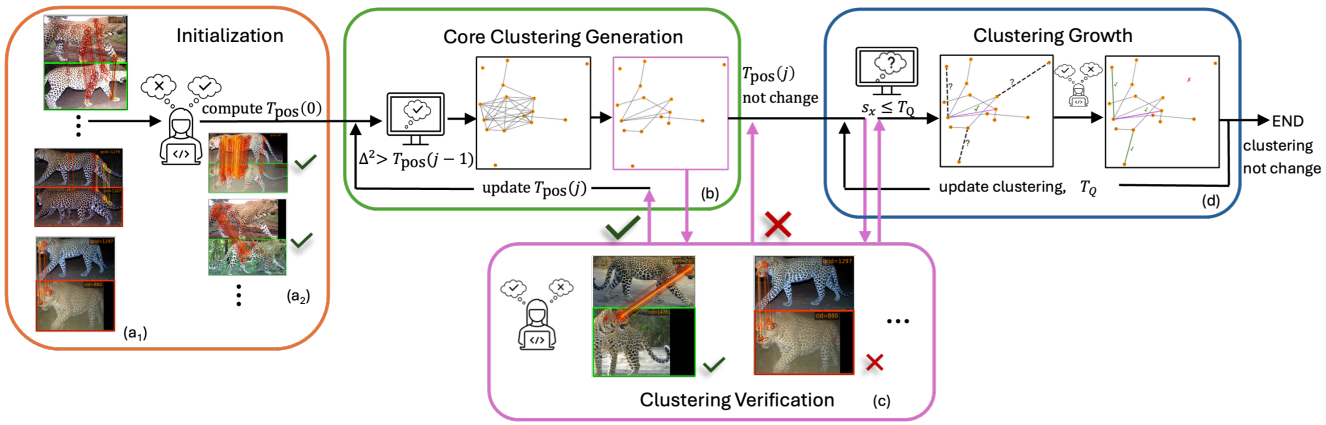

Fig. 4. This figure illustrates the flowchart of our human-in-the-loop solution for identifying individual leopards. Our algorithm consists of four stages: initialization, core clustering generation, clustering verification, and clustering growth. Image similarity scores for segmented leopard image pairs are computed and serve as input to our algorithm. In (a_1), examples of these image pairs with highlighted similarity regions are shown. To enable algorithmic decisions, human input is used to determine an initial positive threshold, $T_{\text{pos}}(0)$. Examples of image pairs with large similarity scores are displayed in (a_2). Then, the algorithm generates a core clustering based on image pairs whose similarity scores exceed T_{pos}. A subset of the core cluster graph and its minimum spanning tree are shown in (b). During core clustering generation, the algorithm periodically requests human verification of a random subset of positive image pairs, as shown in (c). If a human confirms that the pairs are true positives, the core clustering generation continues and the threshold is updated. When the threshold does not change or verification fails, the algorithm begins identifying potential positive image pairs using an internal evaluation measure. Image pairs confirmed by a human are incorporated to update the clustering. This process repeats until the clustering no longer changes, as shown in (d).

is assumed to belong to an individual animal. Thus, every image pair is initially unverified by an expert and unknown by the algorithm, denoted as (X, X). Subsequently, some pairs are confirmed by a human as true positives or true negatives, denoted as $(T, +)$ and $(T, -)$, respectively. Some pairs are automatically identified as positive by the algorithm but remain unverified by a human, denoted as $(X, +)$. Occasionally, some automatically identified positive image pairs are verified as *false positives*, denoted as $(F, +)$, indicating that they belong to different individuals.

B. Expanded Silhouette Score

Before elaborating on our algorithm, we first discuss our expanded definition of the silhouette score, which plays a crucial role in our new method. As previously mentioned, the accuracy of the traditional silhouette score can be affected by extreme image similarity scores. For example, given an image, if the mean similarity score with its nearest neighbor cluster is close to zero, the silhouette score approaches 1, regardless of how poorly the image fits within its own cluster. Analogously, if some images within a cluster have very large similarity scores to the image, the silhouette score also approaches 1, even when the image is significantly different from the rest of the cluster. To address these issues, we restrict the minimum of $\bar{\Delta}_i$ for $I_x \notin C_i$ and the maximum of $\bar{\Delta}_i$ for $I_x \in C_i$. That is, the mean similarity score $\bar{\Delta}_i$ for an image I_x with a cluster C_i from (3) is modified to

$$\bar{\Delta}_i = \begin{cases} \frac{1}{|C_i|} \sum_{I_y \in C_i} \max\left(\Delta(I_x, I_y)^2, \Delta_{\min}\right), & \text{if } I_x \notin C_i \\ \frac{1}{|C_i|-1} \sum_{I_y \in C_i} \min\left(\Delta(I_x, I_y)^2, \Delta_{\max}\right), & \text{if } I_x \in C_i \end{cases} \tag{5}$$

where the lower bound $\Delta_{\min}$ and the upper bound $\Delta_{\max}$ are defined as $\exp(\mu_{\text{pos}} - 2\sigma_{\text{pos}})$ and $\exp(\mu_{\text{pos}} + 2\sigma_{\text{pos}})$, respectively. The corresponding variables μ_{pos} and σ_{pos} are the mean and standard deviation, respectively, of the logarithm of squared similarity scores for all nonzero positive image pairs in a given clustering. The value of 2 was empirically selected based on similarity score distributions observed in the experimental dataset and other public datasets.

To calculate the silhouette score of an isolated image, an appropriate value, denoted $\bar{\Delta}'_o$, of the mean similarity score for the image with its own cluster, needs to be specified because (5) is not defined for isolated images. To determine a reasonable value for $\bar{\Delta}'_o$, we use (4), substituting the mean silhouette score for nonisolated images, denoted $\bar{s}_{\text{niso}}$, for s_x, and $\Delta_{\min}$ for $\bar{\Delta}_{\text{nn}}$. That is,

$$\bar{\Delta}'_o = \frac{1}{1 - \bar{s}_{\text{niso}}} \Delta_{\min}, \tag{6}$$

where

$$\bar{s}_{\text{niso}} = \frac{1}{\sum_{C_{\text{niso}} \subset \mathcal{C}} |C_{\text{niso}}|} \left(\sum_{C_{\text{niso}} \subset \mathcal{C}} \sum_{I_y \in C_{\text{niso}}} s_y \right). \tag{7}$$

If $I_x \in C_{\text{iso}}$, the silhouette score s_x of the image I_x is computed using (4), with $\bar{\Delta}_o$ replaced by $\bar{\Delta}'_o$. An example of an image belonging to an isolated cluster with low confidence is shown in Fig. 3(b). By constraining the mean similarity score of an image with a cluster to a bounded range and redefining the silhouette score for isolated images, we obtain a more accurate measure of how well an image fits within its own cluster. This also improves the robustness of our HITL-based core clustering generation, clustering verification, and clustering growth algorithms, presented in the following subsections.

C. HITL-based Core Clustering Generation and Verification

Generally, traditional clustering algorithms automatically group images with large image similarity scores into the same identities. However, the key question is: what similarity

Algorithm 1: HITL-based core clustering generation

Input : Similarity score matrix for an image dataset $\mathcal{I}$
Number of human decisions for initialization, N_{init}
Output: Clustering, $\mathcal{C} \leftarrow \{\}$
Identity matrix for image pairs,
$\mathcal{Q}(I_x, I_y) = (\text{X}, \text{X})|(\text{X}, +)|(\text{T}, +)|(\text{F}, +)|(\text{T}, -)$
//(X,X):(unverified,unknown), (X,+):(unverified,pos.),
//(T,+):(true,pos.), (F,+):(false,pos.), (T,-):(true,neg.)
Set of unchecked positive pairs, $\mathcal{S}_{\text{ncheck}} \leftarrow \{\}$

```
/*Initialization*/
Initialize a set of positive image pairs, S_pos ← {}
Initialize Q(I_x, I_y) ← (X, X), ∀ I_x, I_y ∈ I; iteration, j ← 0
for i = 1 : N_init do
    Select an image pair I_x, I_y ∈ I, randomly using (8)
    if (I_x, I_y) is human-identified as the same then
        Q(I_x, I_y) ← (T, +), S_pos ← S_pos ∪ (I_x, I_y)
    else
        Q(I_x, I_y) ← (T, −)
Calculate initial threshold T_pos(j = 0) using (9) with S_pos
/*Core clustering generation*/
repeat
    j ← j + 1
    for each I_x, I_y ∈ I, if Δ(I_x, I_y) > T_pos(j − 1) do
        S_pos ← S_pos ∪ (I_x, I_y)
    Generate core clustering C based on positive set S_pos
    Build a set of minimum spanning trees (MSTs) for C
    for each (I_x, I_y) in MSTs, if Q(I_x, I_y) == (X, X) do
        Q(I_x, I_y) ← (X, +), S_ncheck ← S_ncheck ∪ (I_x, I_y)
    Calculate Δ̄'_o using (6) and update T_pos(j) using (11)
    /*Clustering verification*/
    Update Q, S_ncheck using Algorithm 2
    Generate C by including positive image pairs (I_x, I_y)
      where Q(I_x, I_y) == (X/T, +), ∀ I_x, I_y ∈ I
until T_pos(j) does not change or false positive pair found;
```

score is sufficiently large for the algorithm to confidently identify an image pair as positive? For this, we propose HITL-based core clustering generation and clustering verification algorithms, where the algorithm makes decisions guided by human input. Specifically, we use a human to determine an initial positive to automatically identify image pairs as positive if their similarity scores exceed this threshold, as shown in Fig. 4, initialization stage. This is done by randomly sampling high similarity score image pairs and asking the human if they have the same identity. Because the distribution of true positive image pairs varies significantly, it is impossible to guarantee a perfect threshold. If the threshold is too low, the algorithm may misidentify an image pair from different individuals as belonging to the same identity, resulting in a false positive. To detect such errors, the algorithm periodically asks a human to verify randomly selected subsets of positive pairs to ensure consistency, as shown in Fig. 4, clustering verification stage. If no misidentified pairs are found, the algorithm adaptively lowers the threshold to autonomously identify additional positive image pairs until the threshold can no longer be decreased, as shown in Fig. 4, core clustering generation stage. This approach efficiently reduces human involvement in the subsequent HITL-based clustering growth algorithm while limiting the risk of misidentification. The pseudocode for HITL-based core clustering generation and clustering verification algorithms is shown in Algorithm 1 and 2, respectively.

In the initialization stage (Algorithm 1 Line 1-10), the primary goal is to establish an appropriately large positive threshold that prevents negative image pairs with relatively large similarity scores from exceeding this threshold. For this, the algorithm randomly selects image pairs with large similarity scores from a dataset. Specifically, the pairs are randomly selected using a weighted probability distribution, where an image pair with a larger similarity score is more likely to be chosen. Additionally, we impose an upper limit on the probability weight to prevent the selection from being dominated by image pairs with extremely large similarity scores. That is, the initial probability, denoted P_{init}, of an image pair (I_x, I_y) from an image dataset $\mathcal{I}$ initially being chosen is given by

$$P_{\text{init}} \propto \min\left(\Delta(I_x, I_y)^2, \exp\left(\mu_{\text{all}} + 3\sigma_{\text{all}}\right)\right), \tag{8}$$

where $\propto$ is the proportionality symbol. The variables μ_{all} and σ_{all} are the mean and standard deviation, respectively, of the logarithm of the squared similarity scores for all nonzero image pairs in the image dataset $\mathcal{I}$. The upper bound is defined as $\exp\left(\mu_{\text{all}} + 3\sigma_{\text{all}}\right)$. The value of 3 was empirically selected based on similarity score distributions observed in the experimental dataset and other public datasets.

To determine an initial positive threshold, we use the average similarity score for human-confirmed true positive image pairs. That is, the positive threshold, denoted T_{pos}, at the initial iteration, denoted $j = 0$, is given by

$$T_{\text{pos}}(j=0) = \exp\left(\frac{1}{|\mathcal{S}_{\text{pos}}|} \sum_{(I_x, I_y) \in \mathcal{S}_{\text{pos}}} \ln\left(\Delta(I_x, I_y)^2\right)\right), \tag{9}$$

where $\mathcal{S}_{\text{pos}}$ is the set of true positive image pairs confirmed by a human at initialization, and (I_x, I_y) is an image pair within the set $\mathcal{S}_{\text{pos}}$. After determining the initial positive threshold, a clustering is established based on automatically identified positive image pairs.

While the initialization stage focuses on generating an appropriately large threshold, we also introduce a HITL-based clustering verification algorithm to ensure clustering accuracy. The pseudocode is provided in Algorithm 2. In this stage, the algorithm randomly samples a subset of automatically identified positive image pairs for human verification to ensure the reliability of autonomous decision. To efficiently verify positive pairs, a set of MST edges is constructed for a clustering, as discussed in Section II-C. These positive image pairs correspond to the non-redundant largest similarity connection within their respective clusters. This efficiently removes less visually informative image pairs within clusters. As a result, expert annotators can more easily verify their consistency, thereby reducing unnecessary annotation effort on additional cases. Additionally, because image pairs with smaller similarity scores are more likely to be misidentified, the random selection of automatically identified positive pairs within the MST set prioritizes pairs with lower similarity scores for human validation. For this, the verification selection probability is defined to be inversely proportional to

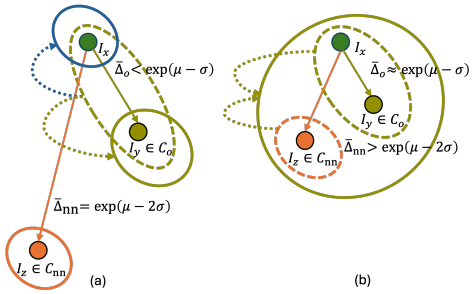

Fig. 5. This figure illustrates two specific cases between an image and its own cluster and its nearest neighbor cluster. Images I are shown using solid circles while clusters C are shown using different colored ellipses. The length of a solid arrow is inversely proportional to the mean similarity score $\bar{\Delta}$ of an image with a cluster. In (a), the value of $\bar{\Delta}_o$ for an image I_x with its own cluster C_o is small. This occurs when a small initial positive threshold is used during the core clustering generation, causing the image pair (I_x, I_y) to be automatically identified as positive. Fortunately, during the clustering verification, the algorithm selects this image pair for human validation, which confirms it as a false positive, and separates them. The dashed arrow shows this process. In (b), the value of $\bar{\Delta}_{\text{nn}}$ for an image I_x with its nearest neighbor cluster C_{nn} is large and close to $\bar{\Delta}_o$. Our clustering growth algorithm identifies the image pair (I_x, I_z) as a potential positive pair because the silhouette score s_x of image I_x is low. Then, a human confirms them as a true positive pair, resulting in the merging of C_o and C_{nn}, as indicated by the dashed arrow.

the squared similarity scores. That is, the probability of an unchecked positive image pair (I_x, I_y) being chosen, denoted P_{val}, is given by

$$P_{\text{val}} \propto \frac{1}{\Delta(I_x, I_y)^2}, \quad \text{where } (I_x, I_y) \in S_{\text{ncheck}}. \tag{10}$$

The variable $\mathcal{S}_{\text{ncheck}}$ is a set of positive image pairs that have not been checked by a human within the MST set in a given clustering. An example of a misidentified positive image pair due to a low initial positive threshold at the core clustering stage is shown in Fig. 5(a). In the subsequent clustering verification phase, the algorithm selects this pair for human validation, where it is confirmed as a false positive, and then separates them.

If no false positive image pairs are found, the algorithm continues the core clustering generation stage, in Algorithm 1 Line 11-24. In this stage, the main objective is to adaptively reduce the positive threshold to automatically identify additional positive image pairs. Once a clustering is established during the initialization stage, the expanded silhouette score, which is a reliable measure for evaluating clustering quality, is used to determine a new positive threshold. This threshold reflects the average similarity scores between positive images and their corresponding clusters in the current clustering. The process iterates until the positive threshold can no longer be reduced or misidentified cases are detected by the HITL-based verification method.

Recall that, $\bar{\Delta}'_o$ reflects the mean similarity scores of images with their own clusters in the current clustering. Thus, $\bar{\Delta}'_o$ serves as a reasonable update for the next positive threshold. Meanwhile, we impose a lower limit on the positive threshold to prevent it from approaching the similarity scores of typical nonzero negative image pairs. That is, the positive threshold

Algorithm 2: HITL-based clustering verification

Input : Set of unchecked positive image pairs, $\mathcal{S}_{\text{ncheck}}$
Number of human decisions for validation, N_{val}
Identity matrix for image pairs,
$\mathcal{Q}(I_x, I_y) = (\text{X}, \text{X})|(\text{X}, +)|(\text{T}, +)|(\text{F}, +)|(\text{T}, -)$
//(X,X):(unverified,unknown), (X,+):(unverified,pos.),
//(T,+):(true,pos.), (F,+):(false,pos.), (T,-):(true,neg.)
Output: Updated $\mathcal{Q}, \mathcal{S}_{\text{ncheck}}$

```
for i = 1 : N_val do
    Select a pair (I_x, I_y) ∈ S_ncheck randomly using (10)
    if (I_x, I_y) is human-identified as the same then
        Q(I_x, I_y) ← (T, +)
    else
        Q(I_x, I_y) ← (F, +)
    S_ncheck ← S_ncheck − (I_x, I_y)
```

T_{pos} at iteration j is set to $\bar{\Delta}'_o$ from the current clustering only if it satisfies the following condition:

$$T_{\text{pos}}(j) = \bar{\Delta}'_o, \quad \text{only if} \quad \Delta_{\text{neg}} < \bar{\Delta}'_o < T_{\text{pos}}(j-1). \tag{11}$$

The variable Δ_{neg} is a value of similarity scores for nonzero negative image pairs at iteration j given by $\exp(\mu_{\text{neg}} + 2\sigma_{\text{neg}})$, where μ_{neg} and σ_{neg} are the mean and standard deviation, respectively, of the logarithm of the squared similarity scores for all nonzero negative image pairs from the current clustering. The value of Δ_{neg} reflects the dominant distribution of negative image pairs. Thus, setting this as the lower bound effectively prevents the positive threshold from becoming too small and mistakenly including negative cases. This adjustment increases the number of algorithmic decisions, reducing the need for human involvement while maintaining high clustering accuracy. However, some true positive image pairs in the dataset have small image similarity scores, making it difficult for the algorithm to identify them automatically. To address this, we introduce a HITL-based clustering growth algorithm in the next subsection.

D. HITL-based Clustering Growth

Our HITL-based core clustering generation and clustering verification algorithms produce a core clustering where most true positive image pairs are automatically identified, only requiring minimal human decision-making. These positive pairs typically exhibit large image similarity scores, enabling confident algorithmic decisions. However, some true positive image pairs have small similarity scores, making them indistinguishable from true negative pairs. For this, we propose a HITL-based clustering growth algorithm, shown in Fig. 4. In this stage, we aim to determine an appropriate adaptive questionable threshold that allows the algorithm to identify indistinguishable pairs as potential positives for human confirmation. Recall that the silhouette score measures how well an image fits within its assigned cluster by comparing it to its nearest neighbor cluster. In our setting, images assigned to the same cluster exhibit strong similarity, so a small silhouette score indicates that an image may also belong to its nearest neighbor cluster, particularly when both $\bar{\Delta}_o$ and $\bar{\Delta}_{\text{nn}}$ are large. Therefore, an adaptive questionable threshold is defined

Algorithm 3: HITL-based clustering growth

Input : Similarity score matrix for an image dataset $\mathcal{I}$
Clustering, $\mathcal{C}$
Identity matrix for image pairs,
$\mathcal{Q}(I_x, I_y) = (\mathrm{X}, \mathrm{X})|(\mathrm{X}, +)|(\mathrm{T}, +)|(\mathrm{F}, +)|(\mathrm{T}, -)$
//(X,X):(unverified,unknown), (X,+):(unverified,pos.),
//(T,+):(true,pos.), (F,+):(false,pos.), (T,-):(true,neg.)
Set of unchecked positive image pairs, $\mathcal{S}_{\mathrm{ncheck}}$
Output: Updated $\mathcal{C}$

repeat
  Compute silhouette score s_x, $\forall\ I_x \in \mathcal{I}$, using (4)
  Calculate questionable threshold T_Q using (12)
  Init. a set of potential positive cluster pairs, $\mathcal{S}_{\mathrm{ppos}} \leftarrow \{\}$
  for each image $I_x \in \mathcal{I}$ if $s_x \leq T_Q$ **do**
    Find its own and nearest neighbor clusters, C_o, C_{nn}
    if $(C_o, C_{\mathrm{nn}}) \notin \mathcal{S}_{\mathrm{ppos}}$ **then**
      $\mathcal{S}_{\mathrm{ppos}} \leftarrow \mathcal{S}_{\mathrm{ppos}} \cup (C_o, C_{\mathrm{nn}})$
  for each cluster pair $(C_p, C_q) \in \mathcal{S}_{\mathrm{ppos}}$ **do**
    Determine image pair (I_x, I_y), s.t.,
    $\Delta(I_x, I_y) = \max\{\Delta(I_p, I_q) \mid I_p \in C_p, I_q \in C_q\}$
    if (I_x, I_y) is human-identified as the same **then**
      $\mathcal{Q}(I_x, I_y) \leftarrow (\mathrm{T}, +)$
    else
      $\mathcal{Q}(I_x, I_y) \leftarrow (\mathrm{T}, -)$
  /*Clustering verification*/
  Update $\mathcal{Q}, \mathcal{S}_{\mathrm{ncheck}}$ using Algorithm 2
  Generate $\mathcal{C}$ by including positive image pairs (I_x, I_y)
  where $\mathcal{Q}(I_x, I_y) == (\mathrm{X/T}, +),\ \forall\ I_x, I_y \in \mathcal{I}$
until $\mathcal{C}$ does not change;

relative to a representative small silhouette score. Accordingly, given a clustering, a questionable threshold, denoted T_Q, is defined using (4), where $\exp(\mu_{\mathrm{pos}} - \sigma_{\mathrm{pos}})$ and Δ_{min} correspond to $\bar{\Delta}_o$ and $\bar{\Delta}_{\mathrm{nn}}$, respectively. That is,

$$T_Q = 1 - \frac{1}{\exp(\sigma_{\mathrm{pos}})}. \tag{12}$$

The questionable threshold is therefore adaptively updated according to the clustering state.

If the silhouette score of an image falls below this threshold, the algorithm identifies this image together with an image from its nearest cluster as a potential positive pair for human confirmation. A human then confirms whether the image pair belongs to the same individual (true positive) or from different individuals (true negative). An example of potential positive image pairs identified by this algorithm is shown in Fig. 5(b), resulting in the two clusters being merged. After all potential positive pairs are confirmed by a human annotator, the clustering is updated based on newly identified positive pairs and the adaptive questionable threshold is updated accordingly. This iterative process continues until no further new positive image pairs are identified. Additionally, to mitigate the risk of incorrect algorithmic decisions, our clustering verification procedure continues to periodically request human validation for a random subset of unchecked positive image pairs during clustering growth. The pseudocode is shown in Algorithm 3.

IV. Results

Our HITL-based individual leopard identification solution is evaluated using a dataset of camera trap images of African leopards provided by *Panthera* [8], an organization dedicated to wild cat conservation. The images were collected over several years from many camera traps strategically placed throughout the leopards' habitat. This dataset includes 1,555 images (N) taken from 676 individual leopards (K), manually identified by *Panthera* animal experts based on distinct spot patterns on each leopard's left and right sides. The distribution of images per individual is shown in Fig. 6. Several factors contribute to the dataset's relatively large individual-to-image ratio (K/N), which means that many leopards appear in only a small number of images. These factors include the elusive nature of African leopards, the limited number of camera traps distributed across their vast habitat, and the capture of a single image per camera trap activation. In addition, because of varying leopard poses and environmental factors such as different weather conditions or occlusion by vegetation, some true positive pairs become indistinguishable from true negative pairs. This is reflected in the distribution of nonzero image similarity scores for true positive and true negative pairs, as shown in Fig. 7. The overlap between true positives with small similarity scores and true negatives with large similarity scores emphasizes the need for human involvement in individual leopard identification. This challenge is effectively addressed by our HITL-based algorithm.

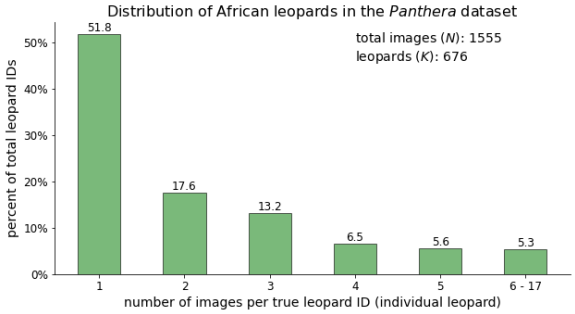

Fig. 6. The figure illustrates the distribution of image number per individual leopard, provided by *Panthera*. Due to the unique spot patterns on each side of a leopard and only a single image taken per activation for camera traps, more than half of the true leopard IDs are represented by a single image.

Our HITL-based individual leopard identification algorithm begins by using an open-source Python implementation of a pretrained Mask R-CNN model [9] to segment leopard bodies in images. It then employs the Hotspotter application [10] to compute image similarity scores for all segmented image pairs, generating an $N \times N$ image similarity score matrix as the input for the subsequent stages. This combination of the pretrained Mask R-CNN model and Hotspotter provides effective leopard object detection and viewpoint-invariant local feature matching for unlabeled leopard camera trap images. It is important to point out that our proposed HITL-based individual identification algorithm requires only a similarity score matrix as input. The choice of objection segmentation approaches and feature detection and matching methods can be selected based on the dataset characteristics. Next, the algorithm applies our HITL-based core clustering generation and clustering verification techniques to construct a core clustering and determine the corresponding number of clusters. Finally,

TABLE I
ALGORITHM ACCURACY ($K = 676, N = 1555$)

Approach[1]	human decisions	clustering accuracy	predicted k	correct clusters[3] (images)[4]	unfound[5] true positives
1-nn	1189	0.909	747	620 (75.7%)	7.02%
2-nn	2284	0.987	693	659 (94.1%)	1.37%
3-nn	3472	0.989	692	660 (95.0%)	1.16%
4-nn	4693	0.989	692	660 (95.0%)	1.16%
HITL[2]	**∼609**	**0.986**	**696**	**657 (93.6%)**	**1.66%**

[1] m-nn: each image and its top-m most similar images based on similarity scores
[2] HITL: 99.9% of cases achieve identical results with 609 human decisions on average
[3] correct cluster: contains ALL images of a single leopard and no other leopard images
[4] (images): the percentage of dataset images belonging to correct clusters
[5] unfound: the percentage of same-leopard pairs not clustered together by the algorithm

to identify indistinguishable image pairs that are challenging for automated approaches, the algorithm performs our HITL-based clustering growth and verification procedures to refine the clustering and update the cluster number, further improving identification accuracy.

The results of applying our algorithm to the *Panthera* dataset are summarized in Table I, along with a comparison against a baseline in which a human checks the m-nearest neighbors (m-nn) [10]. Human decisions are simulated by an infallible expert. In the m-nn baseline, a human compares each image with its m most similar images, based on the highest similarity scores computed by Hotspotter, to determine whether they belong to the same identity. It is important to note that, unlike the automated termination mechanism in our algorithm, the m-nn method requires manual specification of the number of nearest neighbors to be checked. As is known, increasing the number of nearest neighbors m generally improves identification accuracy, but substantially increases human effort. This makes the m-nn method difficult to balance accuracy and efficiency. We terminate the baseline m-nn approach at 4 because this is the first value at which no additional true positive image pairs are found. Clustering accuracy is evaluated using the adjusted mutual information score [27], which quantifies the agreement between a generated clustering and the ground truth of the dataset. Our algorithm achieves clustering accuracy comparable to the 2-nn technique while significantly reducing human involvement. In most experiments, our approach converges to the same clustering with an accuracy of 0.986, compared to 0.987 for the 2-nn method, while impressively reducing the number of human decisions by 77.3%, from 2284 to just 609 image pairs on average. Misidentification is unlikely to occur because human input is used to determine an appropriate initial positive threshold, enabling the algorithm to make accurate positive identifications. Nevertheless, for both algorithms, a small number of true positive image pairs remain indistinguishable (86 unfound pairs for our method and 71 for the 2-nn technique, where an unfound pair refers to a pair of images from the same leopard that are not clustered together by the algorithm). This is because true positive image pairs captured from significantly different viewpoints often obtain few similar regions, resulting in small similarity scores that make identification challenging. Identifying these pairs would necessitate a significant increase in human involvement due to the large number of true negatives that would also require verification. In addition, we compared our HITL-based identification algorithm with a fully automated algorithm [7]. The proposed HITL-based algorithm produces a clustering with 696 clusters and achieves an accuracy of 0.986, compared with the fully automated algorithm's clustering of 718 clusters with an accuracy of 0.957. Additionally, to avoid misidentifying indistinguishable image pairs, the HITL-based verification and clustering growth stages are employed. Such errors commonly occur in fully automated algorithms.

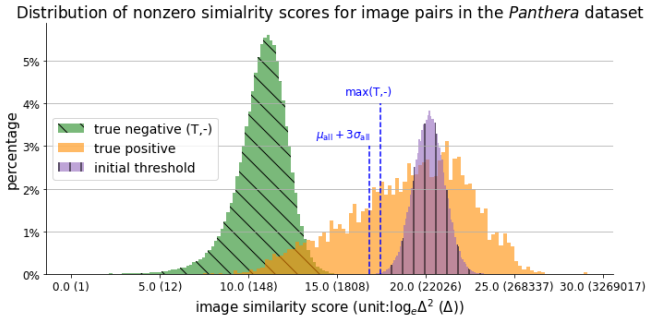

Fig. 7. This figure shows the probability density functions (PDFs) of nonzero image similarity scores for true positive and true negative image pairs in the *Panthera* dataset. To resolve cases where true positive and true negative image pairs have overlapping similarity scores, during the clustering growth stage, our algorithm identifies them as potential positive pairs and requests human confirmation of whether they belong together. Additionally, the PDF of initial positive thresholds is plotted based on 100,000 random executions. Only 0.1% of initial thresholds fall below the maximum similarity score of true negative pairs, indicating a small risk of misidentification during the core clustering generation. Fortunately, this risk is mitigated by our clustering verification algorithm, which uses human validation to correct such misidentified pairs.

To elaborate on the effectiveness of our HITL-based individual leopard identification algorithm, we first demonstrate that it generates a reasonable initial positive threshold. This allows the algorithm to robustly identify image pairs as positive when their similarity scores exceed this threshold. Fig. 7 presents the distribution of initial positive thresholds based on 100,000 random executions. Most thresholds exceed the maximum image similarity scores of true negative pairs (negative maximum), ensuring that all automatically identified positive pairs are correct. Only 0.1% of thresholds fall below

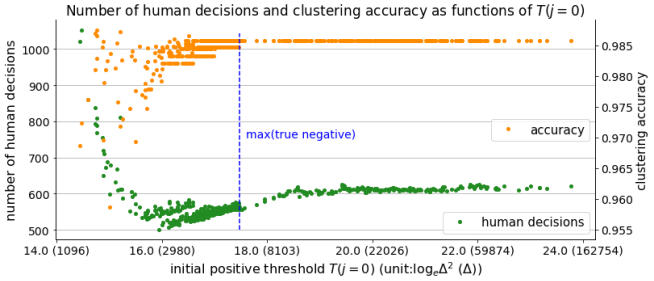

Fig. 8. This figure shows the number of human decisions and clustering accuracy as functions of the initial positive thresholds. To assess the robustness of our algorithm across an extreme range of initial thresholds, we present results from 500 random cases with an initial threshold below the maximum similarity score of true negative pairs (negative maximum) and 500 cases above it. Initial thresholds below the negative maximum occur only 0.1% of the time. At very low initial thresholds, the number of human decisions increases sharply to correct false positive image pairs misidentified by the algorithm. Once the initial threshold exceeds the negative maximum, our algorithm consistently produces the same clustering, on average requiring human decisions for only 609 image pairs (0.05% of the total pairs). This efficiency results from our adaptive positive threshold mechanism that enables the algorithm to autonomously identify most true positive pairs.

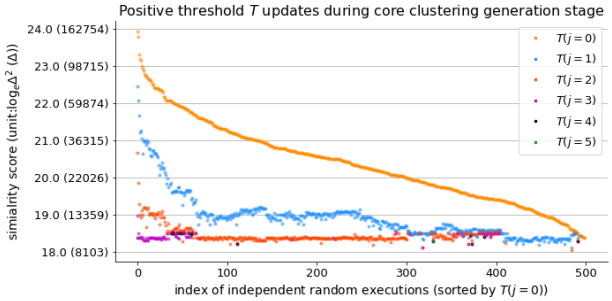

Fig. 9. This figure shows the trend of positive threshold updates starting from various initial values during the core clustering generation process. The results are from randomly selected executions where the initial thresholds were above the maximum similarity score of true negative pairs (accounting for 99.9% of all cases) and are arranged in descending order of their initial values. The points on the y-axis represents the reduction in the positive threshold during one execution. This figure demonstrates that the thresholds consistently converge to a narrow range, regardless of their initial values. This convergence occurs because our algorithm adaptively reduces the positive threshold based on similarity scores of positive pairs within the current clustering. The final threshold enables a large number of true positive pairs to be automatically identified.

the negative maximum; fortunately, our clustering verification algorithm periodically performs human validation on these automatically identified positive pairs, efficiently correcting misidentifications.

To evaluate the reliability of our identification algorithm under different initial thresholds with limited human involvement, we explore how the number of human decisions and clustering accuracy vary with changes in the initial positive threshold, as presented in Fig. 8. To highlight the robustness of our algorithm, the results are shown from two subsets out of 100,000 random executions: one subset of 500 with an initial positive threshold below the negative maximum and the other with 500 above it. It is important to note that the first subset occurs in only 0.1% of all executions, whereas the second accounts for 99.9%. For thresholds below the negative maximum, most true positive pairs are automatically identified due to a small initial positive threshold. However, some true negative pairs with relatively large similarity scores are inevitably misidentified as positive. To address this, our clustering verification algorithm periodically selects some unchecked positive image pairs with small similarity scores for human validation, ensuring the correctness of autonomous decisions. For thresholds above the negative maximum, our core clustering generation algorithm adaptively updates the positive threshold, enabling a large number of image pairs to be automatically identified. As a result, the subsequent clustering growth procedure identifies only a small number of potential positive pairs requiring human confirmation. This process consistently yields the same clustering with an accuracy of 0.986, while requiring human decisions for an average of just 609 image pairs (0.05% of the total pairs).

To understand why our algorithm produces consistent clusterings under large initial positive thresholds with limited human involvement, we closely investigate how the core clustering generation algorithm adaptively reduces the positive threshold, as illustrated in Fig. 9. The results are based on ran-

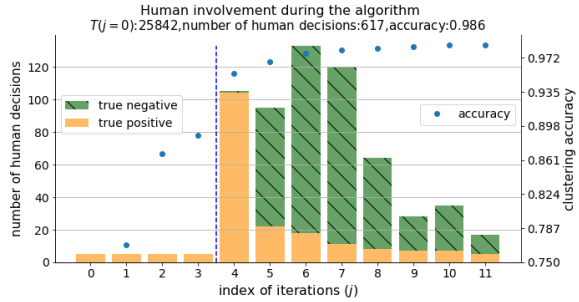

Fig. 10. This figure illustrates the trend of human involvement in distinguishing true positive and true negative pairs, along with the progression of clustering accuracy throughout the algorithm. The results are based on a typical execution where the initial positive threshold is above the maximum similarity score of true negative pairs. In the first four iterations, human involvement is used to determine the initial positive threshold and to validate a subset of automatically identified positive pairs. Afterward, the process falls into the clustering growth and verification stage, as shown on the right side of the dashed line. In this phase, a human confirms the consistency of potential positive pairs and validates unchecked pairs from the earlier stage. The clustering is updated with newly identified positive pairs until no additional pairs can be confirmed. In the final iteration, all verified true positive pairs are from the unchecked positive set, rather than new ones.

domly selected executions, where the initial positive thresholds are above the negative maximum (as occurs in 99.9% of all executions). Recall that the new positive threshold is updated to $\bar{\Delta}'_o$, which effectively reflects the mean similarity score of positive image pairs in the current clustering. Therefore, regardless of the initial value, the positive threshold is adaptively reduced based on the current clustering performance until convergence. The points arranged vertically in the figure represent the positive threshold reduction process during one execution. This adaptive mechanism enables the automatic identification of most true positive pairs with large similarity scores, such that the final positive thresholds consistently converge within a narrow range, while significantly reducing the requirement for human confirmation in the subsequent clustering growth stage.

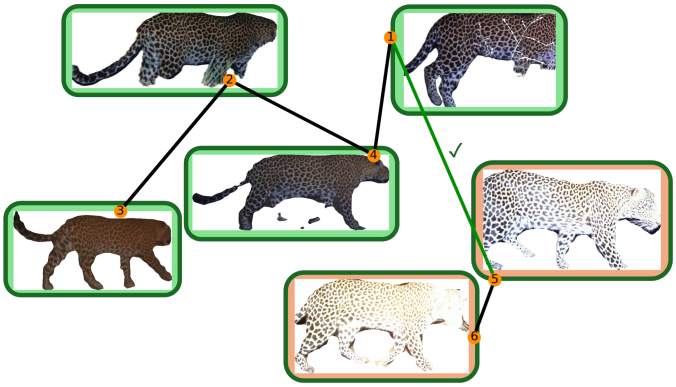

Fig. 11. This figure illustrates how the clustering growth stage can identify clusters eligible for merging. The graph shows one leopard represented by six images, labeled 1 to 6. During the core clustering generation stage, the algorithm initially identifies two separate clusters (images within the same cluster have the same interior border color): one contains Images 1 to 4, and the other includes Images 5 and 6. The edges indicate the layout of the minimum spanning trees for these clusters. In the clustering growth stage, the algorithm identifies the image pair (1,5) as a potential positive, and a human confirms this as a true positive, denoted by a check mark. As a result, all images are merged into a single cluster, indicated by an exterior border color.

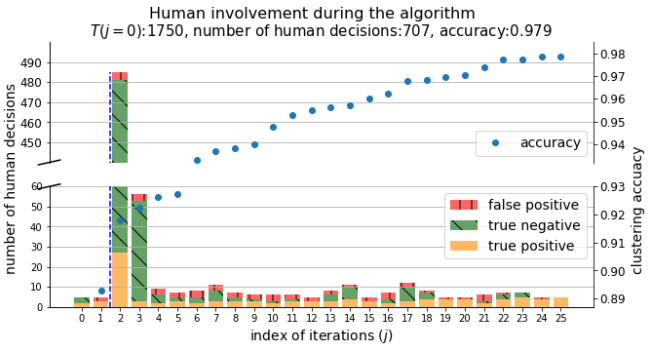

Fig. 12. This figure illustrates how human involvement in distinguishing false positive, true positive, and true negative pairs evolves throughout the algorithm's execution, alongside the progression of clustering accuracy. The results are from a rare case with a very low initial positive threshold (which occurs in only 0.1% of executions). The algorithm begins with the core clustering generation and verification stage, shown on the left side of the dashed line. Due to misidentifications corrected by human validation, the process transitions directly into the clustering growth and verification stage. In this phase, a human confirms potential positive pairs and periodically corrects misidentified positive pairs, further improving identification accuracy. The process continues until no additional false positives are found and no new positive pairs are confirmed. In the final iteration, all verified true positives come from the remaining set of unchecked positives, rather than new ones.

To gain insight into the role of human involvement in our individual identification algorithm, Fig. 10 shows the number of human decisions made for identifying true positive and true negative pairs in each iteration, based on a representative execution where the initial positive threshold is above the negative maximum. At initialization, human input is used to determine the initial positive threshold, allowing the algorithm to confidently make positive identifications for image pairs with large similarity scores. In the following three iterations, human involvement focuses on validating a random subset of automatically identified positive pairs, verifying the correctness of autonomous decisions. At the same time, successful validations allow the positive threshold to be iteratively lowered based on the similarity scores of positive image pairs in the current clustering. Once the threshold can no longer be reduced, the algorithm moves into the clustering growth and verification stage at iteration 4. In the remaining iterations, a human confirms the consistency of potential positive pairs identified by the algorithm, resolving ambiguous cases where true positive pairs are challenging for automated identification. Simultaneously, human validation is periodically performed. The final clustering accuracy reaches 0.986, a significant improvement over 0.888, achieved after the core clustering generation and verification stages alone.

To illustrate how the clustering growth stage successfully identifies all images of a true leopard ID with minimal human involvement, Fig. 11 presents an example where an individual leopard is represented by six images (labeled 1 to 6). The image layout is plotted according to their minimum spanning trees [25], with edge lengths inversely proportional to their image similarity scores. Specifically, our core clustering generation algorithm initially forms two separate clusters based on large similarity score: one cluster includes Images 1 to 4; the other contains Images 5 and 6. After the clustering growth stage, all images are successfully merged into a single unified cluster. This result is achieved by the algorithm identifying the image pair (1,5) as a potential positive because the silhouette score of Image 1 is greater than the questionable threshold and Image 5 has the highest similarity score within its nearest neighbor cluster. A human then confirms that they belong to the same ID, and the clusters are merged.

Finally, we explore the rare case (only occurring in 0.1% of experiments) of a very low initial positive threshold that requires human involvement to correct misidentification. Fig. 12 presents the number of human decisions made in identifying false positive, true positive, and true negative pairs at each iteration for this case. Specifically, at the first iteration $(j = 1)$, a human corrects cases misidentified by the algorithm. Consequently, the process immediately proceeds to the clustering growth and verification stage, without adaptively lowering the positive threshold. Recall that under a low initial threshold, most true positive pairs are automatically identified, while some true negatives with relatively large similarity scores are inevitably misidentified. Because of this, a human in this stage confirms most potential positive pairs as true negatives, while also periodically correcting misidentifications from the core clustering generation stage. The process continues until a randomly selected subset of unchecked positive pairs is fully verified as true positives, and no new true positive pairs are found. As a result, even in this unlikely case of a low threshold, the final clustering has an accuracy of 0.979 and requires human involvement for only 707 image pairs.

Fig. 13 illustrates how, in this rare case, human involvement during the clustering verification stage corrects misidentified image pairs produced by the algorithm. It presents five images (labeled 7 to 11) from four individual leopards: Images 7 and 8 belong to the same leopard, while Images 9 to 11 each represent different individuals. Initially, the algorithm incorrectly identifies these five images as positives due to a

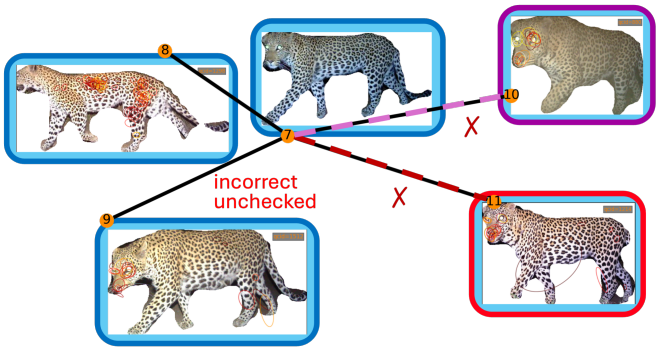

Fig. 13. This figure illustrates how human validation corrects misidentifications made by the algorithm. It shows five images (labeled 7 to 11) from four individual leopards: Images 7 and 8 belong to the same leopard, while Images 9 to 11 each represent different individuals. Due to a very low initial positive threshold, occurring in only 0.1% of executions, the algorithm incorrectly groups all five images into a single cluster, indicated by the same interior border color. The edges represent the layout of the minimum spanning tree within this cluster. During the clustering verification stage, image pairs (7, 10) and (7, 11) are selected for human validation, corrected as false positives, and subsequently isolated, indicated with an "X". However, because our verification mechanism is not designed to exhaustively review all automatically identified positive pairs, the misidentified pair (7, 9) remains unchecked. As a result, Images 7 to 9 remain in the same cluster, while Images 10 and 11 are correctly isolated. (Images in the same cluster after the validation stage have the same exterior border color.)

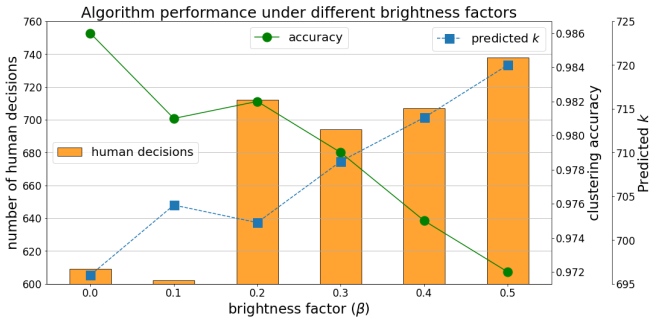

Fig. 14. This figure illustrates the performance of the proposed HITL-based identification algorithm under different values of the brightness factor β. The evaluation considers the number of required human decisions, clustering accuracy measured by adjusted mutual information score, and the estimated number of animal identities (predicted k).

very low initial positive threshold, an event that occurs rarely in our method. Fortunately, this is exactly the kind of issue the clustering verification stage was designed to address. The algorithm randomly selects a subset of unchecked positive pairs for human validation using a weighted probability distribution that favors smaller similarity scores. As a result, image pairs (7,10) and (7,11) are selected and verified as false positives, leading to the isolation of Images 10 and 11. However, because our verification mechanism is not designed to exhaustively verify all automatically identified positive pairs, the misidentified pair (7,9) remains unverified and is incorrectly grouped in the final clustering.

To evaluate the effectiveness of our HITL-based identification algorithm in balancing maximal autonomous identification with minimal human involvement, we conduct ablation experiments by separately disabling the core clustering generation stage and the clustering growth stage. When the core clustering generation stage is disabled, the number of required human decisions increases significantly by 26.4%, from 609 to 770. This is because many positive pairs that would otherwise be automatically identified must instead be manually confirmed. Meanwhile, because of the clustering growth stage, the clustering accuracy remains high at 0.986 with the estimated number of identities (k) of 696. In contrast, when the clustering growth stage is disabled, the identification accuracy decreases by 10.0% to 0.888, while the predicted k increases to 838. This occurs because positive pairs located in the overlap region between the positive and negative similarity score distributions arex indistinguishable to the algorithm. In this setting, only 20 human decisions are required during the verification stage of core clustering generation.

Furthermore, to evaluate the robustness of the HITL-based identification algorithm under varying imaging conditions, we conduct experiments using similarity score matrices generated from overexposed and underexposed leopard images. Illumination variation is a common source of appearance change in real world environments. Specifically, 1/3 of the datasets is randomly selected for underexposure, 1/3 for overexposure, and 1/3 is kept unchanged. For overexposed images, each pixel value is scaled by $1+\beta$, whereas for underexposed images, each pixel value is scaled by $1-\beta$. The brightness factor β varies from 0.1 to 0.5. The corresponding results are presented in Fig. 14. The results demonstrate that the HITL-based identification algorithm remains robust even under extreme lighting conditions (β=0.5). Compared with the original condition, this extreme case results in only a 1.4% decrease in clustering accuracy, from 0.986 to 0.972, and a relatively small increase in the number of required human decisions, from 609 to 683.

It is important to note that the above results regarding human decisions are simulated using an infallible expert, i.e., based on ground truth labels. To evaluate the robustness of our algorithm to occasional practical human annotation errors, we conduct experiments that simulates each type of annotation error (false negative or false positive) at the corresponding stage of the framework. The results are summarized in Table II. Specifically, (a) During the initialization stage, an expert annotator assisted by the Hotspotter results is unlikely to misidentify an image pair. This is because the selected pairs usually have large similarity scores, corresponding to large and clearly visible matching regions, and only a limited number of annotations are required at this stage. Moreover, if a selected pair appears ambiguous, the annotator may request another randomly selected pair. Consequently, an appropriate positive threshold can still be established for subsequent autonomous identification. (b) During the verification stage, an expert annotator validates the consistency of automatically identified positive pairs. Annotation errors are unlikely at this stage because only a limited number of human annotations are required, depending on the number of iterations. In addition, the verified pairs are selected from automatically identified positive pairs whose similarity scores are generally relatively large. The impact of annotation errors depends on the type of misidentification, as discussed below. If a false negative occurs during the verification stage, as shown in Table II (b-1), the

TABLE II
ALGORITHM PERFORMANCE UNDER ANNOTATION ERRORS

Experiment	error (%)	human decisions	clustering accuracy	predicted k
a representative case (occur 99.9%)				
(b-1) (F,-) in verification	1 (1.8%)	770	0.985	697
(c-1) (F,-) in growth	6 (1.0%)	687	0.982	702
(c-1) (F,-) in growth	12 (2.0%)	677	0.978	708
(c-1) (F,-) in growth	18 (3.0%)	671	0.978	708
(c-2) (F,+) in growth	6 (1.0%)	679	0.982	689
(c-2) (F,+) in growth	12 (2.0%)	685	0.977	683
(c-2) (F,+) in growth	18 (3.0%)	703	0.977	676
a rare case (occur 0.1%): $T_{\text{pos}} < \max(\Delta_{\text{neg}})$				
(b-2) (F,+) in verification	0	706	0.983	692
(b-2) (F,+) in verification	1 (1.0%)	720	0.984	693

positive threshold updating process in the core clustering generation stage may terminate prematurely, leading to additional human annotation during the clustering growth stage. In our experiment, when a false negative occurs during the verification step of the first iteration in the core clustering generation stage, the number of human decisions increases by 26.4%, from 609 to 770, while the accuracy remains 0.985. A false positive occurring during the verification stage is extremely rare, as shown in Table II (b-2). This situation can occur only when an exceptionally small positive threshold is established, which appears in only 0.1% of experiments. Because all verified positive pairs are selected from autonomously identified positive pairs, the verification stage can detect and correct an error only when the autonomous identification itself is incorrect. In our experiment, although one false positive occurs, the algorithm still achieves an accuracy of 0.984. (c) During the clustering growth stage, annotation errors are more likely because a relatively large number of human annotation are required. In addition, the expert annotator need confirm whether the indistinguishable image pairs identified by the algorithm belong to the same identity or not. These pairs usually have relatively small similarity scores, corresponding to limited visual matching regions. If a false negative occurs, as shown in Table II (c-1), image pairs belonging to the same individual are inevitably separated into multiple clusters and incorrectly identified as different individuals. In our experiment, when 3% of the human annotations in the clustering growth stage contain false negatives, the number of human decisions increases by 10.0%, while the accuracy decreases by 1.0%. If a false positive occurs, as shown in Table II (c-2), image pairs from different individual are inevitably incorrectly assigned to the same identity. This can further introduce additional false positive pairs, resulting in a more severe effect. Specifically, images involved in false positive pairs may satisfy the conditions for potential positive pairs with their nearest neighbor clusters. If these potential positive pairs are subsequently confirmed as true positives, a larger false positive cluster may gradually formed. In our experiment, when 3% of the human annotations in the clustering growth stage contain false positives, the number of human decisions increases by 15.5%, while the accuracy decreases by 1.0%.

To estimate the practical human annotation time, we refer to the study by Craig Parham [28], who tracked six participants and reported that expert annotators assisted by visualization-based matching typically require approximately 6-16 seconds per image pair. Meanwhile, annotation time varies considerably depending on the annotator's expertise, familiarity with the task, and the difficulty of the image pair. Accordingly, our framework requires approximately 1.69 hours for a human annotator to confirm around 609 image pairs, assuming an average annotation time of 10 seconds per image pair as a midpoint estimate within the reported range. Specifically, during the initialization stage, the annotator confirms the consistency of 10 image pairs, requiring approximately 1 minute. During the core clustering generation stage, the annotator verifies the consistency of 5 automatically identified positive image pairs in each iteration, typically requiring approximately 3 minutes in total. The remaining manually confirmed image pairs are generated during the clustering growth stage, where verification is performed simultaneously.

Finally, our HITL-based individual identification algorithm can effectively generalize to semi-supervised clustering problems with pairwise characteristics involving overlapping distributions. To further validate its generalization, we construct experiments on several simulated distribution and an additional distribution derived from a public camera trap leopard image dataset. First, for the same *Panthera* dataset [8], we maintain the negative distribution and modify the mean and standard deviation of the similarity score distribution for positive pairs. When the mean decreases or the standard deviation increases in positive threshold, the overlap region between the positive distribution and negative distribution becomes larger. This indicates that more human involvement is required for indistinguishable pairs. For example, when the mean decreases by 7.3%, the number of human decisions increases by 17.4%, while the accuracy decreases to 0.982. Similarly, when the standard deviation increases by 19.3%, the number of human decisions increases by 8.5%, while the accuracy decreases to 0.984. Furthermore, our algorithm also works effectively on a public camera-trap leopard image dataset provided by

Wildbook [29]. This dataset contains 590 individual leopards from 3251 images. Our algorithm achieves the accuracy of 0.950 while requiring 2490 human annotations. Additionally, our proposed automated termination mechanism effectively terminate the process, thereby avoiding excessive expert annotation time on positive pairs with extreme small similarity scores. In particular, these pairs are also easily misidentified by human annotators because they contain limited visible matching region.

V. Conclusion

In this paper, we propose an efficient HITL-based algorithm for identifying individual leopards in a dataset of unlabeled camera trap images of African leopards, provided by *Panthera*. While manual identification achieves the highest identification accuracy, it is extremely time-consuming. Automated methods require no human effort but often struggle with ambiguous true positive and true negative image pairs, leading to frequent misidentifications and unacceptably low accuracy. To address this, our algorithm incorporates human involvement throughout the process: determining an initial positive threshold for confident automatic identification (HITL-based core clustering generation), validating positive pairs identified by the algorithm (HITL-based clustering verification), and confirming the consistency of potential positive pairs determined by the algorithm (HITL-based clustering growth). This HITL strategy ensures that our algorithm attains high identification accuracy while requiring minimal human involvement. Additionally, to better estimate how well an image fits its own cluster without dataset labels, we propose an expanded definition of the silhouette score.

Acknowledgments

The authors would like to thank *Panthera* for providing the camera trap data set.

References

[1] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. Van Langevelde, T. Burghardt *et al.*, "Perspectives in machine learning for wildlife conservation," *Nature communications*, vol. 13, no. 1, pp. 1–15, 2022.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.

[3] M. Vidal, N. Wolf, B. Rosenberg, B. P. Harris, and A. Mathis, "Perspectives on individual animal identification from biology and computer vision," *Integrative and comparative biology*, vol. 61, no. 3, pp. 900–916, 2021.

[4] A. Algasov, E. Nepovinnykh, T. Eerola, H. Kälviäinen, C. V. Stewart, L. Otarashvili, and J. A. Holmberg, "Understanding the impact of training set size on animal re-identification," *arXiv preprint arXiv:2405.15976*, 2024.

[5] C. V. Stewart, J. R. Parham, J. Holmberg, and T. Y. Berger-Wolf, "The animal ID problem: continual curation," *arXiv preprint arXiv:2106.10377*, 2021.

[6] R. Shinoda and K. Shiohara, "Petface: A large-scale dataset and benchmark for animal identification," in *European Conference on Computer Vision*. Springer, 2024, pp. 19–36.

[7] C. Guo, A. Miguel, and A. A. Maciejewski, "Automatic identification of individual african leopards in unlabeled camera trap images," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 2460–2471, 2025.

[8] Panthera, "An image dataset of wild african leopards," 2019, https://panthera.org/.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[10] J. P. Crall, C. V. Stewart, T. Y. Berger-Wolf, D. I. Rubenstein, and S. R. Sundaresan, "Hotspotter — patterned species instance recognition," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, 2013, pp. 230–237.

[11] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 337–33 712.

[12] E. Nepovinnykh, I. Chelak, T. Eerola, V. Immonen, H. Kälviäinen, M. Kholiavchenko, and C. V. Stewart, "Species-agnostic patterned animal re-identification by aggregating deep local features," *International Journal of Computer Vision*, pp. 1–16, 2024.

[13] O. Moskvyak, F. Maire, F. Dayoub, A. O. Armstrong, and M. Baktashmotlagh, "Robust re-identification of manta rays from natural markings by learning pose invariant embeddings," in *2021 Digital Image Computing: Techniques and Applications (DICTA)*, 2021, pp. 1–8.

[14] V. Čermák, L. Picek, L. Adam, and K. Papafitsoros, "Wildlifedatasets: An open-source toolkit for animal re-identification," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 5941–5951.

[15] G. Perez, D. Sheldon, G. Van Horn, and S. Maji, "Human-in-the-loop visual re-ID for population size estimation," in *European Conference on Computer Vision*. Springer, 2025, pp. 185–202.

[16] E. Bohnett, J. Holmberg, S. P. Faryabi, L. An, B. Ahmad, W. Rashid, and S. Ostrowski, "Comparison of two individual identification algorithms for snow leopards (*Panthera uncia*) after automated detection," *Ecological Informatics*, vol. 77, p. 102214, 2023.

[17] E. Nepovinnykh, T. Eerola, H. Kälviäinen, and I. Chelak, "NORPPA: Novel ringed seal re-identification by pelage pattern aggregation," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2024, pp. 1–10.

[18] E. Nepovinnykh, V. Immonen, T. Eerola, C. V. Stewart, and H. Kälviäinen, "Re-identification of patterned animals by multi-image feature aggregation and geometric similarity," *IET Computer Vision*, 2025.

[19] J. Boulent, B. Charry, M. M. Kennedy, E. Tissier, R. Fan, M. Marcoux, C. A. Watt, and A. Gagné-Turcotte, "Scaling whale monitoring using deep learning: A human-in-the-loop solution for analyzing aerial datasets," *Frontiers in Marine Science*, vol. 10, p. 1099479, 2023.

[20] D. Sani, M. Khurana, and S. Anand, "Active learning for animal re-identification with ambiguity-aware sampling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 46, 2026, pp. 39 182–39 190.

[21] J. P. Crall, *Identifying individual animals using ranking, verification, and connectivity*. Rensselaer Polytechnic Institute, 2017.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[23] M. Perd'och, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 9–16.

[24] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[25] J. C. Gower and G. J. Ross, "Minimum spanning trees and single linkage cluster analysis," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 18, no. 1, pp. 54–64, 1969.

[26] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[27] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. Association for Computing Machinery, 2009, p. 1073–1080.

[28] J. R. Parham, "Animal detection for photographic censusing," Ph.D. dissertation, 2021.

[29] Botswana Predator Conservation Trust, "Panthera pardus csv custom export. retrieved from african carnivore wildbook," 2022, https://lila.science/datasets/leopard-id-2022/.

Cheng Guo (Student Member, IEEE) received the B.S. degree in Electrical Engineering and Automation from the University of Jinan, Jinan, Shandong, China, in 2015. She received the M.S. and Ph.D. degrees in Computer Engineering at Colorado State University, Fort Collins, CO, USA, in 2025. Her research interests include image processing.

Agnieszka Miguel (Senior Member, IEEE) received the B.S. and M.S. degrees in Electrical Engineering from Florida Atlantic University, Boca Raton, FL, USA, in 1994 and 1996, respectively. She received the Ph.D. degree in Electrical Engineering from the University of Washington, Seattle, WA, USA, in 2001. She is currently a Professor and Chair of Electrical and Computer Engineering at Seattle University, Seattle, WA, USA. Her research interests include data compression, image processing, and engineering education.

Anthony A. Maciejewski (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Ohio State University, Columbus, OH, USA, in 1982, 1984, and 1987, respectively. From 1988 to 2001, he was a Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA. He is currently a Professor of Electrical and Computer Engineering at Colorado State University, Fort Collins, CO, USA. His research interests include robotics, high-performance computing, and engineering education.