Human-in-the-Loop Visual Re-ID for Population Size Estimation

Gustavo Perez[1,2], Daniel Sheldon[2], Grant Van Horn[2], and Subhransu Maji[2]

[1] University of California, Berkeley
gperezs@berkeley.edu
[2] University of Massachusetts, Amherst
{sheldon,gvh,smaji}@cs.umass.edu

Abstract. Computer vision-based re-identification (Re-ID) systems are increasingly being deployed for estimating population size in large image collections. However, the estimated size can be significantly inaccurate when the task is challenging or when deployed on data from new distributions. We propose a human-in-the-loop approach for estimating population size driven by a pairwise similarity derived from an off-the-shelf Re-ID system. Our approach, based on nested importance sampling, selects pairs of images for human vetting driven by the pairwise similarity, and produces asymptotically unbiased population size estimates with associated confidence intervals. We perform experiments on various animal Re-ID datasets and demonstrate that our method outperforms strong baselines and active clustering approaches. In many cases, we are able to reduce the error rates of the estimated size from around 80% using CV alone to less than 20% by vetting a fraction (often less than 0.002%) of the total pairs. The cost of vetting reduces with the increase in accuracy and provides a practical approach for population size estimation within a desired tolerance when deploying Re-ID systems.[3]

Keywords: Human-in-the-loop · Re-ID· Importance sampling

1 Introduction

Computer vision-based re-identification (Re-ID) is increasingly being deployed to analyze large image collections for species-level population monitoring [47,56]. The association of individuals across camera sensors provides valuable data about animal movement, population dynamics, geographic distribution, and habitat use, which in turn can inform conservation goals and ecological models [9, 50, 52]. Computer vision can reduce the manual labor associated with individual identification, and there is significant effort from the community to develop tools that facilitate animal Re-ID across a wide range of animal species [1, 5, 7, 15, 55, 59].

We consider the task of estimating population size, that is, the number of individuals, in a collection of images using a Re-ID system. Beyond surveys, this

[3] Code available at: https://github.com/cvl-umass/counting-clusters

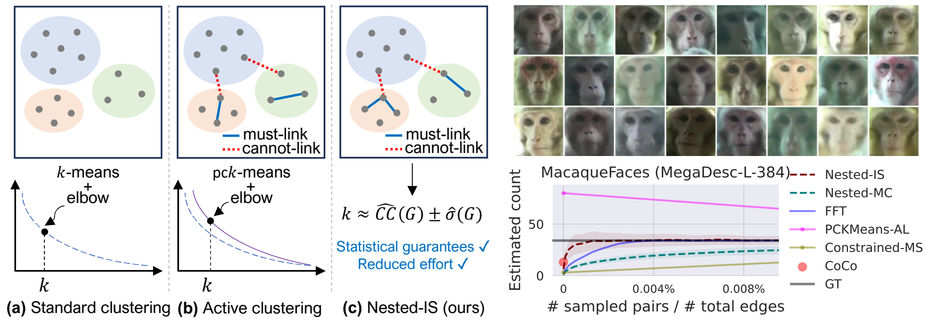

Fig. 1: Estimating Population Size Using a Re-ID System. **(a)** A simple approach involves using k-means clustering on image embeddings derived from the Re-ID system and selecting the optimal k using the "elbow heuristic." **(b)** Active clustering (e.g., pck-means [4]) employs pairwise constraints to enhance clustering accuracy. **(c)** Our method leverages nested importance sampling to produce asymptotically unbiased estimates and confidence intervals on k directly. **(Right)** On the MacaqueFaces dataset [60], our approach (Nested-IS) converges to the true $k = 34$ with fewer constraints than alternative methods, but also provides confidence intervals for the estimate (shown as the shaded red region) for any amount of human feedback.

quantity can inform techniques for category discovery and incremental learning of AI systems [12, 26, 54]. Since the number of individuals is often unbounded, Re-ID is cast as a pairwise classification task and involves predicting whether two images are of the same individual or not. However, estimating the population size from pairwise similarities is non-trivial. One approach is to use a clustering algorithm such as k-means with the "elbow heuristic" [51] to pick k. However, there are many algorithms and heuristics to choose from, making the process subjective. Estimated pairwise similarity can also be a poor approximation to the true similarity, especially when models are deployed out-of-distribution or on challenging tasks where the performance of Re-ID is low. For instance, using k-means with the elbow heuristic on the MacaqueFaces dataset [60] using the state-of-the-art "MegaDescriptor" [55] results in 80 clusters when the true number of individuals is 34, as seen in Fig. 1. Table 1 shows that estimated population size across a variety of animal species and clustering approaches are sometimes off by factor of two or more, which can impact downstream tasks.

Our approach employs statistical estimation to compute the population size using human feedback on pairwise similarity. We develop an estimator based on nested importance sampling, each iteration driven by the approximate pairwise similarity. Feedback in the form of "same" or "not" on a set of sampled edges is used to estimate the number of clusters as well as to provide a confidence interval. We theoretically demonstrate that the estimator is unbiased, i.e., it converges to the true number of clusters in expectation, and an end user can stop screening when the confidence intervals are sufficiently small. Additionally, we contribute a strong baseline based on farthest-first traversal (FFT) to directly estimate the population size, which, although it does not offer statistical guarantees, outperforms alternative active clustering approaches. Our methods can employ any off-the-shelf image similarity and do not involve fine-tuning deep representations, thus can be practical for non-AI experts.

We conduct experiments on seven datasets where the goal is to estimate the number of individuals they contain. These datasets span animal species and exhibit different data distributions, including both short and long-tailed, with varying number of individuals. We also experiment with different image representations from the MegaDescriptor library [55], which provides a unified deep network for various animal Re-ID tasks. The accuracy of the Re-ID system also varies across datasets and are assumed to be unknown, providing a realistic use case for deploying the models. Our approach is more accurate for the same amount of human supervision (sampled pairs) than competing baselines. We also show that the estimator has low bias and produces accurate confidence intervals. In summary, our main contributions are:

- We study the challenges involved in using Re-ID systems for estimating population size across a range of animal Re-ID datasets.
- We propose a novel approach combining nested importance sampling with human-in-the-loop feedback that produces asymptotically unbiased estimates of population size (§ 3).
- We design confidence intervals that provide intuitive feedback for guiding human effort and setting stopping conditions (§ 3.2).
- We report extensive experiments on a benchmark of seven animal Re-ID datasets with different data distributions, demonstrating that our method achieves significantly lower error than strong baselines with similar human effort. In most cases our approach produces estimates close to the true population size using human feedback on less than 0.004% of all pairs (§ 4).
- We use our framework to estimate the number of categories in a dataset for generalized category discovery [12, 26, 54] and measure the impact on clustering accuracy, on animal Re-ID and fine-grained classification (§ 4).

2 Related Work

The task of estimating population size in a dataset is closely related to the problem of grouping individuals using a Re-ID system. However, it is possible to estimate the population size using statistical approaches, such as ours, without fully solving the grouping problem. There is a vast literature in computer vision, especially in face recognition, on building accurate Re-ID systems, as well as using these representations to group individuals. We briefly review these approaches. Our approach focuses on animal Re-ID tasks. This task can be relatively easy for some species that have distinctive patterns but significantly more challenging for others. We review prior work on animal Re-ID and introduce the benchmarks and representations we experiment with.

Re-ID Systems There is significant prior work on person Re-ID based on face and whole-body recognition (see [31] for a more complete survey). Early techniques, such as Eigenfaces [53], have been replaced by modern deep learning approaches [33, 33, 42] trained on large datasets [21, 25]. While significant advances have been made, there are also known issues with poor generalization and bias of Re-ID systems in the presence of demographic shifts [32]. In comparison, techniques for animal Re-ID are less mature. These techniques range from SIFT [35] and Superpoint-based [18] matching approaches (e.g., WildID [7] and HotSpotter [15]) that work relatively well for species with characteristic patterns on their skin or coats (e.g., Zebras and Jaguars), to complex pipelines designed to identify visual characteristics specific to particular species (e.g., whisker patterns of polar bears [1]). Deep learning approaches have also been developed for some species (e.g., Chimpanzees and Bears) but the training datasets are relatively small. Recent work [55] has attempted to train deep learning models across species by consolidating animal Re-ID datasets and has shown that the model generalizes across species and significantly outperforms prior work, including off-the-shelf image representations such as CLIP [44] and DINOv2 [10]. We utilize a set of datasets from their collection (Table 1) and base our population size estimation on the various pre-trained deep networks they provide. Their best-performing model is a Swin-transformer [34] trained with ArcFace loss [17] on the collection of datasets. However, the problem is far from solved – for example, on the WhaleSharkID dataset [24], the performance of the Re-ID system is around 62%, which poses a challenge for population size estimation.

Clustering The population size can be estimated by determining the number of clusters in a dataset. A common approach involves using *clustering algorithms*, such as k-means [22], mean-shift [13], which operate directly on embeddings, or graph-based approaches [8, 19, 49] that incorporate pairwise similarity, and determining the number of clusters based on a *heuristic*. For example, one can compute the within-cluster sum of squares (WCSS) for different values of k in k-means and select the optimal one based on the "elbow heuristic" – the point in the curve where improvements diminish (see Fig. 1). However, there are many clustering algorithms and heuristics available, making the process subjective. We find that population size estimates using these methods can be significantly inaccurate, even with state-of-the-art embeddings (see Table 1). More complex approaches, particularly those incorporating tracking information and sophisticated graph-based clustering (e.g., [19]), have been proposed for grouping in videos. Our primary focus is on image-based approaches using simple clustering algorithms as they are broadly applicable.

Active Clustering Human feedback can be incorporated to improve clustering in various ways. One can fine-tune the deep network using metric learning [27, 30] approaches to improve the underlying similarity using human feedback, for example thorough pairwise [28] and triplet-based [23] learning. However, these methods require significant compute resources and expertise to set various hyperparameters associated with training, and may not be practical for non-experts. Moreover, fine-tuning on uncurated data often encountered in a real-world deployment of the Re-ID system is rarely effective. Hence, we focus on active clustering approaches that incorporate constraints to improve clustering for a fixed embedding. We compare with a constrained k-means algorithms that use

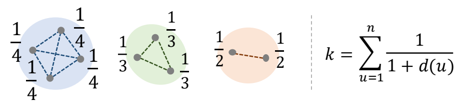

Fig. 2: Counting clusters in a graph. The number of clusters $k = \sum_{u=1}^{n} 1/(1 + d(u))$, where $d(u)$ is degree of node u. In this example $k = 4\times1/4+3\times1/3+2\times1/2 = 3$.

'must-link' and 'cannot-link' constraints within the k-means [4]. Constraints can be selected using a *farthest-first* scheme [4], or by running k-means with a large k and picking constraints to merge the small clusters into larger ones [14]. We also develop a baseline based on the farthest-first traversal (FFT) scheme that is directly aimed at estimating the number of clusters. While these approaches improve over the baseline k-means algorithm they do not provide statistical guarantees on the estimate. Our proposed sampling-based approach results in a consistent estimator of the cluster count and produces accurate confidence intervals (§ 5). The amount of human effort required depends both on the quality of the pairwise similarity as well as the level of precision needed for estimation.

Related Problems Apart from population surveys, estimating the number of clusters is often the primary goal for many tasks. Life-long learning systems must discover and learn novel categories during long-term deployment. In generalized category discovery (GCD) [12, 54], the goal is to cluster images given the labels for a subset of images in the presence of novel categories. This problem is more challenging than traditional semi-supervised learning due to the open-world setting. However, we find that the performance of existing GCD approaches is limited by the accuracy of estimating the number categories in a dataset. Often, existing approaches make the unrealistic assumption that this quantity is known. We show that our approach for estimating the number of clusters has a higher impact on improving GCD than using active clustering algorithms. Our work is also related to approaches that improve statistical estimation by combining human effort and model predictions, such as ISCount [37], DISCount [43], and Prediction-Powered Inference [3]. For example, both ISCount and DISCount use importance sampling to estimate the population mean, and our work extends this idea to a pairwise setting.

3 Method

Assume that we have a collection of images $\mathcal{D} = \{x_i\}_{i=1}^n$ where each image x_i belongs to a cluster $y_i \in \mathcal{Y}$. Our goal is to calculate the total number of clusters $K = |\mathcal{Y}|$ in $\mathcal{D}$. We assume labels $y_i \in \mathcal{Y}$ are unknown but we have access to an image embedding $\Phi(\cdot)$ which can be used to compute an approximate pairwise similarity $\hat{s}(x, z) = f(\Phi(x), \Phi(z))$ between images x and z for some function f (such as the dot product). The true similarity is $s(x, z) = 1$ if images x and z belong to the same cluster and is $s(x, z) = 0$ otherwise. We assume this can be obtained by using human feedback in the form of "same" or "not" for the pair of images. Our goal is to estimate K as accurately as possible using a small amount of human feedback given the approximate similarity $\hat{s}(x, z)$.

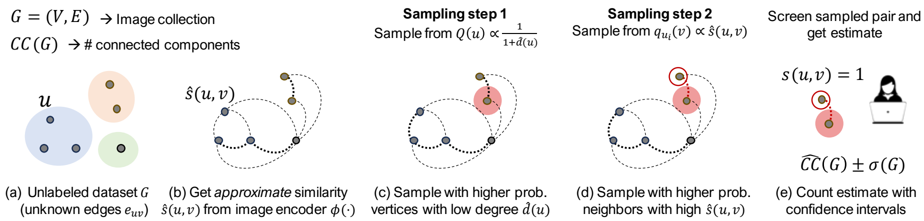

Fig. 3: Proposed Framework for Counting Clusters in a Dataset. (a) We represent dataset as a graph G and estimate the number of connected components for an (unknown) pairwise similarity. **(b)** First, we compute an *approximate* similarity between images using an embedding. **(c)** We sample vertices u_i from the distribution $Q(u)$ which biases the samples towards vertices with low (estimated) degrees. **(d)** We then sample nodes of $v_{i,j}$ from $q_{u_i}(v)$ biased towards neighbors. **(e)** Human feedback on the sampled pairs is used to estimate the number of clusters with confidence intervals.

Lemma 1. *Consider a graph $G = (V, E)$ with vertices $u, v \in V = \{1, ..., n\}$ corresponding to images x_u, with an edge $e_{uv} \in E$ between u and v if the images x_u and x_v belong to the same cluster. Let $d(u) = |\{e_{uv} \in E\}|$ be the degree of vertex u. Then the number of clusters K in the dataset $\mathcal{D}$ is equal to the number of connected components in G, and can be computed as:*

$$K = CC(G) = \sum_{u=1}^{n} \frac{1}{1 + d(u)}. \tag{1}$$

Proof (Proof of Lemma 1). It is easy to see that the G is a collection of cliques, with the clique containing u of size $1+d(u)$. The total contribution of the vertices in this clique is $(1+d(u)) \times (1/(1+d(u))) = 1$, i.e., each clique contributes a total of 1 to Eq. 1, adding up to the total number of cliques or connected components in G. See Fig. 2 for an example.

3.1 Estimation via Nested Sampling

The above lemma provides a way to estimate the number of connected components $\mathrm{CC}(G)$ by sampling.

$$\mathrm{CC}(G) = \sum_{u=1}^{n} \frac{1}{1 + d(u)} = n\mathbb{E}_{u \sim \mathrm{Unif}(1,n)} \left[\frac{1}{1 + d(u)} \right]. \tag{2}$$

Thus, one approach to estimate the number of connected components is to sample N vertices and estimate their degrees by querying a human if an edge exists between the vertex and all other $n - 1$ vertices in the graph. This would require $N \times (n - 1)$ queries, saving over $n \times (n - 1)/2$ queries for exact estimation. However, one can estimate the degree of a vertex by sampling as well. This suggests a two-step Monte Carlo (MC) approach for estimation. First, we sample N vertices $u_1, \ldots, u_N$. For each sampled vertex u_i, we then sample M nodes $v_{i,1}, \ldots, v_{i,M}$ uniformly at random, query a human to obtain $s(u_i, v_{i,j})$ for each potential neighbor, and estimate the degree of the vertex u_i as:

$$\hat{d}(u_i) = \frac{n-1}{M} \sum_{j=1}^{M} s(u_i, v_{i,j}). \tag{3}$$

From this, we can estimate the number of connected components as:

$$\widehat{\mathrm{CC}}_{\mathrm{MC}}(G) = \frac{n}{N} \sum_{i=1}^{N} \frac{1}{1 + \hat{d}(u_i)}. \tag{4}$$

This nested Monte Carlo estimate is asymptotically unbiased for the expected number of connected components, i.e., it converges to the true mean under mild assumptions. One can also construct a sample estimate of the variance and confidence intervals around the mean (we will derive the expressions for a general sampling distribution in the next section). The number of queries required to construct the estimate scales as $N \times M$, which is more efficient than the earlier approach of $N \times n$. However, the variance of the estimator can be high for graphs where the degrees of the vertices vary significantly.

Remark 1. Our problem is related to work on estimating the number of connected components in a graph in sub-linear time. The randomized algorithm proposed in [6, 11] use a similar sampling argument but their cost model is different. They assume the edges in the graph are provided as an adjacency list and run breadth-first-search starting from each node for a number of steps till the size of the connected components exceeds a threshold—a threshold of $1/\epsilon$ gives an $n\epsilon$ additive approximation to CC. In our setting the true edges are unknown—we instead have a noisy pairwise similarity—and have to pay a cost to reveal an edge, and hence the same approach is not applicable.

3.2 Estimation via Nested Importance Sampling

Importance sampling [40] uses a proposal distribution q and replaces the expectation of a quantity $f(x)$ under p as:

$$\mathbb{E}_p\left[f(x)\right] = \mathbb{E}_q\left[\frac{p(x)}{q(x)} f(x)\right]. \tag{5}$$

The equality holds for any proposal distribution q that satisfies $p(x)f(x) > 0 \implies q(x) > 0$. Moreover, one can show that the optimal proposal distribution $q(x) \propto p(x)f(x)$ for which the estimator has zero variance.

We are interested in computing various expectations under a uniform distribution p, and one can show that the optimal proposal distribution to sample nodes for estimating $\mathrm{CC}(G)$ in Eq. 1 is $Q(u) \propto 1/(1+d(u))$, while that to sample edges for estimating the degree $\hat{d}(u)$ in Eq. 3 is $q_u(v) \propto s(u, v)$.

However, sampling from these distributions requires knowledge of the true $s(u, v)$ which is unknown. But we can construct proposals using the approximate pairwise similarity $\hat{s}(u, v) \in [0, 1]$. This can be computed using the similarity between a pair of images estimated from their feature embeddings. Then, we can set $Q(u) \propto 1/(1 + \sum_{v \neq u} \hat{s}(u, v))$ as the proposal distribution to sample a set of vertices I of size N. For each sampled vertex u_i, we sample a set of M

Algorithm 1 Estimating cluster count k using NIS

$N \leftarrow$ Number of sampled vertices
$M \leftarrow$ Number of sampled edges per vertex
$\hat{s}(u, v) \leftarrow$ Approximate pairwise similarity
for i = 1,...,N **do**
    Sample $u_i \sim Q(u)$
    Sample $v_{i,1}, ..., v_{i,M} \sim q_{u_i}(v)$
    $\hat{d}(u_i) \leftarrow \frac{1}{M} \sum_{j=1}^{M} \frac{s(u_i, v_{i,j})}{q_{u_i}(v_{i,j})}$ // Human feedback   ▷ Eq. (7)
end for
$\widehat{\mathrm{CC}}_{\mathrm{NIS}} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{Q(u_i)} \times \frac{1}{1+\hat{d}(u_i)} \right)$ ▷ Eq. (6)

nodes $v_{i,j}$ from the distribution $q_u(v) \propto \hat{s}(u, v)$. The $\widehat{\mathrm{CC}}_{\mathrm{NIS}}(G)$ obtained using importance sampling is:

$$\widehat{\mathrm{CC}}_{\mathrm{NIS}}(G) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{Q(u_i)} \times \frac{1}{1 + \hat{d}(u_i)} \right), \ u_i \sim Q(u), \tag{6}$$

where, $\hat{d}(u_i)$ is estimated as:

$$\hat{d}(u_i) = \frac{1}{M} \sum_{j=1}^{M} \frac{s(u_i, v_{i,j})}{q_{u_i}(v_{i,j})}, \quad v_{i,1}, ..., v_{i,M} \sim q_{u_i}(v). \tag{7}$$

The proposal $Q(u)$ biases the distribution towards isolated vertices, i.e., ones that have low degrees, as they contribute the most toward the number of clusters. The second biases the distribution towards sampling nodes that are likely to be connected, i.e., to have high $\hat{s}(u, v)$. Note the only place human feedback is required is in computing $\hat{d}(u)$ in Eq. 7 consisting of M queries. Hence the overall query complexity of the approach is $N \times M$, similar to the simple Monte Carlo estimator. But if the similarity function $\hat{s}(u, v)$ is a good approximation to the true similarity then we expect the estimator to have a lower variance. Algorithm 1 and Fig. 3 describe the overall scheme.

3.3 Variance and Confidence Intervals

Theorem 1. *Assume $Q(u) > 0$ for all u and $q_u(v) > 0$ for all u, v. Let $\widehat{CC}_{N,M}$ denote the estimator using N sampled vertices and M sampled edges per vertex. For any $M > 0$, the estimator $\widehat{CC}_{N,M}$ is asymptotically normal, i.e.,*

$$\sqrt{N}(\widehat{CC}_{N,M} - \mu_M) \xrightarrow{D} \mathcal{N}(0, \sigma_{1,M}^2) \ \text{as} \ N \to \infty, \tag{8}$$

where $\mu_M = \mathbb{E}[\widehat{CC}_{N,M}] = \mathbb{E}[\widehat{CC}_{1,M}]$ and $\sigma_{1,M}^2 = \mathrm{Var}(\widehat{CC}_{1,M})$. Furthermore, the bias $|\mu_M - CC|$ is $O(1/M)$. Together these facts imply that the estimator is consistent as both N and M go to infinity, i.e., $\lim_{N\to\infty, M\to\infty} \widehat{CC}_{N,M} = CC$.

Asymptotic normality justifies the construction of confidence intervals as follows: let $\hat{\sigma}_{1,M}^2$ be the sample variance of $\frac{1}{Q(u_i)} \times \frac{1}{1+\hat{d}(u_i)}$ and use the 95% confidence interval $\widehat{\mathrm{CC}}_{N,M} \pm z_{0.025} \cdot \hat{\sigma}_{1,M}/\sqrt{N}$.

Proof (Proof of Theorem 1). The estimator $\widehat{\mathrm{CC}}_{N,M}$ is the sample average of N identically distributed copies of

$$\widehat{\mathrm{CC}}_{1,M} = \frac{1}{Q(u_1)} \times \frac{1}{1+\hat{d}(u_1)}. \tag{9}$$

Therefore $\mathbb{E}[\widehat{\mathrm{CC}}_{N,M}] = \mathbb{E}[\widehat{\mathrm{CC}}_{1,M}]$ and the asymptotic normality result holds by the central limit theorem as long as μ_M and $\sigma^2_{1,M}$ are finite. The quantity $\frac{1}{Q(u_1)} \times \frac{1}{1+\hat{d}(u_1)}$ is bounded above by our assumption that $Q(u) > 0$ and $q_u(v) > 0$. Together with the fact that the support of the joint sample space $(u_1, v_{1,1}, \ldots, v_{1,M})$ is finite, implies μ_M and $\sigma^2_{1,M}$ are finite. It remains to show that the bias is $O(1/M)$. This can be proved by the delta method, but also follows from a result of [45], which applies to our setting to assert that the mean-squared error of $\widehat{\mathrm{CC}}_{N,M}$ is $O(\frac{1}{N} + \frac{1}{M^2})$. Because mean-squared error equals variance plus squared bias, this necessitates that the bias, which only depends on M, is $O(\frac{1}{M})$.

Remark 2. [45] describe the optimal allocation of samples when $T = NM \to \infty$ and found that it occurs when N grow proportionally to M^2 in our setting. However, in applications with finite T, the lowest error occurred for $N \propto T^\alpha$ with α between 0.5 and 0.6, which suggests selecting N to be approximately proportional to M may be better in practice (See § 5–Parameters for NIS).

4 Experiments

This section describes the experimental setup, datasets, models used for computing image similarity, baselines, and the evaluation metrics.

Datasets We evaluate our approach on seven animal re-identification datasets, where the goal is to estimate the number of individuals (see Tab. 1) — **Chimpanzee Faces in the Wild** (CTai and CZoo) [20] CTai contains 5,078 images of 72 individuals living in the Taï National Park in Côte d'Ivoire, while CZoo consists of 2,109 recordings of 24 chimpanzees. **IPanda50** [58] contains 6,874 images of 50 giant panda individuals, **OpenCows2020** [2] comprises 4,736 images of 46 Holstein-Friesian cattle individuals, **MacaqueFaces** [60] includes 6,280 face images of 34 rhesus macaque individuals, **WhaleSharkID** [24] features 7693 images with 543 individual whale shark identifications, and **GiraffeZebraID** [41] with 6925 images of 2056 zebra and giraffe individuals in Kenya.

Image Encoding and Similarity We use image embeddings from the MegaDescriptor [55], a Swin-transformer model optimized with ArcFace loss, that beats CLIP [44], DINOv2 [10], and ImageNet-1k [16] pre-trained image encoders on animal Re-ID tasks. Specifically, we consider MegaDescriptor-L-384 for our experiments and present ablation tests with its smaller version MegaDescriptor-B-224 in § 5. The normalized cosine similarity between a pair of embeddings $\Phi(x_u)$ and $\Phi(x_v)$ as used as the proposal distribution with $\tau = 0.5$ (chosen with CZoo):

$$\hat{s}(u,v) = \frac{e^{\hat{c}(u,v)/\tau}}{\sum_v e^{\hat{c}(u,v)/\tau}} \quad \text{where} \quad \hat{c}(u,v) = \frac{\Phi(x_u) \cdot \Phi(x_v)}{||\Phi(x_u)||_2 ||\Phi(x_v)||_2}. \tag{10}$$

Table 1: Population Size Estimation on Animal Re-ID Datasets. The number of images $|\mathcal{D}|$ and individuals $|\mathcal{Y}|$ per dataset along with the estimated population k for a given amount of human effort (sampled pairs). Estimates using connected components (CoCo), Robust Continuous Clustering [48] (RCC), k-means (km), and mean-shift (not shown, but see Fig. 4) exhibit high error rates. k-means improves with the addition of human feedback (pc$k$m) but the error rates remain high. Farthest-first traversal (FFT) outperforms these methods. The proposed Nested-IS (NIS) surpasses these baselines and yields estimates and confidence intervals that often contain the true value. Error rates (%) on each dataset (shown for NMC and NIS) are also lower compared to the baseline and Nested Monte Carlo (NMC). Our method demonstrates significant improvement on the challenging WhaleSharkID [24] and GirrafeZebraID [41].

			Sampled pairs	Estimated k @ $N \times M$ sampled pairs					
								CC±CI (%error)	
Dataset	$\|\mathcal{D}\|$	$\|\mathcal{Y}\|$	$(\frac{N \times M}{\|E\|})$	CoCo	RCC	km→pc$k$m	FFT	NMC	NIS
CTai [20]	5078	72	0.014%	7	258	90 → 87	54	26±4 (64%)	**63±20 (16%)**
CZoo [20]	2109	24	0.004%	3	57	50 → 49	13	08±1 (68%)	**23±07 (17%)**
IPanda50 [58]	6874	50	0.002%	7	251	80 → 79	34	17±3 (66%)	**50±16 (13%)**
OpenCows2020 [2]	4736	46	0.006%	39	220	70 → 69	39	18±3 (62%)	**40±13 (17%)**
MacaqueFaces [60]	6280	34	0.002%	13	90	80 → 77	31	15±2 (55%)	**34±08 (09%)**
WhaleSharkID [24]	7693	543	0.16%	630	1182	70 → 132	251	145±9 (73%)	**543±74 (06%)**
GiraffeZebraID [41]	6925	2056	0.16%	4714	500	100 → 155	271	403±34 (80%)	**1951±370 (08%)**
Average error (%)				69.4%	219.2%	80.4% → 75.4%	38.2%	66.9%	**12.3%**

Clustering Baselines We employ four baselines for estimating the number of clusters. The first is mean-shift clustering [13]. We set the bandwidth parameter as the average distance between samples and their nearest neighbor, and estimate the number of clusters as the number of modes. The second method is k-means [22]. To calculate the *optimal* k, we calculate the within-cluster sum of squares $\mathcal{J}_{\text{km}} = \sum_{i \in G} ||c_i - u_i||_2^2$, where c_i is the cluster center corresponding to u_i. The elbow is identified as the value of k where the slope becomes approximately constant. Third, we use connected components (CoCo) after thresholding the similarity values $\hat{s}(u, v)$ of our graph G. To find the optimal threshold, we use the same procedure as the k-means elbow; that is, we calculate $\mathcal{J}_{\text{km}}$ for different thresholds and return the threshold at the elbow. Lastly, we use robust continuous clustering (RCC) to extract the number of clusters after optimizing the algorithm's objective using the default hyperparameters in [48].

Active Clustering We use active variants of the above approaches. First, we use pairwise constrained k-means (pc$k$-means [4]). For a given k, the objective is:

$$\mathcal{J}_{\text{pckm}} = \sum_{i \in G} ||c_i - u_i||_2^2 + \sum_{(i,j) \in \mathcal{M}} \alpha \mathbb{1}[l_i \neq l_j] + \sum_{(i,j) \in \mathcal{C}} \beta \mathbb{1}[l_i = l_j],$$

where $\mathcal{M}$ and $\mathcal{C}$ are the sets of *must-link* and *cannot-link* constraints, respectively, and l_i denotes the cluster index for i. Scalars $\alpha, \beta > 0$ trade off the k-means objective with the cost of violating constraints. We select empirically $\alpha = \beta = 1$. Given a set of constraints, we can pick the optimal k using the elbow heuristic on the $\mathcal{J}_{\text{pckm}}$ objective. We pick samples based on a farthest-first traversal scheme described below. The second method is constrained mean shift [46] (Constrained-MS) that introduces a density-based integration of the

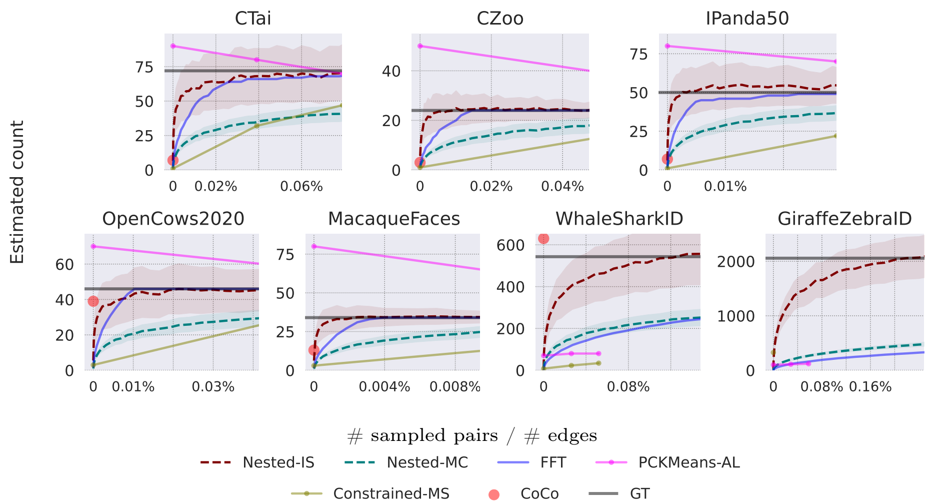

Fig. 4: Performance of Estimating k per Human Effort on Animal Re-ID Datasets. We use the cosine similarity built from the MegaDescriptor-L-384 image embeddings–See § 4. The human effort is measured as the fraction of the sampled pairs and total pairs $|E|$ in the dataset G. Our method estimates the true count with less human effort compared to baselines. Dashed lines indicate the mean estimates and shaded regions indicate the mean 95% confidence interval across 100 trials.

constraints. Lastly, instead of incorporating constraints within k-means one can directly estimate the number of clusters using a farthest-first traversal (FFT) of points and exhaustive comparison with existing individuals. The approach is as follows: 1) Initialize the list of sampled images S as empty and list of discovered individuals as I as empty, 2) Selecting an unsampled point that is farthest from S, i.e., $u = \text{argmax}_u \min_{v \in S} \ d(u, v)$, and add it to the sampled set S. 3) Compare u to all the previously discovered individuals in I. If it matches an individual add it to the corresponding list, else start a new list with u and add it to I. FFT rapidly explore the dataset to find at least one member for each cluster. At each iteration it pays a cost equal to the number of discovered individuals.

Evaluation Metric and Human Effort Error is measured as $|\text{CC} - \widehat{\text{CC}}|/\text{CC}$, the fractional absolute difference between the estimated $\widehat{\text{CC}}$ and the true number of clusters CC. Human effort is measured as the number of pairwise queries used for estimation. To account for variable dataset sizes, we report the number of sampled pairs normalized by the total number of edges $|E|$ in G.

5 Results

We compare our method to the baselines described in § 4 that include connected components (CoCo), pairwise constrained k-means (pc$k$-means), constrained mean shift (Constrained-MS), our farthest-first traversal baseline (FFT), and nested Monte Carlo sampling (Nested-MC) from Eq. (4).

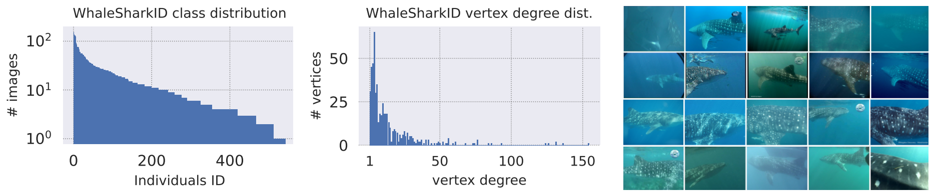

Fig. 5: WhaleSharkID Dataset Statistics. (Left) The dataset is long-tailed with many individuals with a few images. (Center) Histogram of vertex degree distribution–most individuals have less than 5 images. (Right) Sample images from the dataset.

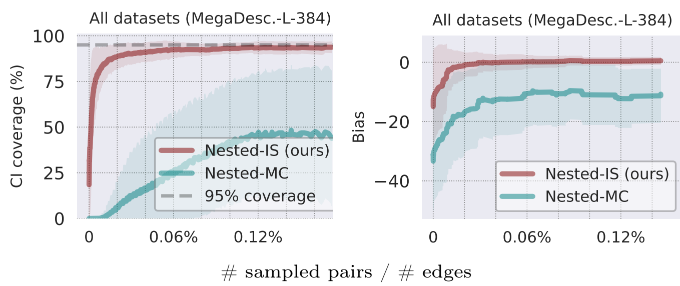

Fig. 6: Confidence Intervals' Coverage and Bias using MegaDescriptor-L-384 feature embeddings on all datasets. (Left) Our method produces a CI coverage close to 95% (when using $z_{0.025} = 1.96$ for a 95% confidence interval–§ 3.3) with less than 0.04% of sampled pairs. (Right) Even though our estimator has a negative bias for a low number of sampled pairs, it rapidly reaches zero compared to Nested-MC.

Baselines Perform Poorly Techniques like k-means and CoCo perform poorly, with about 80% and 69% error on average, respectively, despite using state-of-the-art image embeddings to compute similarity (see Fig. 4 and Tab. 1). For instance, k-means estimates 70 clusters for WhaleSharkID instead of 543, 70 clusters instead of 34 for OpenCows2020, and 50 clusters instead of 24 for CZoo. Estimates using CoCo are slightly better on average but tend to be an underestimate. One reason is that choosing hyperparameters for clustering is particularly challenging. For instance, the WCSS curve as a function of k looks rather smooth, and there is no clear "elbow." Another reason is that the underlying similarity is imperfect. Improving this through deep metric learning approaches would require significant resources and expertise to set various parameters and may not pan out when supervision is limited [38].

Nested-IS outperforms Nested-MC Both Nested-MC and Nested-IS improve over the baselines as more human feedback is collected. As described in § 3.1, Nested-MC samples edges uniformly at random, while Nested-IS is driven by the pairwise similarity and leads to an error reduction at a much faster rate as seen in Fig. 4. Specifically, we get a relative error reduction of 82% over Nested-MC with the same human effort.

Nested-IS outperforms Active Clustering We compare our method to pairwise constrained k-means (pc$k$-means-AL), constrained mean shift (Constrained-MS),

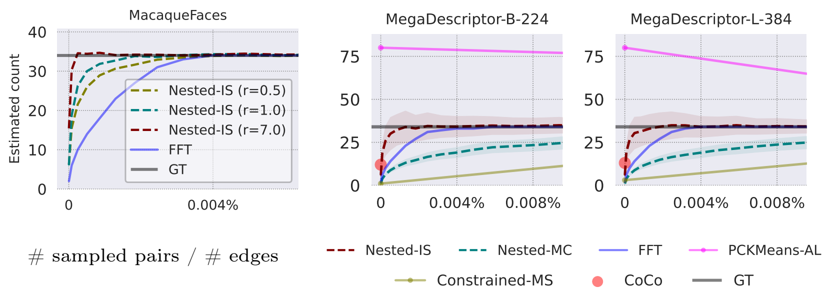

Fig. 7: Ablation Experiments on MacaqueFaces (Left) Our proposed method (Nested-IS) with different sampling ratios $r = M/N$. Increasing the number of sampled neighbors M per sampled vertices N (§ 3.2) improves performance for these datasets. (Center-Right) Performance comparison between MegaDescriptor-B-224 and MegaDescriptor-L384. Although our method is slightly better using superior feature embeddings (L-384), it still outperforms all baselines with B-224.

and the proposed farthest-first traversal approach (FFT) with similar human effort. Fig. 4 shows that Nested-IS handily outperforms active clustering. In WhaleSharkID, despite performing best, our method still requires querying relatively many pairs to get accurate estimates, which can be explained by the difficulty of the task. In Fig. 5–*left-center* we show the histogram of individuals with a clear long-tailed distribution and the histogram of vertex degrees where we can see that in WhaleSharkID most individuals have less than 5 images. The performance of the MegaDescriptor is also the lowest (around 62%) on the Re-ID task on this dataset. However, the performance savings over alternative approaches is significant.

Confidence Intervals are Calibrated In addition to lower error rates, a key advantage of our approach is that it also provides confidence intervals (CIs) (See Fig. 4). To estimate the quality of the CI estimation we calculate the empirical coverage of the CI over 100 runs (i.e., the percentage of times the estimated CIs contains the true count). Our method produces close to 95% coverage using $z_{0.025} = 1.96$ for a 95% CI, as described in § 3.3, when around 0.02% of pairs are sampled across datasets, as shown in Fig. 6–*left*.

Estimation Bias is Low We calculate the empirical bias, denoted as $\textsc{Bias} = \mathbb{E}[\widehat{\mathrm{CC}}_{\mathrm{NIS}}(G)] - \mathrm{CC}(G)$ where $\mathbb{E}[\widehat{\mathrm{CC}}_{\mathrm{NIS}}(G)]$ is the mean count estimate over 100 runs. Fig. 6–*right* shows that the bias is negative initially, but it rapidly drops to 0. The initial negative bias might explain the lower coverage of the CIs.

Parameters for Nested-IS There is a trade-off between the number of sampled vertices N and sampled neighbors per vertex M. Increasing N will reduce the variance of the estimation of the number of clusters, and increasing M will reduce the variance of each vertex degree estimate. In Fig. 7–*left* we show that increasing the the ratio between the number of sampled neighbors per vertex M relative to the number of sampled vertices N (i.e., $r = M/N$) produces more accurate estimations with less human effort. For instance, using $r = 7$ achieves close-to-zero error with 1/5 of the sampled pairs compared to using $r = 0.5$.

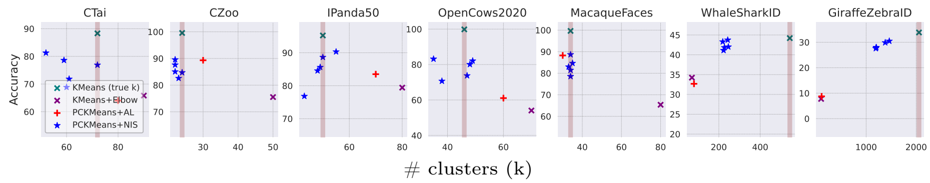

Fig. 8: Measuring Clustering Accuracy. Pairwise constraints (pc$k$-means+AL) improves the clustering accuracy over k-means+elbow, but the clustering accuracy with our estimated k is even better (pc$k$-means+NIS). We plot results with five independent runs for our proposed approach (blue stars). The red shaded line indicates the true number of clusters in the dataset.

Ablation with other MegaDescriptors We test the performance of all baselines using a smaller less-performing version of the MegaDescriptor (MegaDescriptor-B-224). In Fig. 7–*center-right* we show that our method performs similarly with both MegaDescriptor versions.

How Much do we Know about the Clustering? Sometimes we also wish to recover the clustering in a dataset, such as in generalized category discovery [54]. Should human effort be used to estimate the clustering directly using active clustering, or to estimate k for k-means using our approach? To answer this we conduct the following experiment. First, we can measure the accuracy of a clustering as:

$$\textsc{Acc} = \max_{p \in \mathcal{P}(\mathcal{Y})} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[y_i = p(\hat{y}_i)], \tag{11}$$

where $\mathcal{P}(\mathcal{Y})$ is the set of all permutations of the class labels. We compare the accuracy of clustering using k-means with estimated k using the elbow method, using pc$k$-means, and using k-means with k estimated using Nested-IS for tthe same amount of human effort as pc$k$-means. Fig. 8 shows the results. Surprisingly, we find that accurately estimating k has a larger impact on the quality of clustering than active clustering, which suggests that human effort is better spent to estimate k initially. In the Supplementary Material we describe experiments on GCD on fine-grained classification datasets and show a similar trend.

6 Discussion and Conclusion

We propose a human-in-the-loop approach to estimate population size when deploying imperfect Re-ID systems. By carefully selecting a small fraction of pairs to label (often less than 0.002% of all edges), our approach produces unbiased estimates of the population size. A key advantage of our method is that it generates confidence intervals which can be used for guiding human effort. This approach can be implemented on top of any Re-ID system, as it requires only a pairwise similarity between images, making it practical for low-resource settings. Our approach adds to the growing literature on statistical estimation techniques [3, 37, 43] that combine model predictions and ground-truth labels to improve the precision of count estimates. However, we tackle the novel problem of estimating cluster counts, which involves pairwise comparisons.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants #1749854, #1749833, and #2329927. We thank Hung Le and Cameron Musco for initial discussions and the Wildlife Datasets team for publicly releasing the datasets and models for animal Re-ID.

References

1. Anderson, C.J., Da Vitoria Lobo, N., Roth, J.D., Waterman, J.M.: Computer-aided photo-identification system with an application to polar bears based on whisker spot patterns. Journal of Mammalogy **91**(6), 1350–1359 (2010) 1, 4
2. Andrew, W., Gao, J., Mullan, S., Campbell, N., Dowsey, A.W., Burghardt, T.: Visual identification of individual holstein-friesian cattle via deep metric learning. Computers and Electronics in Agriculture **185**, 106133 (2021) 9, 10
3. Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I., Zrnic, T.: Prediction-powered inference. Science **382**(6671), 669–674 (2023) 5, 14
4. Basu, S., Banerjee, A., Mooney, R.J.: Active semi-supervision for pairwise constrained clustering pp. 333–344 (April 2004) 2, 5, 10
5. Beery, S., Morris, D., Yang, S.: Efficient pipeline for camera trap image review. arxiv. arXiv preprint arXiv:1907.06772 (2019) 1
6. Berenbrink, P., Krayenhoff, B., Mallmann-Trenn, F.: Estimating the number of connected components in sublinear time. Information Processing Letters **114**(11), 639–642 (2014) 7
7. Bolger, D.T., Morrison, T.A., Vance, B., Lee, D., Farid, H.: A computer-assisted system for photographic mark–recapture analysis. Methods in Ecology and Evolution **3**(5), 813–822 (2012) 1, 4
8. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient n-d image segmentation. Int. J. Comput. Vis. **70**(2), 109–131 (2006) 4
9. Burton, A.C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J.T., Bayne, E., Boutin, S.: Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. Journal of Applied Ecology **52**(3), 675–685 (2015) 1
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: International Conference on Computer Vision (ICCV) (2021) 4, 9
11. Chazelle, B., Rubinfeld, R., Trevisan, L.: Approximating the minimum spanning tree weight in sublinear time. SIAM Journal on computing **34**(6), 1370–1379 (2005) 7
12. Chiaroni, F., Dolz, J., Masud, Z.I., Mitiche, A., Ben Ayed, I.: Parametric information maximization for generalized category discovery. In: International Conference on Computer Vision (ICCV) (2023) 2, 3, 5
13. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002) 4, 10
14. Craenendonck, T.V., Dumancic, S., Blockeel, H.: Cobra: A fast and simple method for active clustering with pairwise constraints. In: International Joint Conference on Artificial Intelligence (2017) 5
15. Crall, J.P., Stewart, C.V., Berger-Wolf, T.Y., Rubenstein, D.I., Sundaresan, S.R.: Hotspotter—patterned species instance recognition. In: IEEE workshop on applications of computer vision (WACV) (2013) 1, 4

16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (2009) 9
17. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4
18. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 337–33712 (2018) 4
19. ERDdS, P., R&wi, A.: On random graphs i. Publ. math. debrecen **6**(290-297), 18 (1959) 4
20. Freytag, A., Rodner, E., Simon, M., Loos, A., Kühl, H.S., Denzler, J.: Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In: Rosenhahn, B., Andres, B. (eds.) Pattern Recognition. pp. 51–63. Springer International Publishing, Cham (2016) 9, 10
21. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision (ECCV) (2016) 4
22. Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. JSTOR: Applied Statistics **28**(1), 100–108 (1979) 4, 10
23. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: Feragen, A., Pelillo, M., Loog, M. (eds.) Similarity-Based Pattern Recognition. pp. 84–92. Springer International Publishing, Cham (2015) 4
24. Holmberg, J., Norman, B., Arzoumanian, Z.: Estimating population size, structure, and residency time for whale sharks rhincodon typus through collaborative photo-identification. Endangered Species Research **7**, 39–53 (2009) 4, 9, 10
25. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition (2008) 4
26. J, J.K., Paul, S., Aggarwal, G., Biswas, S., Rai, P., Han, K., Balasubramanian, V.N.: Novel class discovery without forgetting. In: European Conference on Computer Vision (ECCV) (2022) 2, 3
27. KAYA, M., BİLGE, H.: Deep metric learning: A survey. Symmetry **11**(9) (2019) 4
28. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition (2015) 4
29. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013) 1
30. Kulis, B.: Metric learning: A survey. Foundations and Trends in Machine Learning **5**(4), 287–364 (2013) 4
31. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., , Hua, G.: Labeled faces in the wild: A survey. In: Advances in Face Detection and Facial Image Analysis. pp. 189–248 (2016) 3
32. Learned-Miller, E., Ordóñez, V., Morgenstern, J., Buolamwini, J.: Facial recognition technologies in the wild: A call for a federal office. Algorithmic Justice League (2020) 4
33. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 4

34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision (ICCV) (2021) 4
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (Nov 2004) 4
36. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. Tech. rep. (2013) 1
37. Meng, C., Liu, E., Neiswanger, W., Song, J., Burke, M., Lobell, D., Ermon, S.: Is-count: Large-scale object counting from satellite images with covariate-based importance sampling. In: AAAI Conference on Artificial Intelligence (2022) 5, 14
38. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) European Conference on Computer Vision (ECCV) (2020) 12
39. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008) 1
40. Owen, A.B.: Monte Carlo theory, methods and examples (2013) 7
41. Parham, J., Crall, J., Stewart, C., Berger-Wolf, T., Rubenstein, D.: Animal population censusing at scale with citizen science and photographic identification. In: SS-17-01. pp. 37–44. AAAI Spring Symposium - Technical Report, AI Access Foundation (2017) 9, 10
42. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015) 4
43. Perez, G., Maji, S., Sheldon, D.: DISCount: Counting in Large Image Collections with Detector-Based Importance Sampling. In: AAAI Conference on Artificial Intelligence (2024) 5, 14
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021) 4, 9
45. Rainforth, T., Cornish, R., Yang, H., Warrington, A., Wood, F.: On nesting monte carlo estimators. In: International Conference on Machine Learning. pp. 4267–4276. PMLR (2018) 9
46. Schier, M., Reinders, C., Rosenhahn, B.: Constrained mean shift clustering. In: International Conference on Data Mining (SDM). SIAM (2022) 10
47. 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 in Ecology and Evolution **10**(4), 461–470 (2019) 1
48. Shah, S.A., Koltun, V.: Robust continuous clustering. Proceedings of the National Academy of Sciences **114**(37), 9814–9819 (2017) 10
49. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence **22**(8), 888–905 (2000) 4
50. Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J.T., Burton, C., Townsend, S.E., Carbone, C., Rowcliffe, J.M., Whittington, J., et al.: Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. Frontiers in Ecology and the Environment **15**(1), 26–34 (2017) 1
51. Thorndike, R.: Who belongs in the family? Psychometrika **18**, 267–276 (1953) 2
52. Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., van Langevelde, F., Burghardt, T., et al.: Perspectives in machine learning for wildlife conservation. Nature communications **13**(1), 792 (2022) 1

53. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of cognitive neuroscience **3**(1), 71–86 (1991) 3
54. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: IEEE Conference on Computer Vision and Pattern Recognition (2022) 2, 3, 5, 14
55. Čermák, V., Picek, L., Adam, L., Papafitsoros, K.: WildlifeDatasets: An Open-Source Toolkit for Animal Re-Identification. In: IEEE workshop on applications of computer vision (WACV) (2024) 1, 2, 3, 4, 9
56. Vidal, M., Wolf, N., Rosenberg, B., Harris, B.P., Mathis, A.: Perspectives on individual animal identification from biology and computer vision. Integrative and comparative biology **61**(3), 900–916 (2021) 1
57. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 1
58. Wang, L., Ding, R., Zhai, Y., Zhang, Q., Tang, W., Zheng, N., Hua, G.: Giant panda identification. IEEE Transactions on Image Processing **30**, 2837–2849 (2021) 9, 10
59. Weideman, H., Stewart, C., Parham, J., Holmberg, J., Flynn, K., Calambokidis, J., Paul, D.B., Bedetti, A., Henley, M., Pope, F., et al.: Extracting identifying contours for african elephants and humpback whales using a learned appearance model. In: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision. pp. 1276–1285 (2020) 1
60. Witham, C.L.: Automated face recognition of rhesus macaques. Journal of Neuroscience Methods **300**, 157–165 (2018), measuring Behaviour 2016 2, 9, 10

A Experiments on Fine-Grained Classification Datasets

In this section we present experiments on four fine-grained image classification datasets, where the goal is to estimate the number of categories — **Caltech-UCSD Birds** (CUB) [57] consists of 11,788 images of 200 bird species, **Stanford Cars** [29] contains 16,185 images of 196 car models, **FGVC Aircraft** [36] comprises 10,000 images of 100 aircraft models, and **Oxford Flowers** [39] includes 8,189 images of 102 flower categories. Although the names of categories are known, we use the same pairwise setting as the Re-ID tasks.

A.1 Performance of Estimating k

Here we show the estimated k as a function of the number of sampled pairs, as with the Re-ID datasets in Fig. 4. Similarly to Re-ID tasks, our method outperforms all the baselines when using feature embeddings from an ImageNet-pretrained ResNet50, DINO ViT-B/8, and CLIP ViT-L/14. We calculate the similarity as described in § 4.

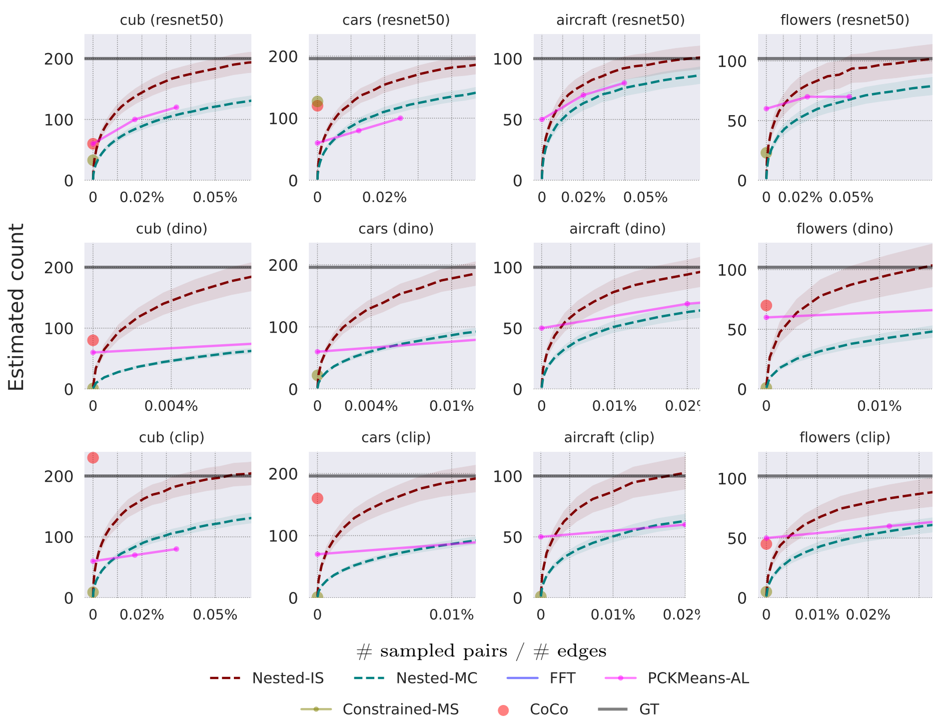

Fig. A1: Performance of Estimating k per Human Effort across fine-grained classification datasets. We use the cosine similarity built from ImageNet pretrained ResNet50, DINO ViT-B/8, and CLIP ViT-L/14 image embeddings. The human effort is measured as the fraction of the sampled pairs and total pairs $|E|$ in the dataset G. Our method estimates the true count with less human effort compared to the other baselines. Dashed lines indicate the mean estimates and shaded regions indicate the mean 95% confidence interval across 100 trials.

A.2 Measuring Clustering Accuracy

Similarly to the Re-ID datasets (See Fig. 8), we find that accurately estimating k has a bigger impact on the quality of clustering than active clustering, which suggests that human effort is better spent to estimate k initially.

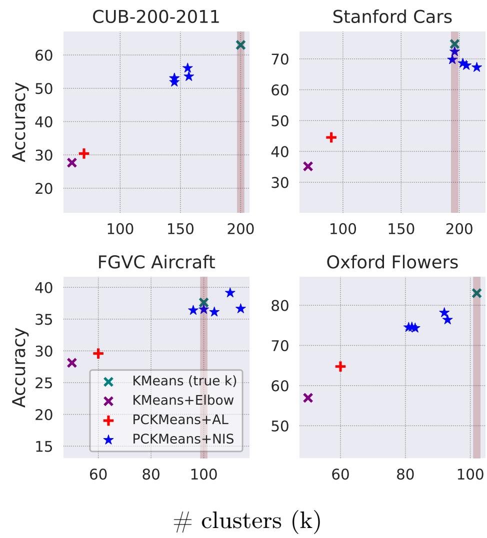

Fig. A2: Clustering accuracy on fine-grained classification datasets using the right number of clusters. Using the improved k and pairwise constraints (pc$k$-means+AL) improves the clustering accuracy over k-means+elbow, while clustering accuracy with our estimated k improves accuracy further (pc$k$-means+NIS). We plot results with five estimated ks and constraints from our proposed approach (blue stars). The red shaded line indicates the true number of clusters in the dataset.