Information Fusion 133 (2026) 104323

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

From species-specific models to universal re-ID: a survey of animal re-identification

Cigdem Beyan *, Anil Osman Tur , Ehsan Karimi

Department of Computer Science, University of Verona, Strada Le Grazie 15, 37134, Verona, Italy

ARTICLE INFO

Keywords:
Animal re-identification
Individual identification
Wildlife monitoring
Agricultural vision

ABSTRACT

Automated animal re-identification (Re-ID) has become an essential tool for wildlife ecology, conservation management, and precision livestock farming. Recent progress in deep representation learning, transformer architectures, multimodal learning, and vision-language modeling has accelerated the development of scalable, non-invasive systems for identifying individuals across images and videos. This survey provides a comprehensive review of animal Re-ID research published between 2020 and 2025, encompassing 41 peer-reviewed works. We propose a structured taxonomy of animal Re-ID methods and provide an integrated analysis of approaches, datasets, and evaluation practices. We also highlight persistent challenges, including domain shift, temporal variability, annotation scarcity, and inconsistent evaluation protocols, and outline broad future research directions toward universal, temporally robust, and ecologically meaningful animal Re-ID systems. This survey provides a unified foundation for advancing robust and deployable solutions in the coming decade.

1. Introduction

Accurate identification of individual animals is fundamental to wildlife conservation, ecological research, and sustainable livestock management. Traditional methods, such as physical tagging or manual photo cataloging by experts, are resource-intensive, invasive, and scale poorly when applied to large or widely dispersed populations [1,2]. Consequently, automated animal re-identification (Re-ID), particularly through advances in computer vision and deep learning, has emerged as a transformative tool, enabling non-invasive and scalable tracking from images and videos [3–5].

In the context of computer vision and deep learning, Re-ID refers to recognizing the same individual across different images or video sequences, often captured at different times, viewpoints, devices, or environments. This objective differs from species classification or detection because identities must be matched within a species. Early Re-ID research focused on human surveillance [6], resulting in extensive benchmarks and architectures optimized for person Re-ID [7,8]. These studies established robust techniques for feature extraction, metric learning, and domain adaptation, e.g., Rami et al. [9], Bak and Carr [10], Wu et al. [11], Li et al. [12]. However, directly transferring these advances to animals is nontrivial.

Animal Re-ID poses unique challenges that limit the direct transfer of human Re-ID methods. Unlike humans, animals lack consistent visual markers such as clothing or accessories, exhibit greater pose variability, and are often observed in unconstrained natural environments. Their images can show strong intra-class variation due to pose, viewpoint, illumination, seasonal changes, occlusion, and background clutter [13–15]. Datasets are typically smaller, long-tailed, and expensive to annotate, and open-set conditions where new individuals appear over time are naturally common, yet many existing models perform poorly under such scenarios [16–18]. Beyond these environmental and observational challenges, animals also undergo intrinsic biological changes, such as growth, molting, seasonal coat variation, and aging, which introduce appearance shifts independent of viewpoint or environmental factors. Combined with habitat clutter and limited training data, these biological variations further complicate the task, making animal Re-ID inherently difficult.

Historically, automated animal Re-ID evolved from expert-driven photo-catalogue matching and handcrafted visual descriptors toward learned visual embeddings. Early computer-assisted systems relied on contour matching and local invariant features tailored to specific species [3]. Although these methods demonstrated the feasibility of automated Re-ID, they remained constrained by species specificity and limited generalization. With the maturation of deep representation learning, recent research has explored a broader set of strategies. State-of-the-art (SOTA) pipelines predominantly rely on convolutional and transformer-based architectures (e.g., [19,20]), typically paired with metric-learning objectives to construct discriminative embeddings (e.g., [13,21]). Beyond these foundations, pose-aware,

* Corresponding author.
E-mail addresses: cigdem.beyan@univr.it (C. Beyan), anilosman.tur@univr.it (A.O. Tur), ehsan.karimi@studenti.univr.it (E. Karimi).

https://doi.org/10.1016/j.inffus.2026.104323
Received 17 December 2025; Received in revised form 19 March 2026; Accepted 24 March 2026
Available online 26 March 2026

pattern-based, and geometric-consistency models improve robustness under extreme viewpoint and appearance variation [13,21–23], while temporal architectures integrate video cues for more stable representations [15,16,24,25]. Additional work investigates domain adaptation and cross-dataset generalization [18,19,26], few-shot learning [27] to mitigate annotation scarcity, and self-supervised [28] or unsupervised methods [29] that eliminate reliance on identity labels. Multimodal approaches incorporate metadata, pose or orientation cues, thermal imagery, or vision-language alignment [18,30–32], and semi-automated systems integrate human-in-the-loop feedback for greater reliability under open-set conditions [33]. In some settings, Re-ID is embedded within larger tracking or video-association frameworks (e.g., [34]). Collectively, these advances reflect a shift toward more robust, scalable, and generalizable animal Re-ID systems by addressing the limitations outlined above.

These methodological developments are closely tied to the availability of diverse datasets spanning wildlife and agricultural contexts, which provide the empirical foundation for evaluating and comparing animal Re-ID systems. In wildlife-focused automated Re-ID, curated datasets exist for species such as sea turtles [17], manta rays [35], polar bears [36], yaks [37], and insects, including honeybees [16] and bumblebees [38], capturing diverse fine-grained identity cues in natural imagery. In agricultural applications, automated Re-ID supports herd management, welfare monitoring, and productivity analysis, with representative datasets covering Holstein cows [28,39], cattle Re-ID in farm environments [30], and pig Re-ID benchmarks [19,34].

From a benchmarking perspective, most automated animal Re-ID studies still focus on a single species (e.g., [29,36]), reflecting the ecological and agricultural settings in which datasets are typically collected. While species-specific datasets enable targeted analysis, they also risk encouraging texture memorization and limit the assessment of how well models generalize beyond the conditions in which they were trained. Although early multi-species Re-ID efforts have appeared (e.g., [22,40]), models are still typically trained and evaluated separately for each species. Even benchmarks that combine multiple datasets [4,41] rarely include evaluation protocols that test whether a model trained on one species (or one dataset) can successfully identify individuals from another. As a result, the field still lacks systematic assessments of cross-species and cross-dataset generalization. Only recently has it become feasible to move toward unified Re-ID models that operate across multiple species, including previously unseen categories. This shift is driven by visual-language models (VLMs) and multimodal large language models (MLLMs) [18], which enable prompt-based adaptation and species-agnostic representation learning.

Building on these developments in methodology, datasets, and benchmarking practices, this survey reviews research progress in automated animal Re-ID from 2020 to 2025; a period marked by rapid advances in deep learning, multimodal modeling, and evaluation resources. We provide an integrated analysis of methods, datasets, and emerging benchmarking frameworks, highlight persistent challenges, and outline key research directions to guide future work toward more robust, scalable, and generalizable animal Re-ID systems.
The contributions of this survey paper are as follows.

- We systematically analyze recent animal Re-ID methods in depth, detailing their architectures, learning formulations, and technical innovations.
- We organize the field along five methodological axes, providing a unified framework for comparing approaches in feature representation, metric learning, temporal modeling, multimodal integration, and alternative training regimes.
- We review all major animal Re-ID datasets, categorizing them into four habitat-based groups, and analyze their species coverage, scale, annotation types, capture conditions, and evaluation protocols.
- We critically compare existing benchmarking frameworks, including evaluation protocols, toolkits, and cross-species benchmarks, and analyze their strengths, limitations, and implications for standardization.
- Based on limitations in current methods, datasets, and benchmarks, we identify key challenges and outline future priorities for advancing robust, generalizable, and ecologically grounded animal Re-ID.

The remainder of this paper is organized as follows. Section 2 introduces the conceptual foundations of animal Re-ID, including key terminology, problem definitions, and distinctions from person Re-ID. Section 3 describes our survey methodology and outlines the central research questions guiding this work. Section 4 presents a high-level methodological landscape, organizing recent approaches according to feature representation, metric learning, temporal modeling, multimodal integration, and alternative learning regimes. Section 5 provides a detailed review of state-of-the-art (SOTA) methods. Section 6 surveys animal Re-ID datasets and benchmarking resources, grouped by habitat and acquisition context, and analyzes their annotation types, evaluation protocols, and public availability. Section 7 synthesizes major challenges and open problems encountered across methods, datasets, and evaluation practices. Section 8 outlines future research directions and recommendations. Finally, Section 9 concludes the survey.

2. Foundations of animal re-ID

Accurately identifying individual animals is essential in both ecological research and agricultural management. Traditionally, this has relied on manual or physical identification techniques. In agricultural settings, farmers use ear tags and electronic identifiers [42–44], while wildlife researchers depend on radio/GPS collars, PIT tags, or expert-curated photo catalogues that record distinctive morphological features [45–48]. These approaches provide reliable identity labels, which are then applied to photographs collected during routine monitoring, forming the annotated image datasets that underpin automated animal Re-ID studies [19,28,30,34,39]. Building on these labeled visual datasets, automated animal Re-ID has emerged as an interdisciplinary problem at the intersection of computer vision, ecology, and behavioral science, aiming to recognize individuals across images or video sequences in a non-invasive manner [3].

Before the recent wave of transformer-based and multimodal approaches, automated animal Re-ID evolved through several foundational stages spanning ecology, classical computer vision, and early deep learning. Early photo-identification systems in wildlife research date back to the 1990s, when semi-automated pipelines were developed to assist manual identification of individuals based on distinctive morphological patterns. For example, Whitehead [49] introduced computer-assisted matching of sperm whale flukes, demonstrating the feasibility of contour-based identification in ecological monitoring. Similar approaches relied on expert-curated catalogs of dorsal fins, pelage patterns, or facial features, often requiring manual landmark annotation and domain expertise.

Subsequently, classical computer vision techniques were adopted to reduce manual effort. Handcrafted feature descriptors such as SIFT [50] and SURF were combined with geometric verification and database retrieval frameworks [3]. These pipelines extracted affine-invariant local features and aggregated them using techniques such as bag-of-visual-words or Fisher vectors, enabling scalable matching for patterned species. However, they were typically species-specific, sensitive to illumination and deformation, and required careful parameter tuning.

The introduction of deep convolutional neural networks (CNNs) marked a significant paradigm shift. Inspired by advances in person Re-ID, researchers began adopting Siamese architectures and triplet-loss-based metric learning to learn discriminative embeddings directly from data. As reviewed in Ravoor and T.s. b. [3], early deep-learning-based animal Re-ID systems focused primarily on single-species settings and CNN backbones, often adapting human Re-ID pipelines with minimal architectural modification. These methods improved robustness to pose

and viewpoint variation but remained largely species-specific, single-modal, and limited in cross-dataset generalization. Collectively, these foundational works established the transition from handcrafted biometrics to learned visual embeddings, setting the stage for the methodological diversification observed after 2020.

While early deep learning-based approaches borrowed architectural principles from person Re-ID, the two domains differ substantially in data regimes, sources of variation, and evaluation practice. Person Re-ID (most commonly studied as pedestrian Re-ID [51,52]) benefits from large multi-camera benchmarks with abundant identities, frequent re-appearances, and rapid intra-session changes, where intra-class variation is mainly driven by external factors such as pose, clothing, illumination, or occlusion [7,53,54]. In contrast, animal Re-ID typically operates with far fewer identities, sparser observations, and data collection protocols driven by ecological or farming constraints [3–5,40,55]. Moreover, it must contend with intrinsic biological changes over time such as growth, molting, and seasonal coat variation, as well as habitat clutter, severe domain shifts, and a wide range of morphologies across species. These factors make animal Re-ID inherently more complex and elevate transferability as an important concern, necessitating rigorous benchmarking protocols to evaluate generalization and operational reliability.

2.1. Existing surveys on automated animal re-ID

To date, only one dedicated survey has systematically examined automated animal Re-ID: the work of Ravoor et al. [3]. Their survey reviews deep-learning-based approaches up to mid-2020, focusing on multi-species Re-ID, connections to animal tracking, and parallels with person Re-ID, with particular emphasis on cross-camera scenarios. It also integrates insights from the tracking literature and identifies open-set identification as an important future direction. At the time of its publication, the field was dominated by CNN-based architectures, species-specific datasets, and single-modal pipelines. Transformer-based models, attention mechanisms, multimodal approaches, including VLMs and MLLMs, and universal cross-species Re-ID frameworks had not yet appeared for animal Re-ID. Since then, the methodological landscape has evolved considerably: new datasets and benchmarks have been released, evaluation protocols have diversified, and datasets have become larger, more complex, and increasingly time-aware. These developments motivate our survey, which reviews research progress from 2020 to 2025 and highlights emerging directions such as cross-species generalization, transferability, multimodal modeling, and visual-language integration.

In addition to this early animal-focused survey, a more recent work [56] provides a broader overview of transformer-based approaches across multiple Re-ID domains, including person, vehicle, object, and animal Re-ID. However, its coverage of animal Re-ID is necessarily brief, organizing methods primarily around global body images, key local regions, and auxiliary cues. As acknowledged by the authors, transformers are still seldom explored in animal Re-ID, and the survey covers only a limited set of methods and datasets. Due to its timing and focus, it does not include several studies and emerging directions that have gained prominence more recently, including CNN-based approaches, novel strategies beyond traditional metric learning, unsupervised and self-supervised Re-ID, domain adaptation, and multimodal Re-ID with visual-language integration. It also does not address newer evaluation paradigms, such as time-aware evaluation, cross-species generalization, and unified networks capable of handling any animal species. In contrast, our survey provides a comprehensive and detailed review of these recent developments, emphasizing methodological novelties, first-time implementations, and limitations for each animal Re-ID approach.

3. Survey methodology

Our survey follows a structured multi-stage screening pipeline, summarized in Fig. 1. We conducted a database search in *Scopus* covering the period January 2020 to September 2025. The query was applied to the title, abstract, and keyword fields and included both hyphenated and non-hyphenated variants of “animal re-identification”, as well as terminology commonly used in ecological literature (e.g., “individual identification”). The search was restricted to the Computer Science subject area to reduce irrelevant documents. After duplicate removal, the retrieved records were manually screened.

A publication qualified for inclusion if it examined automated animal Re-ID using images or videos as the primary modality, with optional additional modalities (e.g., metadata, text, or alternative imaging technologies), and if its main objective was the identification of individual animals using machine/deep learning techniques. Importantly, studies primarily centered on animal detection or tracking, where Re-ID served only as a secondary component (e.g., [57,58]), were considered out of scope. This filtering ensures that the final set of works reflects advances in automated, model-based identification rather than manual identification or ecological analysis alone.

Following the screening and selection process, the collected papers were grouped into three categories based on their contributions: *(a) Methods:* works proposing new architectures or learning formulations for animal Re-ID, including, for example, CNNs-based, transformer-based architectures, metric learning, and multimodal learning; *(b) Datasets:* papers introducing new animal Re-ID datasets in e.g., wildlife, zoo, farm, or small-species contexts, typically providing identity annotations, metadata, and defined evaluation splits; *(c) Benchmarks and Toolkits:* papers proposing evaluation protocols, unified toolkits, or cross-species benchmarks that standardize preprocessing, dataset loading, splitting strategies, and performance assessment across multiple existing datasets. Some works span multiple categories, for example, by proposing a new method while releasing a new dataset, or by combining a benchmark protocol with baseline model evaluations.

3.1. Decoding methods, datasets and benchmarks

In reviewing the methodological contributions of animal Re-ID studies, we adopt a structured, in-depth approach that goes beyond summarizing high-level ideas. For each work, we explain its relevance to our survey, analyze technical contributions in detail, and highlight the core takeaway. This includes the underlying architectures or learning formulations, the problem settings addressed (e.g., images or videos, single- vs. multi-species), and the innovations introduced, such as new algorithms, post-processing strategies, evaluation protocols, as well as limitations and weaknesses. This comprehensive methodological review enables our survey to systematically capture novelties, first-time implementations, and practical implications of each approach, providing researchers with detailed insights into SOTA techniques and identifying gaps for future exploration.

For papers introducing datasets and benchmarks, we follow a similarly structured and detailed review strategy. Beyond summarizing the dataset, we critically examine key characteristics that are relevant for animal Re-ID research. These include the species covered, the number of individual animals, the type of environment where data were collected (e.g., wild, zoo, underwater, or internet-sourced), the number of images or videos, the time span over which data were gathered (ranging from days to multiple years), and whether the dataset spans multiple locations or events. We also consider practical aspects such as public availability and any accompanying metadata or standardized splits. This approach allows our survey to provide a comprehensive understanding of existing resources, highlighting strengths, limitations, and opportunities for future dataset development in animal Re-ID.

This survey is guided by several key questions: *i)* which architectures and learning paradigms have been proposed for animal Re-ID; *ii)* what problem settings are studied in the literature, and what technical challenges arise in each; *iii)* what novel algorithms, modeling strategies, and evaluation protocols have emerged since 2020; *iv)* what strengths, limitations, and practical weaknesses characterize existing methods; *v)* what

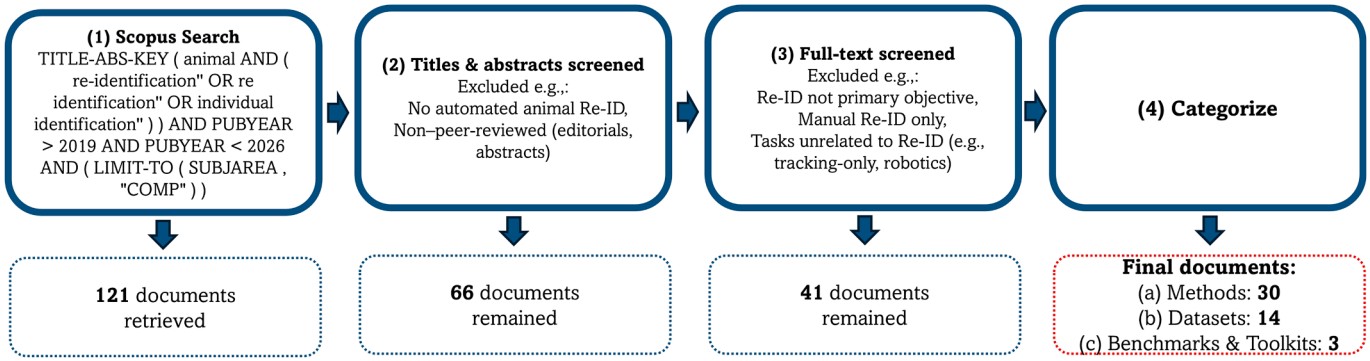

Fig. 1. Survey methodology pipeline, illustrating the staged process of database querying, screening, eligibility assessment, and final categorization of included papers.

species, environments, dataset scales, and annotation types are covered by available datasets, and to what extent these resources are standardized and publicly accessible; *vi)* what methodological and technological trends, as well as emerging research directions, can be observed; *vii)* what benchmarking frameworks and evaluation toolkits exist, and how they support reproducibility and standardized assessment; and *viii)* how current methods can be organized into a coherent taxonomy that clarifies the methodological landscape of animal Re-ID.

4. Animal re-ID method landscape

This section provides a high-level overview of the methodological landscape in automated animal Re-ID from 2020 to 2025. Recent progress in this field has produced a diverse ecosystem of approaches, driven by challenges such as data scarcity, information fusion, domain shift, annotation costs, and the need for robust and transferable systems. We organize this landscape along five aspects: (1) feature representation (4.1), (2) metric learning (4.2), (3) video and temporal modeling (4.3), (4) multimodal learning (4.4), and (5) alternative or additional training regimes (4.5). Feature representation encodes visual characteristics into discriminative embeddings. Metric learning defines how these features are compared and structured in the embedding space. Video and temporal modeling leverage sequential information to improve robustness under motion, occlusion, and pose variation. Multimodal learning expands the information sources beyond raw appearance, integrating complementary cues such as pose, motion, or text. Alternative or additional training regimes explore self-supervised, unsupervised, and domain-adaptive strategies that enhance scalability and generalization. These categories are not mutually exclusive; many studies integrate multiple paradigms, but they provide a useful conceptual structure for organizing the field. Fig. 2 provides a visual summary of this taxonomy and illustrates representative animal datasets across major habitats.

4.1. Feature representation

Techniques for extracting discriminative and invariant embeddings from animal imagery, including CNNs, transformer/attention models, and pose- or pattern-aware architectures, are foundational for representing animal identity under real-world shifts (e.g., pose, illumination, species, background). In the context of feature representation, three primary families of methods can be highlighted:

CNN-based. Convolutional architectures augmented with part-aware or attention mechanisms have proven effective for animal Re-ID, e.g., in [13,23,60], while modules such as posture-guided attention and Spatial Transformer Networks (STN) [61] further enhance feature robustness and discriminability under challenging scenarios, as observed in [27,39,62].

Transformer-based. Transformers and hybrid CNN-Transformer architectures capture long-range context and cross-view consistency that are less accessible to purely CNN-based Re-ID methods [20,28,30]. Within this family, ViT backbones enhanced with camera or site tokens, deformable attention mechanisms that selectively focus on the most informative regions, and multi-granularity feature fusion that integrates representations across layers or scales further improve robustness to viewpoint changes, background clutter, and illumination variation. These transformer-driven strategies have demonstrated strong gains in feature quality and discriminability across diverse animal Re-ID settings [24,63,64].

Local descriptors and geometry-based. Methods such as [65,66] exemplify this direction by extracting distinctive local texture patches from the animal's surface appearance, combining affine-invariant region detectors and deep local descriptors with geometry-consistent feature aggregation and Fisher Vector aggregation, enabling scalable Re-ID across taxa without species-specific retraining.

Overall, CNN-based backbones remain central to feature learning in animal Re-ID and are widely used across many recent approaches [13,21,23,30,31,35,39,60,62,67]. However, cross-domain or visually heterogeneous habitats often benefit from pipelines that integrate local-descriptor aggregation (e.g., keypoint or pattern-based matching) or attention-driven and transformer models that capture broader contextual relations. In addition, geometric alignment and multi-image feature aggregation strategies further enhance identification in patterned species, where enforcing spatial or temporal consistency across views reinforces appearance-based cues.

4.2. Metric learning

Metric learning forms a core component of animal Re-ID, where appropriately designed metric heads transform generic feature representations into discriminative embedding spaces suitable for similarity-based matching. Many animal Re-ID pipelines employ Siamese or triplet-network architectures trained with distance-based objectives such as contrastive or triplet loss, often paired with semi-hard negative mining to improve training stability and embedding quality [13,21,24,40,64,68].

4.3. Video and temporal modeling

Although most animal Re-ID research focuses on still images, several works and datasets leverage temporal continuity in videos to enhance identity embeddings, e.g., Chan et al. [16], Lamping et al. [24], Williams et al. [25], Yu et al. [28]. These methods use temporal cues not for explicit tracking but to stabilize representations under pose variation, occlusion, or motion blur. For example, Williams et al. [25] aggregates information across short video segments (tracklets) to obtain more consistent identity features and uses temporal constraints to merge visually fragmented tracks. Such approaches demonstrate how short-term

Fig. 2. Overview of the animal Re-ID landscape. (a) Methodological taxonomy summarizing recent approaches. **(b)** Representative examples from major animal Re-ID datasets [13,14,17,59], illustrating the diversity of species, habitats, capture conditions, and visual characteristics encountered.

temporal structure can improve identification without requiring multi-camera or long-term tracking. Similarly, Yu et al. [28] illustrates how fusing short-term sequences from multiple cameras can strengthen identity embeddings, with both supervised and self-supervised approaches benefiting from the additional temporal context.

4.4. Multimodal learning

Animal Re-ID increasingly benefits from leveraging multiple, complementary sources of information. These include purely visual cues such as appearance, color patterns, and pose or keypoint-based features, as well as cross-modal signals like textual descriptions or semantic priors from language models. Integrating these modalities enables richer representations and often improves generalization and robustness to occlusion, viewpoint variation, and domain shifts.

One prominent example is Jiao et al. [18], which proposes a universal animal Re-ID framework for unseen species that integrates visual prompts from image triplets with textual guidance generated by GPT-4 through CLIP's shared language-vision embedding space. A text-guided attention module refines the visual embeddings used in triplet-based learning, enabling strong zero-shot adaptation to novel species and establishing a new benchmark for open-world, multimodal animal Re-ID.

By contrast, several works instead fuse multiple cues derived from visual information. Perneel et al. [30] fuse complementary visual-derived cues: pose, behavioral metadata (e.g., standing, lying left, lying right), and spine orientation to achieve robust identification under varying activities. Other studies adopt similar ideas by integrating structural or texture-based information: for instance, Moskvyak et al. [31] incorporates facial and body landmarks encoded as heatmaps, while works such as [22,35,65,66] leverage segmentation masks and pelage pattern textures.

4.5. Alternative / additional training paradigms

Several alternative or complementary training regimes have emerged to reduce reliance on manual labels, improve generalization, and support scalable deployment in animal Re-ID.

Few-shot learning. Methods aim to identify new individuals when only a small number of labeled images are available per class. Wahltinez and Wahltinez [27] exemplify this setting by demonstrating that a metric-learning model can be rapidly adapted to new species using only a few annotated examples per individual, achieving strong performance across both invertebrate and mammalian datasets. This illustrates the potential of few-shot approaches to reduce annotation effort while maintaining reliable identification across diverse taxa.

Self-supervised learning (SSL). SSL approaches can produce discriminative embeddings for animal Re-ID without relying on full identity annotations. For instance, Yu et al. [28] leverages tracklet-based pseudo-labels with the normalized temperature-scaled cross-entropy loss (NT-Xent [69]) to train identity embeddings across multiple cameras, enabling accurate individual identification without directly using ground-truth identities.

Unsupervised learning. Such methods aim to learn discriminative animal embeddings without relying on any identity annotations, which is particularly valuable when labeling is expensive or infeasible. Zhang et al. [29] introduce the first unsupervised animal Re-ID method. This method employs feature-aware noise contrastive learning, combining pseudo-labeling via clustering and dual contrastive losses to encourage robust embeddings.

Content-based image retrieval (CBIR). Such studies adapt classical image retrieval principles to animal Re-ID, avoiding deep metric-learning objectives entirely. For instance, Nepovinnykh et al. [22,65,66] exemplify this direction by combining affine-invariant local features, deep descriptors, and Fisher Vector aggregation to enable modular, training-light pipelines.

Semi-automated / Human-in-the-loop Systems. Combines manual annotation (e.g., attribute coding) with automated feature extraction and matching to reduce labeling effort and support the identification of new individuals under limited labeled data. For example, Kulits et al. [33] provides an intuitive annotation interface and semi-automated matching: the system detects and crops elephants' ears, extracts contour features, and presents ranked candidate matches for human confirmation, enabling large-scale monitoring by non-experts.

Domain Adaptation. DA in animal Re-ID is typically performed as unsupervised adaptation using unlabeled target data. For example, Dubourvieux et al. [26] proposes a cumulative multi-domain adaptation framework that preserves source performance while improving specialization and generalization across multiple farms. The method in-

Table 1
Animal Re-ID methods by backbone, key characteristics, and multi-dataset evaluation. Abbreviations: ML: Metric Learning; FSL: Few-Shot Learning; SSL: Self-Supervised Learning; UL: Unsupervised Learning; DA: Domain Adaptation; MML: Multimodal Learning; VBR: Video-Based Re-ID; CBIR: Content-Based Image Retrieval; HIL: Human-in-the-Loop.

Ref.	Backbone	Key Characteristics / Learning Paradigm	Multiple Datasets?
[21]	CNN	ML, Siamese network, triplet loss, segmentation and pelage-pattern	×
[35]	CNN	ML, kNN matching, triplet loss, pose variance loss	✓
[31]	CNN	ML, body landmark heatmaps, center loss, triplet loss	×
[13]	CNN	ML, pose-aware feature learning, attention mechanism, triplet loss	×
[60]	CNN	ML, attention mask from Faster R-CNN [70], Cascade R-CNN [71], RetinaNet	×
[23]	CNN	ML, behavior categories (standing, sitting, lying), pose-constrained multi-similarity loss, based on OSNet [72]	✓
[30]	CNN	ML, behavior and orientation metadata, pose-invariant matching	×
[67]	CNN	ML, lightweight, triplet loss, raw, segmented, and shape-based inputs	×
[39]	CNN	ML, global and part-based features with spatial attention	×
[62]	CNN	ML, lightweight architecture based on ShuffleNet v2 [73] and BNNeck [74]	✓
[32]	CNN	ML, lightweight Siamese framework, thermal images, knowledge distillation	×
[64]	CNN + ViT	ML, deformable attention, triplet loss with hard mining	×
[20]	ViT	ML, cross-attention, global-local feature integration	✓
[63]	ViT	ML, global ViT and part-aware multi-granularity features	×
[24]	ViT	ML, head-based transformer, semi-hard triplet mining, YOLOv5 cropping	×
[19]	ViT	ML, oriented bounding box detection	×
[75]	CNN + ViT	ML, hierarchical spatial-frequency fusion transformer with cross-attention	✓
[68]	CNN	ML, systematic benchmarking of triplet vs. Proxy-NCA [76]	✓
[40,55]	CNN	ML, benchmarking Siamese, triplet, and contrastive learning	✓
[25]	CNN	VBR, classifier-based clustering with temporal context	×
[18]	CLIP + ViT	MML, CLIP [77] and GPT-4 guided multimodal embeddings	✓
[27]	CNN	FSL, triplet loss and semi-hard mining	✓
[28]	CNN	SSL, tracklet-based pseudo-labels, NT-Xent loss [69]	×
[29]	CNN	UL, feature-aware noise contrastive learning, DBSCAN clustering	×
[65]	CNNs	CBIR, segmentation with Mask R-CNN [78], pelage-pattern extraction with U-Net [79], affine-invariant descriptors [80,81], feature aggregation with Fisher Vectors	×
[22,66]	CNN	CBIR, local pattern similarity with global geometric consistency using point matching and homography estimation	✓
[33]	CNN	HIL, semi-automated, Faster R-CNN [70] for ear detection, CurvRank [82], ear contour matching	×
[26]	CNN	DA, unsupervised multi-domain adaptation with pseudo-labeling, domain-specific batch normalization, and source-guided loss, no metric learning	×

tegrates source-guided training, domain-specific batch normalization, and automated clustering hyperparameter tuning, enabling scalable adaptation across diverse farm environments.

5. Detailed review of the methods

This section presents an in-depth discussion of the works in each methodological category, highlighting their architectures, technical contributions, novelties, and limitations. Furthermore, Table 1 provides a structured overview of these studies, reporting the main methodological components (e.g., backbone architecture, loss formulation, feature representations, and key components) as well as the evaluation setup, including whether the methods were tested across multiple datasets. The text follows the order of aspects from the previous section, with works selected based on their main contributions. The datasets, evaluation metrics, and data splits used in these works, as well as studies focused exclusively on such aspects, are detailed in the next section.

Among the earliest **CNN**-based Re-ID frameworks, Siamese networks were widely adopted to model visual similarity. For example, Nepovinnykh et al. [21] proposes a framework based on segmentation and pelage pattern matching. The pipeline first isolates the seal using DeepLab [83], followed by a filter-based method [84] to extract pelage patches. A Siamese network trained with triplet loss learns patch-level similarities, and a topology-aware matching step preserves spatial consistency across the animal's body. Although the method demonstrates strong performance on its target dataset, it depends heavily on accurate segmentation and high-quality masks for reliable deployment.

Subsequent studies explored pose-invariant solutions. Moskvyak et al. [35] introduce a CNN-based embedding framework that combines triplet learning with an additional pose variance regularization term, encouraging embeddings of the same individual under different viewpoints to cluster tightly. In a related work, the authors incorporate body landmark heatmaps into feature learning to explicitly guide pose-aware representation [31]. By augmenting RGB inputs with landmark likelihood maps and employing joint metric-learning objectives, the model improves robustness to viewpoint variation and visual ambiguity. While these approaches demonstrate notable gains over baseline CNN models, they are evaluated primarily on manta ray datasets, leaving cross-species generalization an open question.

Other approaches extended pose-awareness by incorporating explicit part-based representations under extreme viewpoint variation. Li et al. [13] integrate pose keypoint estimation into a deep embedding framework, extracting both global and part-level features guided by detected body regions. By jointly training these representations with metric learning objectives, the model improves robustness to pose and camera variation. The method is evaluated under both manually normalized ("plain") and fully automatic ("wild") settings, consistently outperforming prior baselines, although a noticeable performance drop remains in realistic wild scenarios.

To reduce reliance on manually annotated keypoints, subsequent work introduced attention-based mechanisms that automatically highlight informative animal regions. In contrast to pose-annotated approaches such as [13], Cheng et al. [60] propose DeFAt, an additive masking module that leverages intermediate feature maps from standard object detectors to emphasize relevant animal regions during embedding learning. The module improves robustness to background clutter without requiring explicit keypoint annotations and is compatible with various detection and Re-ID backbones. While it demonstrates consistent

gains over keypoint-based baselines, its effectiveness remains dependent on reliable object detection.

Building on pose- and region-aware designs, later works explicitly modeled posture categories to better handle intra-class variation caused by animal poses. He et al. [23] propose MPFNet, a posture-aware architecture that distinguishes between standing, sitting, and lying poses to mitigate posture-induced variation, achieving improvements over prior pose-guided methods on the same datasets. The framework combines pose-specific feature extraction with a posture discriminator that guides feature fusion, and introduces a pose-constrained metric learning objective to enhance intra-class compactness. Although effective in addressing pose diversity, the method relies on a separately trained posture classifier, is limited to coarse pose categories, and introduces additional computational overhead.

A similar idea has been explored in livestock Re-ID, where [30] proposes an orientation- and behaviour-aware embedding framework for Holstein-Friesian cattle, incorporating pose-related metadata into the matching process. By conditioning similarity comparisons on estimated orientation and behavioural state, the method substantially improves robustness to viewpoint variation in unconstrained farm environments. Although effective, its performance depends on reliable estimation of auxiliary cues and may degrade in visually uniform or low-contrast scenarios.

For smaller animals or insects, CNN-based models remain effective for feature extraction, although limited dataset size and image variability constrain generalizability. Borlinghaus et al. [67] introduce BumbleNet, a lightweight embedding framework tailored to Bombus terrestris, and compare raw-image, segmented-image, and shape-based inputs. Their results indicate that CNN-based embeddings outperform size-based metrics, with raw-image models achieving the strongest performance. However, reliance on raw imagery may introduce sensitivity to spurious cues, and the findings may not readily generalize to larger datasets, long-term monitoring, or other insect species.

Attention mechanisms combined with global and part-based representations enable CNNs to better capture discriminative regions under challenging poses or occlusions. In this line, Chen et al. [39] introduces the Global and Part Feature Network (GPN) and its spatially-aware variant GPN-ST for cattle face Re-ID. These models jointly learn holistic and local representations, while GPN-ST further integrates a spatial transformer module to adaptively emphasize informative regions. Even though attention improves feature localization, extreme poses or heavy occlusion can still hinder precise alignment.

Lightweight CNN architectures have also been explored to enable efficient deployment on edge devices while supporting the identification of new individuals without retraining. In this context, Wang et al. [62] proposes ShuffleNet-Triplet, a compact embedding framework based on ShuffleNet v2 [73] designed for resource-constrained environments. By combining metric learning with classification-based supervision, the method produces discriminative embeddings that allow query samples to be matched against a gallery without retraining. The framework demonstrates strong efficiency and competitive accuracy on livestock datasets. However, the method still faces challenges in distinguishing highly similar individuals, relies on relatively limited datasets for benchmarking, and focuses solely on visual cues without incorporating multimodal information. Moreover, its robustness under diverse and uncontrolled farm environments remains largely untested.

Another lightweight Re-ID system is presented in Ashok Kumar et al. [32], combining YOLOv5-based knowledge distillation for detection with a Siamese network for individual recognition. The approach is evaluated on thermal imagery of elephants and achieves high accuracy while enabling real-time, resource-efficient deployment for wildlife monitoring and conflict mitigation. By leveraging thermal imaging, the method addresses the practical challenge of reliable animal identification under low-visibility conditions. However, reliance on thermal sensors may limit applicability in environments where such equipment is unavailable, and the pairwise Siamese matching strategy may scale poorly to very large populations or complex multi-individual scenarios.

A synthesis of these CNN-based approaches shows a clear evolution from early Siamese formulations toward more specialized feature representations that incorporate spatial, structural, and contextual cues. Several works enhance robustness to pose and viewpoint variation through pose-aware embeddings, part-based representations, and attention mechanisms that focus on discriminative regions. Other studies further improve performance by integrating auxiliary information such as body landmarks, behavioral metadata, or orientation cues, enabling more reliable matching in unconstrained settings. In addition, lightweight CNN architectures and embedding-based retrieval frameworks support efficient deployment and allow new individuals to be added without retraining. Collectively, these developments demonstrate that CNN-based methods can effectively capture discriminative identity cues while adapting to diverse animal species and application scenarios.

Despite these advances, CNN-based animal Re-ID methods exhibit several common limitations. Many approaches rely on additional supervision signals such as segmentation masks, keypoints, landmarks, or metadata, making them sensitive to the quality and availability of such annotations. Several methods are evaluated on limited datasets or specific species, raising concerns about cross-species generalization and robustness in real-world conditions. Performance can degrade under challenging scenarios such as occlusion, extreme poses, low-quality imagery, or visually similar individuals. Furthermore, certain designs introduce additional computational overhead or depend on reliable detection and preprocessing steps, which may limit scalability and deployment in unconstrained environments.

Different from CNN-based animal Re-ID, **transformer** architectures have been explored to enhance feature representations and aggregate information across multiple images. For example, Li et al. [64] presents a hybrid CNN-transformer framework that employs deformable attention to refine and fuse multi-image features for animal Re-ID. The method is trained with metric-learning objectives and evaluated on the ATRW dataset [13], where it improves performance over several earlier models. Interestingly, although transformer-based fusion is designed to emphasize informative regions, empirical results show that simple mean fusion performs better. This observation suggests that current transformer fusion strategies may still require further refinement or additional supervision to fully exploit feature importance.

To capture both global and local dependencies, cross-attention mechanisms have been explored in transformer-based animal Re-ID. Zhang et al. [20] propose the CATLA transformer, which introduces a hierarchical locally aware network redesign that integrates global and local token representations to better model animal body structure and pose variation. Evaluated across multiple species and tasks, including full-body and face Re-ID, CATLA achieves competitive performance and demonstrates improved generalization across viewpoints. However, the model remains challenged by species with limited distinctive visual cues, such as chimpanzees with highly similar facial and fur patterns.

By combining global self-attention with part-level embeddings, transformers can capture both holistic contextual information and fine-grained identity cues in patterned species. Bai et al. [63] investigate transformer-based architectures, including ViT for global feature modeling, as well as the Multi-Granularity Network (MGN) [85] and the Neighbor Transformer (NFormer) [86]. Both MGN and NFormer were originally proposed for person Re-ID: MGN extracts complementary global and part-level representations, while NFormer explicitly models interactions across input images rather than processing a single image. The study shows that a ViT + MGN integration, which combines transformer-based global representations with part-aware embeddings, improves Re-ID performance on Amur tiger datasets. However, the evaluation is limited to captive and camera-trap imagery, and some transformer variants (e.g., NFormer) require additional fine-tuning to achieve stable performance. Moreover, despite focusing on the same species, the

study does not provide a comparison with earlier attention-based approaches such as DeFAt [60].

Transformers can also be trained on specific anatomical regions to leverage distinctive visual cues for smaller animals. Lamping et al. [24] apply a ViT architecture to head-focused imagery extracted from video sequences of chickens, exploiting features such as combs and wattles for individual identification. Their experiments compare transformer models with CNN baselines and show that transformer-based embeddings can effectively capture discriminative head features. However, the approach focuses exclusively on head regions and may overlook informative cues from the body or plumage. Moreover, because head crops were manually curated and filtered using a high-confidence detector, the evaluation is constrained, and practical deployment would require reliable automatic head detection prior to Re-ID.

Orientation-aware transformer pipelines have also been explored to improve robustness in livestock Re-ID scenarios. Odo et al. [19] propose a ViT-based framework for pig identification that integrates orientation-aware detection to align animals more accurately in top-down farm imagery. By using oriented bounding boxes instead of conventional axis-aligned detections, the approach reduces background interference and improves identification accuracy under both open- and closed-world evaluation settings. This design choice proved a crucial improvement in performance and highlights that current CNN- and transformer-based Re-ID models still struggle to suppress background noise without geometric alignment. Experiments across datasets collected from commercial farms further demonstrate improved cross-site robustness. However, scalability remains limited by the small number of farms considered, and the method relies on accurate orientation-aware detection.

Integrating spatial and frequency information has recently been explored to enhance transformer embeddings for challenging cross-species Re-ID scenarios. Zheng and Wang [75] propose a hierarchical spatial-frequency fusion transformer that combines spatial CNN features with complementary frequency-domain representations within a unified CNN-transformer architecture. By jointly modeling multi-scale spatial structure and frequency cues, the framework improves robustness to noise, appearance variation, and inter-species differences. Experiments across multiple datasets show SOTA performance and strong cross-domain generalization, while maintaining high inference efficiency. However, the approach introduces additional feature extraction and fusion stages, increasing architectural complexity compared with standard transformer pipelines.

Beyond individual designs, these transformer-based approaches highlight a shift toward richer feature representations that integrate global and local information. Compared to CNN-based methods, they leverage self-attention mechanisms to capture broader contextual dependencies and fine-grained spatial relationships. Several works demonstrate improved robustness to pose variation and complex appearance patterns, particularly when combined with multi-granularity representations. Overall, these approaches show the potential of transformer-based models to improve representation quality in animal Re-ID.

Despite their strong performance, transformer-based animal Re-ID approaches present several practical challenges in real-world ecological and agricultural settings. These models typically require large and diverse training datasets and benefit from pretraining, which may not be readily available for many species where labeled data are scarce and costly to obtain. In addition, their higher computational and memory demands compared to lightweight CNN-based methods can limit real-time deployment on edge devices commonly used in camera traps, farms, or remote monitoring systems. Furthermore, transformer architectures often rely on carefully tuned training strategies and large-scale data distributions, which may not transfer well across species, habitats, or acquisition conditions. These factors highlight a trade-off between representational power and deployability that remains an open challenge for practical animal Re-ID applications.

Metric learning plays a central role in animal Re-ID, training networks to produce embeddings that bring images of the same individual closer while pushing apart images of different individuals. Various loss functions have been systematically evaluated, often in combination with different backbones, providing practical guidance for model selection.

For instance, Dlamini and van Zyl [68] systematically compares class-aware and pairwise metric-learning losses across several CNN architectures and multiple wildlife datasets, including settings with very limited samples per individual. Their results show that triplet loss generally outperforms Proxy-NCA [76], although the performance gap remains relatively small. While no single architecture-loss combination proves universally optimal, the study identifies VGG-11 with triplet loss as a strong baseline and offers useful benchmarking insights for metric-learning-based animal Re-ID systems. Similarly, Schneider et al. [40] systematically compare Siamese and triplet-loss networks for animal Re-ID across multiple species. Their results show that triplet-loss-based models consistently outperform Siamese formulations and can achieve high mean average precision even for visually challenging species. The study highlights the generalizability of deep metric learning for animal Re-ID and provides practical benchmarking insights for dataset preparation and evaluation. In a follow-up study, the authors extend the comparison to include contrastive learning [55], again finding that triplet loss provides more stable and reliable performance across species.

These studies highlight the central role of metric learning in animal Re-ID and provide empirical guidance on the choice of loss functions and training strategies. Across multiple datasets and species, triplet-loss-based formulations consistently demonstrate strong and stable performance, often outperforming alternative objectives such as Siamese or contrastive learning. These findings reinforce the effectiveness of embedding-based similarity learning as a general framework for animal Re-ID and suggest that performance is more sensitive to data characteristics and training setup than to specific backbone architectures.

These benchmarking studies also reveal several limitations of metric-learning-based approaches. Performance differences between loss functions are often marginal, indicating that no single formulation is universally optimal across datasets and species. In addition, metric learning relies on carefully constructed training samples and sufficient intra-class variation, which can be challenging in datasets with limited examples per individual. Finally, while these studies provide useful guidance under controlled settings, their conclusions may not fully translate to more complex real-world scenarios involving domain shifts, open-set conditions, or large-scale deployments. Beyond image-based approaches, leveraging **temporal information in videos** allows animal Re-ID systems to exploit motion cues and temporal consistency, improving identity tracking across frames. Williams et al. [25] introduce Classifier-Based Clustering (CBC), a video-based Re-ID framework that combines appearance features with temporal constraints to assign consistent identities across tracklets. By preserving temporal information during clustering rather than collapsing tracks into single representations, CBC improves robustness in scenarios involving occlusion, motion blur, or interactions among multiple animals. Experiments show strong performance compared with alternative clustering strategies, highlighting the benefit of jointly exploiting spatial, temporal, and appearance cues in short video sequences. However, the method relies on approximate knowledge of the number of individuals, incurs additional computational cost due to iterative clustering, and is evaluated under relatively controlled detection conditions.

Although relatively limited in number, video-based animal Re-ID approaches highlight the importance of incorporating temporal information to complement appearance-based representations. By leveraging motion cues and temporal consistency across frames, these methods can improve identity association in challenging scenarios such as occlusion, interactions between multiple animals, and short-term appearance changes. This suggests that temporal modeling provides a valuable additional signal beyond static image features for animal Re-ID.

Current video-based approaches remain limited in scope and present several challenges. They often rely on assumptions such as approximate knowledge of the number of individuals and high-quality tracking or

detection, which may not hold in real-world deployments. In addition, the iterative or clustering-based formulations can introduce higher computational costs compared to image-based methods. The limited number of existing studies and their evaluation under relatively controlled conditions further highlight the need for more extensive investigation of temporal modeling in diverse and unconstrained environments. Addressing the open-world animal Re-ID problem, Jiao et al. [18] introduces the first universal framework for category-generalizable identification across diverse species. The approach adopts a **multimodal learning** paradigm that combines visual and textual guidance to adapt to unseen categories. Specifically, visual prompts derived from image triplets are integrated with a frozen CLIP encoder [77], while textual cues generated by GPT-4 guide attention toward discriminative visual features. This vision-language design enables zero-shot adaptation to novel species and establishes a benchmark for open-world animal Re-ID. Experiments demonstrate strong cross-species generalization and superior performance compared with domain generalization and conventional Re-ID baselines. However, the framework depends on large pre-trained vision-language and language models, requires triplet-based prompts and class-level labels, and incurs additional computational overhead, while current evaluations rely on relatively limited open-world test splits.

Multimodal approaches introduce a promising direction for animal Re-ID by extending beyond purely visual representations. By leveraging joint vision-language models, these methods enable open-world identification and adaptation to previously unseen species, addressing a key limitation of traditional closed-set Re-ID systems. This paradigm suggests that incorporating semantic information can enhance generalization and flexibility in cross-species scenarios.

Current multimodal approaches present several practical challenges. They rely heavily on large pre-trained vision-language and language models (e.g., CLIP and GPT-4), which introduces dependencies on external pretrained systems. In addition, the use of prompt-based learning and auxiliary supervision increases methodological complexity and computational overhead. The limited number of existing studies and their evaluation on relatively restricted open-world settings further suggest that the robustness and scalability of multimodal animal Re-ID remain insufficiently explored.

A **few-shot learning** framework that generalizes across phylogenetically diverse species, from invertebrates such as sea stars to large mammals, is introduced in Wahltinez and Wahltinez [27]. The method trains a deep embedding network using triplet loss so that images of the same individual map to nearby points in feature space, while different individuals are separated. Re-ID is then performed through nearest-neighbor search in this embedding space, allowing new individuals to be added to the gallery without retraining the model. Importantly, the study introduces a time-aware data-splitting strategy that groups images from the same acquisition event to prevent temporal leakage and produce more realistic evaluation protocols. Extensive image augmentation and dropout regularization help mitigate overfitting in the few-shot setting, while careful hyperparameter tuning is required to achieve stable performance. Although the method demonstrates strong few-shot and cross-species generalization, practical deployment may still require initial labeled examples per individual and can be sensitive to hyperparameter choices. Performance may also degrade in highly unconstrained field conditions, and scalability to very large populations remains an open challenge.

Although still limited in number, few-shot approaches introduce an important direction for animal Re-ID by enabling identification with minimal labeled data. By learning embedding representations that generalize across species and individuals, these methods allow new identities to be incorporated without retraining, which is particularly valuable in ecological settings where annotation is costly. This suggests that few-shot learning can provide a flexible and scalable alternative to traditional fully supervised pipelines.

Current few-shot approaches also present several limitations. They still require at least a small number of labeled examples per individual, and their performance can be sensitive to hyperparameter choices and training conditions. In addition, robustness may degrade under highly unconstrained field conditions, and scalability to very large populations remains challenging. The limited number of existing studies further indicates that the effectiveness of few-shot learning for animal Re-ID has not yet been extensively validated across diverse datasets and species.

For **self-supervised** Re-ID, Yu et al. [28] use tracklet-based pseudo-labels to train a CNN-based embedding network with the NT-Xent contrastive loss [69]. The approach exploits intra-tracklet consistency to learn discriminative representations without requiring identity annotations, while multi-camera tracklets further enhance feature quality for zero-label Re-ID. The learned embeddings are later evaluated with ground-truth cow identities to measure identification performance. Results show that both the self-supervised and supervised settings benefit from multi-view integration. However, reliance on tracklets, motion patterns, and farm-specific interactions may limit generalizability, highlighting the need for cross-site validation.

Self-supervised approaches offer a promising direction for animal Re-ID by reducing reliance on costly identity annotations. By exploiting temporal consistency and multi-view information, these methods can learn meaningful embedding representations without explicit labels, making them particularly attractive for large-scale or continuously collected datasets. This suggests that self-supervised learning can provide a scalable alternative to fully supervised pipelines in annotation-scarce environments.

Current self-supervised approaches also exhibit several limitations. Their performance depends heavily on the availability and quality of tracklets, as well as consistent motion patterns and controlled acquisition settings. As a result, generalization across different farms, camera setups, or environmental conditions may be limited. The small number of existing studies further indicates that the robustness of self-supervised animal Re-ID remains insufficiently explored.

The method in [29] represents the first **unsupervised** framework for animal Re-ID, targeting scenarios where identity labeling is expensive or unavailable. Unlike person Re-ID, where unsupervised approaches are widely studied, this work remains one of the few attempts in the animal Re-ID domain, with comparisons primarily drawn from person Re-ID methods adapted to animal datasets [87–90]. Their Feature-Aware Noise Contrastive Learning (FANCL) framework combines contrastive representation learning with clustering-based pseudo-label generation to enable training without identity annotations. By introducing feature-aware noise perturbations and consistency constraints, the method encourages embeddings to capture more global identity cues rather than relying on local visual patterns. Experiments show that FANCL achieves performance comparable to supervised baselines and surpasses existing unsupervised approaches, demonstrating promising scalability for annotation-scarce settings. However, the approach may still overemphasize activation-map regions at the expense of holistic context, and it has so far been validated only on a single species.

Unsupervised approaches represent an important direction for reducing reliance on identity annotations. By combining contrastive learning with clustering-based pseudo-labeling, these methods enable representation learning without manual labeling, which is particularly valuable in ecological scenarios where annotation is costly or infeasible. This suggests that unsupervised learning has the potential to scale animal Re-ID systems to larger datasets and previously unstudied species.

Current unsupervised approaches remain limited in scope and present several challenges. Their performance depends on the quality of pseudo-label generation and feature consistency, which may be difficult to maintain in highly variable or noisy data. In addition, the reliance on clustering and auxiliary constraints can introduce sensitivity to dataset characteristics and training dynamics. The evaluation on a single species further indicates that the generalization of unsupervised

animal Re-ID methods remains largely unverified across diverse datasets and conditions.

Despite the conceptual similarity between CBIR and Re-ID, **CBIR** approaches have rarely been applied to animal Re-ID. Nepovinnykh et al. [65] introduce NORPPA, a CBIR-inspired pipeline that exploits the permanent pelage patterns of Saimaa ringed seals for identification. The framework follows a modular retrieval pipeline in which the animal is first segmented, distinctive pelage patterns are extracted, and invariant local descriptors are aggregated to enable similarity-based matching. Because identification relies on descriptor retrieval rather than learned identity embeddings, new individuals can be added to the database without retraining, supporting scalable and partially human-in-the-loop workflows. NORPPA outperforms earlier methods such as [21]; however, its performance remains sensitive to image quality, occlusions, and segmentation accuracy, and the approach may transfer poorly to species with less distinctive or stable visual patterns.

Consequently, the NORPPA framework was extended to support broader cross-species applications by combining local pattern similarity with global geometric consistency [22,66]. This extension integrates local descriptor matching with geometric verification to enforce spatial consistency between candidate matches. By jointly considering appearance similarity and global structure, the approach improves retrieval accuracy compared with earlier baselines [21,65]. However, performance remains sensitive to image quality and the distinctiveness of species-specific patterns. These results highlight the potential of geometry-aware, multi-image Re-ID pipelines for scalable wildlife monitoring.

Taken together, CBIR-based approaches offer an alternative to metric-learning pipelines by relying on explicit pattern descriptors and similarity retrieval rather than learned identity embeddings. Unlike metric-learning methods, which require training data and optimization of embedding spaces, CBIR frameworks enable direct matching and allow new individuals to be added without retraining. This makes them particularly suitable for species with stable and distinctive visual patterns, where handcrafted or descriptor-based representations remain effective. These results suggest that CBIR methods can provide a complementary solution to deep metric learning in specific animal Re-ID scenarios.

CBIR-based approaches also exhibit several limitations. Their performance is highly dependent on image quality, accurate segmentation, and the presence of distinctive and stable visual patterns, which may not generalize across species. In addition, the reliance on handcrafted or descriptor-based representations can limit adaptability to complex variations in pose, illumination, and appearance. The limited number of studies further indicates that the scalability and robustness of CBIR-based animal Re-ID remain insufficiently explored in diverse and unconstrained environments.

A different line of research explores **semi-automated**, human-in-the-loop systems that combine manual annotation with automated visual matching to support large-scale wildlife monitoring and open-set evaluation [33]. The ElephantBook platform integrates with EarthRanger [91], a real-time conservation data management system, and uses computer vision methods to detect elephant ear regions and match distinctive ear contours for individual identification. By combining automated ranking with human verification, the framework improves usability and scalability for non-expert conservation practitioners. However, the approach remains sensitive to image quality, relies on accurate manual coding, and is primarily applicable to species with distinctive contour features, limiting broader generalization and fully automated open-set deployment.

Semi-automated, human-in-the-loop approaches provide a complementary paradigm to fully automated Re-ID systems by explicitly incorporating expert knowledge into the identification process. By combining automated ranking with manual verification, these frameworks improve usability and reliability, particularly in conservation settings where accuracy is critical and full automation may not yet be feasible. This suggests that hybrid systems can serve as a practical intermediate solution for large-scale wildlife monitoring. However, the approach remains sensitive to image quality, relies on accurate manual coding, and is primarily applicable to species with distinctive contour features. In practice, the need for human intervention can limit scalability and throughput in large-scale deployments, particularly when monitoring large populations or continuous data streams. Furthermore, dependence on consistent image acquisition and species-specific visual cues may restrict generalization across habitats and species, highlighting challenges for fully automated and broadly applicable open-set Re-ID systems.

While most animal Re-ID methods assume that training and test data originate from the same environment, real-world deployments often involve domain shifts between farms, camera systems, or habitats. **Domain adaptation** techniques aim to address this challenge by leveraging unlabeled target data to improve generalization while preserving performance on the source domain. In this context, Dubourvieux et al. [26] introduces Cumulative Unsupervised Multi-Domain Adaptation (CUMDA), a framework for Holstein cattle Re-ID that extends unsupervised domain adaptation from single- to multi-target scenarios by combining source-guided training and domain-specific batch normalization (DSBN). By enabling cumulative adaptation across multiple farms while maintaining source-domain knowledge, the approach improves robustness to domain shifts and reduces the catastrophic forgetting often observed in conventional UDA methods [92]. Experiments across heterogeneous livestock datasets demonstrate improved cross-domain generalization, although performance still lags behind fully supervised training due to large appearance variations and limited color cues. Nevertheless, CUMDA represents one of the first multi-target domain adaptation frameworks for animal Re-ID and offers a promising direction for scalable deployment in heterogeneous farm environments.

Domain adaptation approaches address a critical challenge in animal Re-ID by explicitly modeling domain shifts across environments such as farms, camera systems, and habitats. By leveraging unlabeled target data while preserving source-domain knowledge, these methods improve robustness to distribution changes in real-world deployments. However, their performance still lags behind fully supervised training, particularly under large appearance variations and limited color cues. In addition, their effectiveness depends on the quality of pseudo-labeling and the similarity between source and target domains, which may vary significantly in practice. The limited number of studies and evaluations across a small number of environments further indicates that the scalability and robustness of domain-adaptive animal Re-ID remain insufficiently explored.

6. Datasets and benchmarks

Studies reviewed in Section 5 have evaluated animal Re-ID models across a wide range of species, reflecting diverse appearance patterns and imaging conditions. Among large mammals, the Amur tiger represents one of the most established datasets [13,60], alongside elephants [32,33], lions, giraffes, and zebras [27,68]. Marine species such as manta rays and humpback whales have also been explored [35,40], while smaller animals like bumblebees [67], fruit flies [40], and chickens [24] highlight the adaptability of Re-ID methods to fine-grained recognition tasks. Notably, cattle Re-ID has become a major benchmark for livestock monitoring [26,39]. Such diversity demonstrates the growing generalization of animal Re-ID across species with different texture complexity, imaging modalities, and ecological settings (see Table 2).

Studies have leveraged a variety of input modalities to capture distinctive features. These include thermal imaging [32], pose information represented as keypoints [13], and face or body landmarks encoded as heatmaps [31]. Other approaches exploit segmentation masks and pelage or skin pattern textures to enhance identity discrimination [22,35,65,66]. In livestock monitoring, body keypoints capturing orientation (e.g., spine angle from tail implant to withers) and behavioral states such as standing or lying are also used [30]. Despite this diversity, most inputs remain RGB images/videos, typically processed

Table 2
Animal species used to evaluate the Re-ID methods reviewed in Section 5.

Species	Ref.	Species	Ref.
Elephant	[32,33,75]	Lion	[20,68]
Yak	[75]	Chimpanzee	[20,27,40,55,68,75]
Amur tiger	[13,20,23,27,40,55,60,63,64,68,75]	Zebra	[27,68]
Bumblebee	[67]	Giraffe	[27]
Fruit fly	[40,55,75]	Panda	[68,75]
Macaque	[75]	Red panda	[20,29]
Cattle	[26–28,30,39,62]	Dog	[23]
Chicken	[24]	Manta ray	[31,35]
Humpback whale	[35,40,55,75]	Ringed seal	[21,22,27,65,66]
Whale shark	[22,66]	Pig	[19,25]
Sea star	[27]	Golden monkey	[20]
Meerkat	[20]		

by CNN or transformer (ViT) backbones. Some recent methods leverage textual descriptions associated with images to enable multimodal Re-ID learning [18].

Below, we first introduce the evaluation protocols and metrics, and then review the existing datasets, benchmarks, and toolkits covered in our survey. A summary of these resources is provided in Tables 3 and 4.

6.1. Evaluation protocol and metrics

Animal Re-ID experiments typically adhere to a protocol that partitions the available data into training, validation, and testing subsets. The training set is used to learn discriminative representations, often supervised by individual identity labels, while validation images may be used for hyperparameter tuning or model selection. For testing, the data are further partitioned into two non-overlapping subsets: the gallery and the query. The gallery contains one or more reference images per individual, while the query set includes distinct images of the same individuals captured under different conditions (e.g., pose, viewpoint, or time). During evaluation, each query image is compared against all gallery images, producing a ranked list based on, e.g., visual similarity or feature-space distance. Correct matches appearing at higher ranks indicate better model performance.

Animal datasets often lack explicit camera information, so gallery-query splits are usually defined at the individual level rather than by camera ID, and are constructed randomly or in a time-aware manner to reduce overestimation bias. Moreover, due to strong appearance asymmetry in many species, such as when the left and right body sides or facial profiles differ substantially, some studies e.g., [13,37] treat the two sides as distinct identities. This design avoids false matches between visually dissimilar sides of the same animal and ensures more realistic performance estimates under asymmetric visual conditions. However, a limitation of this design is that models do not learn cross-side correspondences, which may hinder generalization in real-world scenarios where either body side could be visible.

Most studies adopt a closed-set evaluation protocol, where all test identities are also seen during training. In contrast, open-set Re-ID reflects more realistic field scenarios in which new, previously unseen individuals may appear at test time. In this setting, the system must not only identify known individuals but also correctly reject unknown ones, a capability explored, e.g., in [18,19,27,33]. Additionally, although time-aware Re-ID aspires to model appearance changes across time, such as those caused by growth, aging, or environmental variation, this capability remains largely conceptual in current systems. For instance, in Wahltinez and Wahltinez [27], the concept is applied as a dataset splitting strategy to prevent performance overestimation. Specifically, images collected during the same event are kept together in either the training or testing set, and separate time-aware splits are created for each collection event. This ensures that temporally correlated images of the same individuals do not leak between training and testing, simulating scenarios such as catch-and-release in ecological monitoring without explicitly modeling long-term appearance changes.

Performance in animal Re-ID is typically evaluated using two standard metrics: the Cumulative Matching Characteristic (CMC) and the mean Average Precision (mAP). The CMC curve measures the probability that a correct match appears within the top-k retrieved results for a given query. Formally, given Q query images, the Rank-k accuracy is computed as:

$$\text{Rank-}k = \frac{1}{Q}\sum_{q=1}^{Q} \mathbb{I}(\text{correct match appears in top-}k) \tag{1}$$

where $\mathbb{I}(\cdot)$ is an indicator function returning 1 if the correct identity appears among the top-k ranked gallery images, and 0 otherwise. Many studies, e.g., Odo et al. [19], Perneel et al. [30], Zuerl et al. [36] report Rank-1 CMC (the percentage of cases where the top-ranked gallery image corresponds to the correct identity) though some also include higher ranks (e.g., Rank-5 or Rank-10) for a more complete assessment.

The mAP metric, instead, evaluates both precision and recall by averaging the precision across recall levels and all queries, offering a more holistic measure of retrieval quality:

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{m_q}\sum_{k=1}^{n_q} P_q(k)\cdot \text{rel}_q(k) \tag{2}$$

where $P_q(k)$ is the precision at rank k, $\text{rel}_q(k)$ indicates whether the kth ranked gallery image matches the query, and m_q is the number of relevant gallery images for query q.

As noted earlier, most animal Re-ID datasets lack explicit camera information, so correctly matched instances are uniformly considered during evaluation. Moreover, for the animal datasets that include visually similar samples captured from slightly different viewpoints, the CMC-based Rank-k accuracy can sometimes be inflated by such easy matches, making mAP a more informative and robust measure of overall performance.

6.2. Review of the datasets

The datasets introduced between 2020 and 2025 can be broadly grouped into four categories reflecting their capture environments and biological diversity: *(a)* wildlife and semi-wild datasets, *(b)* zoo-based datasets, *(c)* farm, livestock, and domestic-animal datasets, and *(d)* invertebrate and small-species datasets.

6.2.1. Wildlife and semi-wild datasets

ATRW (Amur Tiger). Li et al. [13] is a large-scale, wildlife-specific benchmark for Amur tiger Re-ID, addressing the scarcity of high-quality datasets for non-domestic, non-rigid animal species. It comprises 8076 high-resolution video clips captured across multiple wild zoos using synchronized surveillance and SLR cameras, from which 3649 annotated bounding-box images representing 92 individual tigers were obtained. Because the left and right flanks of each tiger show distinct stripe patterns and are rarely visible together, ATRW treats each side as a separate entity, yielding 182 entities for Re-ID evaluation. The dataset includes identity labels, bounding boxes, 15-point pose keypoints, and view orientation annotations, together with benchmark protocols for detection, pose estimation, and both plain (manually cropped) and wild (automatically detected) Re-ID. Despite its limited number of individuals, an inherent constraint of endangered species, ATRW offers substantial variability in pose, illumination, background, and occlusion.

Table 3
Summary of recent animal Re-ID datasets (2020–2025), organized by habitat and acquisition context.

Dataset	Species	#IDs	#Images	Time Span	Source	Available Annotations	Open set?	Time Aware?	Publicly Available?
Wildlife and Semi-Wild Datasets									
ATRW [13]	Amur tiger	92	3649	–	Wild zoos	IDs, bounding boxes, pose keypoints, view orientation	✓	×	✓
YakReID-103 [37]	Yak	103	2247	–	Wild (highland pastures)	IDs, bounding boxes, direction-based pose (left/right)	✓	×	✓
SealID [14]	Saimaa ringed seal	57	2080	10 years	Wild (lake; boat surveys + camera traps)	IDs, segmentation masks, pelage patches, metadata (GPS, timestamps)	✓	×	✓
SeaTurtleID 2022 [17]	Sea turtle	438	8729	13 years	Wild (underwater snorkeling)	IDs, timestamps, masks, bounding boxes, body-part orientations	✓	✓	✓
Zoo-Based Datasets									
RedPanda43 [29]	Red panda	43	3487	–	Zoo	IDs; environment labels; train/query/gallery split	×	×	×
Common Toads [93]	Bufo bufo	376	7245	–	Field rescue operations (controlled single-animal videos)	IDs, bounding boxes, camera-angle labels	×	×	×
Mixed-Species Zoo [93]	Turtles, camels, penguins, goats, toads	160	825	Weeks	Zoo environment	IDs, species labels, manual bounding boxes	×	×	×
iPanda-50 [59]	Giant panda	50	6874	–	Captive (panda bases; streaming cameras)	IDs, bounding boxes, eye locations, curated keyframes	×	×	✓
PolarBearVidID [15]	Polar bear	13	138,363 frames	–	Six zoos	IDs, bounding boxes, 100-frame video sequences, camera metadata	×	×	✓
Farm, Livestock, and Domestic Animal Datasets									
Holstein Dairy Cow [94]	Holstein dairy cow	13	1485	-	Dairy farm	IDs only	×	×	✓
Holstein Face [39]	Dairy cattle	3000	130,000	–	Farm	IDs, face bounding boxes, viewpoints (L/F/R), imaging-condition metadata	✓	×	×
MultiCam Cows2024 [28]	Holstein-Friesian cattle	90	101,329	7 days	Farm (multi-camera CCTV)	Tracklet IDs, bounding boxes, camera metadata	×	✓	✓
MPDD-192 [23]	Dog	192	1657	–	Natural outdoor settings	IDs, posture labels (standing/sitting/lying)	×	×	✓
Invertebrate and Small-Species Datasets									
Honeybee [16]	Honeybee	181 / 4949	8962 / 109,654	12 days	Observation hive (controlled)	IDs (barcode), track IDs, timestamps, pose keypoints	✓	✓	✓
Animal-Identification-from-Video [5]	Pigs, koi fish, pigeons	93	20,490 clips (2,379 frames)	9–24 s per video	Pixabay videos (controlled, non-wild)	IDs, bounding boxes	✓	×	✓

SealID. Nepovinnykh et al. [14] targets the endangered Saimaa ringed seal, offering 2080 images from 57 individuals collected during annual molting seasons (2010–2019) in natural lake habitats. The dataset provides a curated database (430 images) and a query set (1650 images), capturing high variability in pose, illumination, pelage condition, contrast, and viewpoint. SealID includes expert-verified identity labels, pixel-level segmentation masks refined from Mask R-CNN [78] outputs, and a complementary patch dataset of pelage patterns to support fine-grained pattern matching in deformable, low-contrast species.

SeaTurtleID2022 (Loggerhead Sea Turtle). Adam et al. [17] introduces a long-span benchmark for sea turtle Re-ID, consisting of 8729 images of 438 loggerhead sea turtles collected over 13 years. Each image includes identity labels, timestamps, segmentation masks, bounding boxes, and body-part orientations. The dataset defines two ecologically motivated, time-aware splits: closed-set and open-set to reflect realistic population dynamics. Baseline evaluations are provided across classic feature-based methods (e.g., [95], SuperPoint [96]) and deep learning models (e.g., ArcFace [97], FaceNet [98], Mask R-CNN [78], Hybrid Task Cascade [99], Mask2Former [100]). Importantly, the study demonstrates that random splits inflate performance, underscoring the necessity of time-aware evaluation for wildlife Re-ID.

iPanda-50 (Giant Panda). Wang et al. [59] is a video-derived giant panda Re-ID dataset built from long streaming videos recorded at panda conservation and breeding centers. Key frames are extracted using SSIM-based filtering and manually verified by keepers to ensure identity correctness. Tight bounding-box cropping yields 6874 images of 50 pandas captured under varied poses, viewpoints, illumination, occlusions, and enclosure conditions. The dataset includes identity labels and eye-location annotations, reflecting the discriminative role of facial regions. Although organized as a closed-set identification task, iPanda-50 provides a challenging benchmark for fine-grained recognition of highly similar, non-rigid animals.

PolarBearVidID (Polar Bear). Zuerl et al. [15] is a video-based Re-ID dataset for a non-human species, introducing 1431 sequences (≈138k frames) of 13 individually identified polar bears recorded across six zoos. Videos were captured using tripod-mounted and handheld cameras positioned to cover large enclosure areas, resulting in high variability in pose, movement, lighting, occlusions, enclosure geometry, and viewpoints. Expert biologists manually annotated key frames with bounding boxes and identity labels; each labeled frame was expanded into an 8 s sequence (100 frames) via a detection–tracking pipeline. The dataset follows a closed-set protocol without temporal splits. Since it is video-

based, it enables motion-aware deep learning approaches for animal Re-ID in zoo and wildlife monitoring.

6.2.2. Zoo-based datasets

RedPanda43 (Red Panda). This custom red panda dataset [29] contains 3487 images of 43 individually microchipped animals captured in both indoor and outdoor zoo environments. It reflects moderate variation in pose, illumination, and background. Identities are split into 20 for training and 23 unseen individuals for testing, under a closed-set protocol. The dataset contains only identity labels; no bounding boxes, masks, or pose annotations are provided.

Mixed-species zoo and common toads. Fruhner and Tapken [93] introduce two benchmark datasets to study cross-species transfer in Re-ID. The first includes 7245 cropped images of 376 common toads captured during field rescue operations under varied viewpoints. The second is a mixed-species zoo dataset with 825 images of 160 individuals from five species (turtles, camels, penguins, goats, and toads). Both datasets contain identity labels and support the evaluation of human Re-ID models such as OSNet-AIN [101]. Notably, the study reveals that training on the heterogeneous multi-species dataset can improve performance on individual species compared to species-specific training alone, suggesting beneficial cross-species feature transfer. While the datasets expand the diversity of benchmarks available for animal ReID, they remain relatively small, and the toad dataset lacks a fixed train/test split, which might complicate direct reproducibility. Nonetheless, the work provides valuable insights into model adaptability, dataset design, and practical constraints when applying human Re-ID architectures to multi-species animal Re-ID.

6.2.3. Farm, livestock, and domestic animal datasets

Holstein dairy cow. Li et al. [94] introduce an early dairy cow dataset comprising 1485 annotated side-view images of 13 Holstein cows collected at a commercial farm under naturally varying lighting, contamination, and background conditions. Images are annotated only with identity labels. Despite its small scale, it is one of the first publicly available cow Re-ID datasets captured in unconstrained farm environments.

Holstein face. Chen et al. [39] present a large-scale cattle face dataset with 130,000 images from 3000 Holstein cows. Images were collected in an open-farm environment using multiple camera viewpoints and automatically cropped with an improved Faster R-CNN + NASNET-A detector. The dataset includes identity labels and left/frontal/right view annotations but lacks temporal metadata or long-term tracking.

MultiCamCows2024. Yu et al. [28] is a large, multi-camera Re-ID dataset with 101,329 images of 90 Holstein-Friesian cows collected over seven days using three ceiling-mounted CCTV cameras. It provides cropped images, full CCTV footage, bounding boxes, tracklet-level identities verified by humans, and camera-view metadata. The dataset captures realistic farm challenges such as varying viewpoints, lighting, movement patterns, and class imbalance.

YakReID-103. Zhang et al. [37] is the first yak Re-ID dataset, collected in outdoor highland pastures. It contains 2247 images of 103 yaks, annotated with bounding boxes, direction-based pose labels (left vs. right), and identity information. Left and right body profiles are treated as separate entities, resulting in 182 labeled identities. The dataset provides identity-disjoint train/test splits and includes both “simple” and “hard” subsets, though these difficulty levels are manually curated rather than temporally or environmentally defined.

MPDD-192 (Dog posture dataset). He et al. [23] introduces 1657 images of 192 individual dogs captured in natural outdoor settings across three major postures: standing, sitting, and lying. Images reflect substantial viewpoint variability and were curated by removing blurry or occluded samples. Each image includes identity and posture labels, making the dataset well/suited for evaluating pose robustness in full-body animal Re-ID. Despite its modest scale, it fills an important gap by providing one of the first multi-pose dog Re-ID datasets collected in unconstrained real environments.

6.2.4. Invertebrate and small-species datasets

Honeybee. Chan et al. [16] introduces two complementary datasets collected from 12 days of video captured at a hive entrance. The long-term dataset contains 8962 abdomen-aligned images of 181 barcoded bees, while the short-term dataset provides 109,654 images across 4949 tracklets with self-supervised pseudo-labels. These datasets enable evaluation across same-hour, same-day, and cross-day conditions and support supervised, self-supervised, and hybrid training regimes.

Animal identification from video. Kuncheva et al. [5] introduces an annotated video-based dataset comprising five short clips featuring pigs, koi fish, and pigeons, totaling 20,490 annotated instances across 93 identities. Each frame is labeled with identity tags and bounding boxes, and the videos present challenging conditions such as multiple animals per frame, frequent occlusions, strong intra-class variability, and occasional annotation noise. To avoid data leakage caused by near-duplicate adjacent frames, the authors provide frame-sequential train/test splits aligned with the temporal structure of each clip. The dataset is accompanied by baseline results using 26 classical classifiers on RGB, Histogram of Gradients (HOG), and Local Binary Pattern (LBP) features, as well as object detection and tracking benchmarks. While the dataset offers a standardized and transparent resource for evaluating animal Re-ID and tracking methods, it remains limited in scale, species diversity, and ecological realism due to its reliance on short, publicly sourced stock video footage rather than field-collected material.

Fig. 3 complements Table 3 by visually summarizing the landscape of animal Re-ID datasets in terms of identity count, image volume, and evaluation protocols. The figure shows clear domain-dependent differences: wildlife datasets are typically smaller but more likely to include open-set settings, zoo datasets are predominantly closed-set, farm datasets can be large-scale yet rarely support open-set or time-aware evaluation, and invertebrate datasets are heterogeneous in scale. Overall, the figure highlights that dataset growth has not been matched by equally broad adoption of open-set and time-aware protocols, underscoring an important benchmarking gap in the literature.

6.3. Review of the benchmarks and toolkits

Kuncheva et al. [41] present a benchmarking study for animal Re-ID and propose a general evaluation protocol suitable for any partially annotated video or image collection. Using the previously released Animal Identification from Video dataset [5], they benchmark 25 classifiers across five feature representations (RGB, HOG, LBP, autoencoder embeddings, and MobileNetV2). A key contribution is the use of frame-sequential train/test splits, which preserve temporal continuity and prevent data leakage from adjacent frames. Their results consistently show that simple linear models, particularly LDA with RGB color moments, outperform CNNs and MobileNetV2, even with augmentation and tuning. The authors also release all annotations and code, supporting reproducibility. Limitations include the small scale of the dataset (one short video per species), low overall accuracies caused by open-set and high-variability conditions, lack of temporal modeling beyond splitting, and the evaluation of only two deep learning architectures.

WildlifeDatasets [4] introduces a unified benchmarking toolkit addressing fragmentation and inconsistency across animal Re-ID evaluation. It provides standardized access to 31 publicly available wildlife datasets with automated downloading, preprocessing, and comprehensive support for closed-set, open-set, disjoint-set, and time-aware splits. The toolkit integrates a wide range of feature extractors, from local descriptors such as SIFT and SuperPoint to CNN and transformer

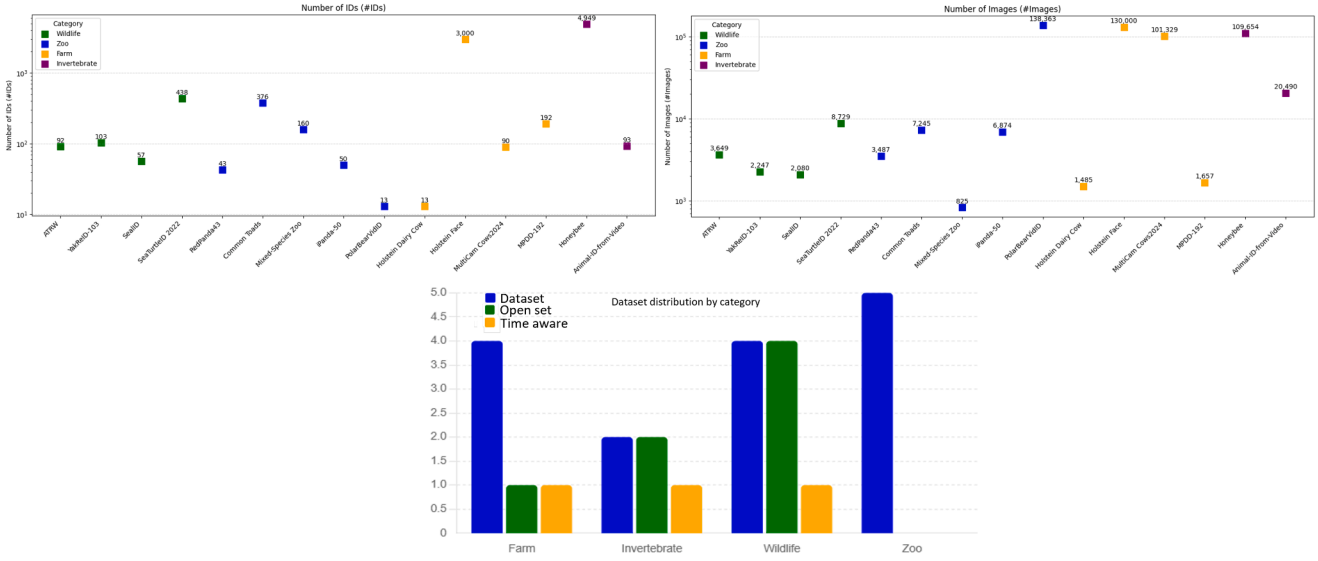

Fig. 3. Visual summary of animal Re-ID datasets across major acquisition contexts. The figure shows the number of identities and number of images for each dataset (top), along with the distribution of total datasets, open-set datasets, and time-aware datasets by category (bottom).

Table 4
Comparison of animal Re-ID benchmarking frameworks.

Benchmark	Purpose	Data Used/Provided	Protocols	Baselines	Code
Kuncheva et al. [41]	Evaluation protocol; classifier benchmarking	Uses 5-video dataset (93 IDs)	Closed-set; frame-sequential CV	25 classical + DL models	✓
Wildlife Datasets [4]	Unified toolkit; large-scale benchmarking	31 public datasets; MegaDescriptor	Closed-set, open-set, disjoint, time-aware	SIFT, SuperPoint, ArcFace, CLIP, DINOv2, MegaDescriptor	✓
Wildlife-71 [18]	Cross-species open-world benchmark	Wildlife-71 (71 categories; 67 train, 4 unseen test)	Open-world; support-triplet adaptation	TransReID, AGW, CAL, DGReID, UniReID	✓

architectures, and includes GPU-accelerated matching APIs. A central contribution is MegaDescriptor, the first foundation model for animal Re-ID, trained across 29 datasets and released in multiple sizes. MegaDescriptor achieves state-of-the-art performance and consistently outperforms generic vision models such as CLIP [77] and DINOv2 [102]. However, the framework still relies primarily on closed-set evaluations, performance remains dataset-dependent, and methodological choices (e.g., excluding overlapping datasets) may not generalize to field conditions. Proper metadata handling is also essential to avoid data leakage.

Jiao et al. [18] define the ReID-AW ("Re-identify Any Animal in the Wild") task and introduce Wildlife-71, the first benchmark designed to test cross-species generalization to unseen animal categories. Wildlife-71 contains 71 categories (67 for training and four for testing: zebra, seal, giraffe, tiger), with a protocol requiring models to adapt to each unseen species from a single triplet of labeled support images. The authors evaluate a broad suite of baselines, including person Re-ID models, DGReID methods, and their universal model UniReID, which combines CLIP with GPT-4-guided semantic prompts. Their findings show that conventional models struggle to generalize to unseen species, positioning Wildlife-71 as a challenging open-world Re-ID benchmark. Limitations include the small test split (only four species), reliance on support triplets and class labels at test time, and categories sourced from web and non-Re-ID tracking datasets such as GOT-10k [103], which reduces ecological diversity.

Taken together, these works illustrate the rapid evolution of benchmarking practices in animal Re-ID. Early efforts such as Kuncheva et al. [41] focused on protocol design and comparative evaluation within a single, small-scale dataset, highlighting the need for more systematic and reproducible assessment. WildlifeDatasets [4] subsequently addressed this fragmentation by introducing the first unified benchmarking toolkit, enabling standardized preprocessing, consistent evaluation protocols, and cross-dataset experimentation at scale. Finally, Jiao et al. [18] extended benchmarking into the open-world setting with Wildlife-71, explicitly testing generalization to completely unseen species. Overall, the field has progressed from isolated, dataset-specific comparisons to comprehensive, unified, and cross-species benchmarks that more closely reflect the challenges of real-world animal Re-ID.

7. Challenges and open problems

As is noticeable, animal Re-ID has made significant progress in recent years, yet several fundamental challenges remain unresolved. These limitations stem from biological variability, data scarcity, domain shift, and fragmented evaluation practices and collectively shape the open problems outlined below.

Information fusion. Although fusion-based approaches have strengthened animal Re-ID through pose and behavior cues, landmark heatmaps, pelage-pattern and geometric features, temporal tracklets, spatial-frequency fusion, and emerging vision-language guidance, they also introduce challenges rooted in the heterogeneous and uneven nature of ecological data. Many auxiliary modalities are dataset or species-specific, inconsistently annotated, or noisy, forcing models to fuse incomplete or unreliable information. Cross-modality settings, such as RGB-thermal Re-ID [33], further complicate alignment because each modality exhibits different noise characteristics, spatial statistics, and availability (e.g., thermal imagery only at night). Temporal fusion is similarly complex: short-term motion cues captured in video contrast with long-term appearance drift observed in multi-year datasets such as [17], yet current methods fuse only short-term signals. Fusion also must balance species-specific identity patterns (e.g., stripes, pelage patches, facial regions) with species-agnostic representations required for cross-species generalization, as emphasized by universal and open-world models like [18].

Transferability and domain shift. Models frequently fail to generalize across habitats, camera systems, seasons, and species. Many methods are developed and evaluated on a single dataset or population, resulting in dataset-specific overfitting and limited cross-site transferability. Severe domain shifts arise from differences in lighting, background, behavior, and environmental context, and current domain adaptation techniques [26] remain insufficient for the complex, multi-domain variability encountered in ecological monitoring. Although large-scale toolkits, e.g., [4] partially alleviate this issue by unifying data access and evaluation splits, true cross-dataset and cross-species generalization remains unresolved. Benchmarks, e.g., [18] further illustrate the difficulty of adapting models to entirely unseen species.

Temporal drift and longitudinal variability. Animals undergo intrinsic appearance changes due to growth, molting, seasonal coat variation, aging, or injury. Most existing datasets provide only short-term observations, e.g., Honeybee Re-ID spans days [16], and MultiCamCows2024 covers only one week [28]. SeaTurtleID2022 offers a comparatively long 13-year collection period [17], yet species such as loggerhead turtles are extremely long-lived, meaning that even a decade of observations might reflect only a small fraction of their lifespan and long-term appearance drift. Consequently, time-aware Re-ID is rarely implemented beyond preventing temporal leakage in train/test splits, and models still struggle with long-term temporal drift in multi-year wildlife monitoring programs, including those involving turtles, seals, or livestock [14].

Data scarcity, label quality, and annotation burden. Datasets are typically small, unbalanced, and expensive to annotate, especially for rare wildlife species. Although physical tagging or expert knowledge may provide ground-truth identities, assigning these identities to thousands of images remains labor-intensive. Many datasets contain limited identities, sparse viewpoints, or noisy labels caused by occlusion, low image quality, or manual annotation difficulty. For example, Nepovinnykh et al. [14] reports that deformable bodies, wet/dry pelage variation, and partial visibility make per-image identity annotation challenging and require expert verification. While self-supervised and few-shot learning offer potential relief, they remain underdeveloped for multi-species or open-world adaptation. Developing scalable annotation pipelines, including semi-automated, self-supervised, and human-in-the-loop systems, remains an important open problem across ecological applications.

Evaluation and benchmarking gaps. Despite recent advances, evaluation practices remain inconsistent. Many datasets rely on closed-set protocols, even though ecological deployments are inherently open-set, with new individuals appearing over time [28]. Time-aware and cross-species evaluations remain rare, and cross-dataset metrics are poorly standardized. Some benchmarks include animal categories sourced from heterogeneous or non-ecological imagery. For example, Jiao et al. [18] incorporates animal instances extracted from the generic object-tracking dataset [103]. Because such images are not captured in realistic wildlife monitoring conditions, they may reduce the ecological validity of the benchmark. Furthermore, no existing benchmark jointly evaluates appearance, pose, and motion-based cues; even large toolkits, e.g., [4] cannot yet capture the full diversity of real-world monitoring pipelines.

Motion, pose, and behavior modeling. Most models operate on static images, although animals exhibit substantial pose variation and complex motion patterns. Temporal and behavioral cues in videos remain underutilized, with only a few datasets providing rich motion information (e.g., Zuerl et al. [15]). Only limited work explores temporal fusion, tracklet-based representation learning, or behavior-aware embeddings. Robust temporal modeling, especially under occlusion, crowding, or multi-animal interactions, remains a major open research direction.

Explainability, bias, and ethical deployment. Current animal Re-ID models lack interpretability, making it challenging for ecologists or conservation practitioners to assess model confidence or understand failure cases. Datasets exhibit strong sampling biases, often toward particular species, enclosures, farms, or individuals, which raises concerns about ecological validity and fairness [17]. Moreover, deploying Re-ID systems in wildlife or livestock environments requires addressing ethical considerations involving surveillance, animal disturbance, and the reliability of model predictions under uncertain field conditions.

7.1. Synthesis: Mapping challenges to methodological categories

Table 5 summarizes how different methodological categories align with the key challenges identified above. The entries provide a qualitative overview based on evidence reported in the literature, indicating whether a challenge is explicitly addressed within each methodological paradigm. A method category is marked as strong when a challenge is consistently addressed as a primary objective, moderate when it is handled in several works but not systematically, partial when it is only indirectly or conditionally mitigated, and not addressed when it falls largely outside the scope of the approach. The discussions below focus on broader challenge groupings, whereas Table 5 offers a fine-grained, per-challenge mapping. Notably, CNN and Transformer models primarily act as backbone architectures, and many other paradigms (e.g., MML, SSL/UL, and DA) are built upon these representations. Their limited coverage across challenges should therefore be interpreted in the context of this representational role rather than as a limitation in capability. Furthermore, ML is treated as a separate methodological dimension, as it defines the embedding objective independently of the underlying backbone architecture.

A clear pattern emerges for information fusion and multimodal cues: MML approaches are the only paradigm that explicitly and consistently integrates heterogeneous data sources, while most other methods rely primarily on visual representations. CNN- and Transformer-based models can incorporate additional cues, but typically in an implicit or task-specific manner rather than through systematic fusion. ML further supports this process by structuring the embedding space, yet it does not directly address cross-modal integration. HIL approaches provide a complementary mechanism by mitigating incomplete or noisy inputs through manual verification. This indicates that, despite recent progress, robust and generalizable information fusion remains largely confined to explicitly multimodal frameworks, with limited support from other methodological directions.

A similar pattern is observed for transferability and generalization. DA methods are the only approaches that explicitly address distribution shifts by aligning feature spaces across domains, whereas Transformer

Table 5
Mapping between key challenges in animal Re-ID and methodological approaches. Columns denote method categories: CNN (CNN-based feature learning), Transformer (transformer-based feature learning), ML (metric learning objectives), Video (video and temporal modeling), MML (multimodal learning), FSL (few-shot learning), SSL/UL (self-supervised and unsupervised learning), CBIR (content-based image retrieval), HIL (human-in-the-loop systems), and DA (domain adaptation). Symbols indicate the extent to which each method category addresses a given challenge: •• strong, • moderate, △ partial, and - not addressed. See text for their definition.

Challenges	CNN	Transformer	ML	Video	MML	FSL	SSL/UL	CBIR	HIL	DA
Information fusion / multimodal cues	△	△	△	△	••	△	△	△	△	△
Missing / noisy auxiliary data	△	△	△	△	•	△	•	△	•	△
Transferability across domains	△	•	△	△	•	△	•	△	△	••
Cross-species generalization	△	•	△	△	••	•	•	△	△	△
Temporal drift / longitudinal variation	–	△	△	••	△	△	•	–	△	–
Data scarcity	–	–	△	△	△	••	••	△	•	•
Annotation burden / label quality	–	–	•	△	△	•	••	△	••	•
Open-set identification	–	△	•	△	••	•	△	•	•	•
Time-aware evaluation	–	–	△	•	△	△	△	–	△	–
Benchmarking / standardization gaps	–	–	–	–	△	△	△	△	△	△
Pose / viewpoint variation	•	•	•	•	•	△	△	•	△	–
Motion / behavior modeling	△	△	△	••	•	△	△	–	△	–
Explainability / interpretability	–	–	△	△	△	△	△	△	•	–
Bias / ethical considerations	–	–	–	–	△	△	△	–	•	–

and multimodal models tend to improve generalization more implicitly through richer and more flexible representations. ML provides only partial support by learning discriminative embeddings, but does not directly account for domain discrepancies. In contrast, cross-species generalization remains largely confined to multimodal and vision-language approaches, which are specifically designed to operate beyond single-species settings. Most other methods, including conventional CNN- and Transformer-based pipelines, are still evaluated in species-specific scenarios, limiting their ability to generalize across diverse ecological conditions. Overall, this highlights that while progress has been made in improving representation robustness, explicit modeling of domain and species shifts remains limited, leaving transferability as a key open challenge.

Temporal dynamics are primarily addressed by video-based approaches, which explicitly model motion cues and temporal continuity to improve identity consistency across frames. SSL/UL methods partially exploit temporal information through constructs such as tracklets, leveraging short-term consistency without directly modeling long-term dynamics. In contrast, most CNN- and Transformer-based approaches operate on static imagery, capturing temporal variation only indirectly through learned invariances. Similarly, ML contributes by enforcing temporal robustness in the embedding space, but does not explicitly encode temporal relationships. Other paradigms, including CBIR and DA, largely overlook temporal information altogether. Overall, these observations highlight that temporal modeling remains narrowly concentrated in video-based and tracklet-driven methods, with limited integration into general-purpose Re-ID pipelines, particularly for long-term temporal variation.

Data-related challenges are primarily addressed by FSL and SSL/UL, which explicitly aim to reduce dependence on large labeled datasets. HIL systems complement these approaches by improving annotation quality through interactive verification and correction. ML provides additional support by enabling effective similarity learning under limited data regimes, although it still requires labeled examples. In contrast, most CNN- and Transformer-based methods remain largely dependent on supervised training, and DA techniques, while reducing target-domain labeling requirements, still rely on labeled source datasets. Overall, this highlights that current progress is driven mainly by alternative training paradigms, while mainstream representation learning approaches continue to depend heavily on annotated data, leaving data scarcity and annotation cost as persistent bottlenecks.

Evaluation-related aspects remain unevenly addressed across methodological paradigms. Open-set identification is supported by a subset of approaches, particularly MML, FSL, and CBIR methods, while ML further facilitates open-set recognition through embedding-based similarity matching. In contrast, most CNN- and Transformer-based pipelines are still evaluated under closed-set assumptions, limiting their applicability to real-world scenarios where new individuals frequently appear. DA improves cross-domain generalization but does not explicitly account for identity novelty. Time-aware evaluation is incorporated in video-based approaches and certain dataset designs, whereas most methods rely on static train-test splits that overlook temporal variation. More broadly, benchmarking and standardization are not directly addressed by specific methodological paradigms, resulting in fragmented and often inconsistent evaluation practices. Overall, this highlights a disconnect between methodological development and evaluation realism, indicating that more unified, open-set, and time-aware benchmarking frameworks are needed to reflect real-world deployment conditions.

Explainability and ethical considerations remain largely underexplored in current animal Re-ID research. HIL approaches provide a degree of interpretability and user control by incorporating expert validation into the identification process. In contrast, most learning-based methods, including ML frameworks and deep embedding models, operate as largely opaque systems, offering limited transparency in decision-making and little capacity to assess bias or ecological validity. This indicates that, despite advances in performance, issues related to interpretability, fairness, and responsible deployment remain insufficiently addressed.

Overall, Table 5 reveals that existing approaches tend to address individual challenges in isolation, with no single methodological paradigm providing comprehensive coverage across all dimensions. While different approaches offer complementary strengths, these are rarely integrated within unified frameworks. This fragmentation highlights the need for more holistic solutions that combine advances in representation learning, data-efficient training, multimodal fusion, and robust evaluation to address the complex and interdependent challenges of real-world animal Re-ID.

8. Future directions and recommendations

Building on the challenges outlined above, several research directions offer promising opportunities for advancing animal Re-ID toward more robust, scalable, and ecologically meaningful applications.

Multimodal and cross-species models. Future models should integrate complementary modalities, such as audio cues, movement patterns, depth information, or thermal imaging, to overcome the limitations of appearance-only representations. Cross-species and universal models,

e.g., [4,18] already indicate the feasibility of developing generalized feature extractors capable of transferring across habitats and taxa. Expanding these approaches to support a broader set of species will require larger training corpora, improved prompt-based adaptation mechanisms, and the incorporation of structured biological priors, such as taxonomic similarity, morphology, or stripe/spot developmental patterns.

Cross-modality. Cross-modality learning is an important but largely unexplored direction for animal Re-ID. Unlike person Re-ID, where visible-infrared and visible-thermal matching have been studied (e.g., [104, 105]), animal datasets rarely provide paired cross-modality observations of the same individual, limiting progress on modality-invariant embeddings and robustness to illumination or nocturnal conditions. Adapting cross-modality strategies from person Re-ID, such as modality translation, shared embedding spaces, adversarial alignment, or modality-specific attention, represents a promising avenue. Developing paired RGB-thermal, RGB-depth, or RGB-acoustic datasets for wildlife and livestock would enable such methods and support more illumination-invariant and habitat-invariant Re-ID systems.

Self-supervised, weakly supervised, and continual learning. Given the persistent scarcity of labeled data, future systems will increasingly rely on self-supervised, weakly supervised, or semi-supervised learning strategies. Techniques such as contrastive learning, cluster-based pseudo-labeling, or lifelong representation learning could enable models to learn from continuous streams of unlabeled wildlife data. This is particularly important for long-term deployments where new individuals appear over time and manual labeling cannot keep pace.

Long-term and temporal modeling. Temporal robustness remains largely unaddressed in current systems. Future work should investigate architectures capable of modeling long-range temporal dependencies, incorporating tracklets, motion cues, trajectories, seasonal variation, and age-related appearance drift. Time-aware benchmarks, e.g., [17] offer an initial foundation, though many long-lived species may require multi-decade monitoring to fully capture appearance evolution. Future models should explicitly learn long-term semantic cues (e.g., expected growth curves, molting cycles, or size changes) to build identity representations resilient to biological development and aging.

Ecologically valid benchmarks and standardized evaluation. As benchmarking evolves, there is a pressing need for datasets that capture real ecological conditions, including environmental variability, partial observability, imperfect camera placement, group movement, and cross-camera transitions. Future benchmarks should incorporate unified and reproducible evaluation protocols, including open-set, time-aware, and cross-species scenarios, and cover a wider range of habitats and taxa. Toolkits, e.g., [4] may serve as a foundation for community-driven expansion.

Synthetic data and data augmentation. Generative models can help alleviate data scarcity by producing synthetic examples reflecting rare poses, aging effects, or challenging weather and lighting conditions. Photorealistic 3D animal models, physics-based animation, and diffusion models can support controlled simulation pipelines. Intentionally exposing models to systematically perturbed or adversarial scenarios, such as severe occlusion, extreme illumination, or low-resolution inputs, to evaluate robustness and identify brittleness in Re-ID pipelines.

Explainability, reliability, and ethical deployment. As animal Re-ID transitions into field deployment, models must provide interpretable outputs and uncertainty estimates to ensure trustworthiness for conservation practitioners and ecologists. Future systems should incorporate explainable AI tools, confidence calibration, and mechanisms to detect out-of-distribution individuals. Ethical considerations, such as minimizing disturbance to wildlife, preventing misuse of monitoring data, and addressing biases in species or population representation, should be integral to model design, training, and dataset curation.

These directions collectively highlight that future progress will depend on unified benchmarks, multimodal and generalizable models, scalable self-supervised learning, and careful consideration of ecological and ethical requirements.

9. Conclusions

Animal Re-ID has undergone rapid development in the last five years, driven by advances in deep learning, improved datasets, and the emergence of unified benchmarking resources. Our survey reviewed recent studies, highlighting major methodological paradigms ranging from CNN- and transformer-based models to multimodal and self-supervised approaches. We examined datasets introduced between 2020 and 2025, categorized them by ecological context, and analyzed their species diversity, annotation quality, temporal span, and evaluation protocols. We also compared the first unified toolkits and benchmarks, such as WildlifeDatasets [4] and Wildlife-71 [18], which signal a shift toward standardized, cross-species, and more ecologically realistic evaluation.

Despite this progress, substantial challenges remain. Transferability across habitats and species is still limited; long-term temporal drift is underexplored; and most datasets remain small, short-term, or ecologically constrained. Benchmarking practices vary widely, hindering fair comparison and real-world deployment. Addressing these gaps will require unified evaluation protocols, scalable self-supervised and multimodal learning, temporal modeling capable of handling lifelong appearance changes, and careful consideration of ecological and ethical constraints.

As research continues to expand toward universal, multimodal, and deployable animal Re-ID systems, we anticipate rapid progress in building robust technologies that support large-scale wildlife monitoring, conservation decision-making, and sustainable livestock management. This survey provides a comprehensive foundation to guide future work in the field.

CRediT authorship contribution statement

Cigdem Beyan: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization; **Anil Osman Tur:** Writing – original draft, Methodology, Formal analysis; **Ehsan Karimi:** Writing – original draft, Data curation.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Cigdem Beyan, Anil Osman Tur reports financial support was provided by Research Council of Norway. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the Research Council of Norway (Computer vision to expand monitoring and accelerate assessment of coastal fish (CoastVision), project number 325862.

References

[1] D. Connor, J. Khan, E. Murray, W. Sanderson, C. Turnbull, M. Vincent, Marine Monitoring Handbook, 2001. March 2001.
[2] A.B. Irvine, R.S. Wells, M.D. Scott, An evaluation of techniques for tagging small odontocete cetaceans, Fish. Bull. 80 (1) (1982) 135–143.
[3] P.C. Ravoor, T.S.B. Sudarshan, Deep learning methods for multi-species animal re-identification and tracking: a survey, Comput. Sci. Rev. 38 (2020) 100289.
[4] V. Čermák, L. Picek, L. Adam, K. Papafitsoros, Wildlifedatasets: an open-source toolkit for animal re-identification, in: IEEE WACV, 2024.
[5] L.I. Kuncheva, F. Williams, S.L. Hennessey, J.J. Rodriguez, A benchmark database for animal re-identification and tracking, in: IEEE IPAS, 2022.
[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: IEEE CVPR, 2016, pp. 1335–1344.
[7] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: IEEE CVPR, 2017, pp. 1367–1376.
[8] Q. Leng, M. Ye, Q. Tian, A survey of open-world person re-identification, IEEE TCSVT 30 (4) (2019) 1092–1108.
[9] H. Rami, M. Ospici, S. Lathuilière, Online unsupervised domain adaptation for person re-identification, in: IEEE/CVF CVPR, 2022, pp. 3830–3839.
[10] S. Bak, P. Carr, One-shot metric learning for person re-identification, in: IEEE CVPR, 2017, pp. 2990–2999.
[11] W. Wu, D. Tao, H. Li, Z. Yang, J. Cheng, Deep features for person re-identification on metric learning, Pattern Recognit. 110 (2021) 107639.
[12] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: ACCV, Springer, 2012, pp. 31–44.
[13] S. Li, J. Li, H. Tang, R. Qian, W. Lin, ATRW: a benchmark for Amur tiger re-identification in the wild, in: ACM MM, 2020.
[14] E. Nepovinnykh, T. Eerola, V. Biard, P. Mutka, M. Niemi, M. Kunnasranta, H. Kälviäinen, SealID: saimaa ringed seal re-identification dataset, Sensors (2022) 7602.
[15] M. Zuerl, R. Dirauf, F. Koeferl, N. Steinlein, J. Sueskind, D. Zanca, I. Brehm, L.v. Fersen, B. Eskofier, PolarBearVidID: a video-based re-identification benchmark dataset for polar bears, Animals 13 (5) (2023) 801.
[16] J. Chan, H. Carrión, R. Mégret, J.L. Agosto Rivera, T. Giray, Honeybee re-identification in video: new datasets and impact of self-supervision, in: VISIGRAPP, 2022.
[17] L. Adam, V. Čermák, K. Papafitsoros, L. Picek, SeaTurtleID2022: a long-span dataset for reliable sea turtle re-identification, in: IEEE WACV, 2024.
[18] B. Jiao, L. Liu, L. Gao, R. Wu, G. Lin, P. Wang, Y. Zhang, Toward re-identifying any animal, Neurips 36 (2023) 40042–40053.
[19] A. Odo, N. McLaughlin, I. Kyriazakis, Re-identification for long-term tracking and management of health and welfare challenges in pigs, Biosyst. Eng. 251 (2025) 89–100.
[20] Z. Zheng, Y. Zhao, A. Li, Q. Yu, Wild terrestrial animal re-identification based on an improved locally aware transformer with a cross-attention mechanism, Animals 12 (24) (2022) 3503.
[21] E. Nepovinnykh, T. Eerola, H. Kalviainen, Siamese network based pelage pattern matching for ringed seal re-identification, in: IEEE WACVw, 2020.
[22] E. Nepovinnykh, I. Chelak, T. Eerola, V. Immonen, H. Kälviäinen, M. Kholiavchenko, C.V. Stewart, Species-agnostic patterned animal re-identification by aggregating deep local features, in: IJCV (2024).
[23] Z. He, J. Qian, D. Yan, C. Wang, Y. Xin, Animal re-identification algorithm for posture diversity, in: IEEE ICASSP, 2023.
[24] C. Lamping, G. Kootstra, M. Derks, Transformer-based similarity learning for re-identification of chickens, Smart Agric. Technol. 11 (2025) 100945.
[25] F.J. Williams, S.L. Hennessey, L.I. Kuncheva, Animal re-identification in video through track clustering, in: PAA (2025).
[26] F. Dubourvieux, G. Lapouge, A. Loesch, B. Luvison, R. Audigier, Cumulative unsupervised multi-domain adaptation for holstein cattle re-identification, Artif. Intell. Agric. 10 (2023) 46–60.
[27] O. Wahltinez, S.J. Wahltinez, An open-source general purpose machine learning framework for individual animal re-identification using few-shot learning, Methods Ecol. Evol. 15 (2) (2024) 373–387.
[28] P. Yu, T. Burghardt, A.W. Dowsey, N.W. Campbell, Holstein-Friesian re-identification using multiple cameras and self-supervision on a working farm, Comput. Electron. Agric. 237 (2025) 110568.
[29] J. Zhang, Q. Zhao, T. Liu, Feature-aware noise contrastive learning for unsupervised red panda re-identification, in: IJCNN, 2024.
[30] M. Perneel, I. Adriaens, J. Verwaeren, B. Aernouts, Dynamic multi-behaviour, orientation-invariant re-identification of HOlstein-Friesian cattle, Sensors 25 (10) (2025) 2971.
[31] O. Moskvyak, F. Maire, F. Dayoub, M. Baktashmotlagh, Learning landmark guided embeddings for animal re-identification, in: IEEE WACVw, 2020.
[32] L. Ashok Kumar, D. Karthika Renuka, S. Saravana Kumar, Computer vision based knowledge distillation model for animal classification and re-identification using siamese neural network, J. Intell. Fuzzy Syst. 44 (4) (2023) 5731–5743.
[33] P. Kulits, J. Wall, A. Bedetti, M. Henley, S. Beery, ElephantBook: a semi-automated human-in-the-loop system for elephant re-identification, in: ACM SIGCAS, 2021, pp. 88–98.
[34] M. Wang, M.L.V. Larsen, D. Liu, J.F.M. Winters, J.-L. Rault, T. Norton, Towards re-identification for long-term tracking of group housed pigs, Biosyst. Eng. 222 (2022) 71–81.
[35] O. Moskvyak, F. Maire, F. Dayoub, A.O. Armstrong, M. Baktashmotlagh, Robust re-identification of Manta rays from natural markings by learning pose invariant embeddings, in: DICTA, 2021.
[36] M. Zuerl, R. Dirauf, F. Koeferl, N. Steinlein, et al., PolarBearVidID: a video-based re-identification benchmark dataset for polar bears, Animals 13 (5) (2023) 801.
[37] T. Zhang, Q. Zhao, C. Da, L. Zhou, L. Li, S. Jiancuo, YakReID-103: a benchmark for yak re-Identification, in: IEEE IJCB, 2021.
[38] P. Borlinghaus, F. Tausch, L. Rettenberger, A purely visual re-ID approach for bumblebees and its application to ecological monitoring, Smart Agric. Technol. 3 (2023) 100135.
[39] X. Chen, T. Yang, K. Mai, C. Liu, J. Xiong, Y. Kuang, Y. Gao, Holstein cattle face re-identification unifying global and part feature deep network with attention mechanism, Animals 12 (8) (2022) 1047.
[40] S. Schneider, G.W. Taylor, S.C. Kremer, Similarity learning networks for animal individual re-Identification-beyond the capabilities of a human observer, in: IEEE WACVw, 2020.
[41] L.I. Kuncheva, J.L. Garrido-Labrador, I. Ramos-Pérez, S.L. Hennessey, J.J. Rodríguez, An experiment on animal re-identification from video, Ecol. Inform. 74 (2023) 101994.
[42] A. Pretto, G. Savio, F. Gottardo, F. Uccheddu, G. Concheri, A novel low-cost visual ear tag based identification system for precision beef cattle livestock farming, Inf. Process. Agric. 11 (1) (2024) 117–126.
[43] M. Wutke, D. Debiasi, S. Tomar, J. Probst, N. Kemper, K. Gevers, M.-A. Lieboldt, I. Traulsen, Multistage pig identification using a sequential ear tag detection pipeline, Sci. Rep. 15 (1) (2025) 20153.
[44] D. Babot, M. Hernández-Jover, G. Caja, C. Santamarina, J.J. Ghirardi, Comparison of visual and electronic identification devices in pigs: on-farm performances, J. Anim. Sci. 84 (9) (2006) 2575–2581.
[45] L.C.M. Omeyer, P. Casale, W.J. Fuller, B.J. Godley, K.E. Holmes, R.T.E. Snape, A.C. Broderick, The importance of passive integrated transponder (PIT) tags for measuring life-history traits of sea turtles, Biol. Conserv. 240 (2019) 108248.
[46] M.J. Witt, D. Beton, S. Davey, W.J. Fuller, B.J. Godley, M. Özkan, R. Snape, K.L. Stokes, A.C. Broderick, Phenological shift mitigates predicted impacts of climate change on sea turtle offspring, Endanger. Species Res. 56 (2025) 41–51.
[47] A. Fontaine, A. Simard, V. Simard, H.G. Broders, K.H. Elliott, Using PIT tags to infer bat reproductive status and parturition date: busy nights during lactation, J. Mammal. 105 (2024) 289–299.
[48] V.R. Goswami, M.V. Lauretta, M.D. Madhusudan, K.U. Karanth, Optimizing individual identification and survey effort for photographic capture–recapture sampling of species with temporally variable morphological traits, Anim. Conserv. 15 (2) (2012) 174–183.
[49] H. Whitehead, Computer assisted individual identification of sperm whale flukes, Rep. Int. Whaling Commission Spec. Issue 12 (1990) 71–77.
[50] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
[51] Z. Yang, D. Wu, C. Wu, Z. Lin, J. Gu, W. Wang, A pedestrian is worth one prompt: towards language guidance person re-identification, in: IEEE/CVF CVPR, 2024, pp. 17343–17353.
[52] Z. Yu, L. Li, J. Xie, C. Wang, W. Li, X. Ning, Pedestrian 3D shape understanding for person re-identification via multi-view learning, IEEE TCSVT 34 (7) (2024) 5589–5602.
[53] M. Gou, S. Karanam, W. Liu, O. Camps, R.J. Radke, Dukemtmc4reid: a large-scale multi-camera person re-identification dataset, in: IEEE CVPRw, 2017, pp. 10–19.
[54] X. Shu, X. Wang, X. Zang, S. Zhang, Y. Chen, G. Li, Q. Tian, Large-scale spatio-temporal person re-identification: algorithms and benchmark, IEEE TCSVT 32 (7) (2021) 4390–4403.
[55] S. Schneider, G.W. Taylor, S.C. Kremer, Similarity learning networks for animal individual re-identification: an ecological perspective, Mamm. Biol. 102 (3) (2022) 899–914.
[56] M. Ye, S. Chen, C. Li, W.-S. Zheng, D. Crandall, B. Du, Transformer for object re-identification: a survey, IJCV 133 (5) (2025) 2410–2440.
[57] Q. Guo, Y. Sun, L. Min, A. Putten, E.F. Knol, B. Visser, T.B. Rodenburg, J.E. Bolhuis, P. Bijma, P.H.N. de With, Video-based detection and tracking with improved re-Identification association for pigs and laying hens in farms, in: VISIGRAPP, 2022.
[58] M. Moosa, F.A. Cheikh, A. Beghdadi, M. Ullah, Self-supervised animal detection, tracking & re-identification, in: IPTA, 2024.
[59] L. Wang, R. Ding, Zhai, et al., Giant panda identification, IEEE TIP 30 (2021) 2837–2849.
[60] X. Cheng, J. Zhu, N. Zhang, Q. Wang, Q. Zhao, Detection features as attention (DEFat): a keypoint-free approach to amur tiger re-identification, in: ICIP, 2020.
[61] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, Neurips 28 (2015).
[62] Y. Wang, X. Xu, Z. Wang, R. Li, Z. Hua, H. Song, ShuffleNet-triplet: a lightweight RE-identification network for dairy cows in natural scenes, Comput. Electron. Agric. 205 (2023) 107632.
[63] X. Bai, T. Islam, M.A.H. Bin Azhar, Transformer-based models for enhanced amur tiger re-identification, in: Proc. of IEEE SAMI, 2024.
[64] Z. Li, Z. Yan, W. Tian, D. Zeng, Y. Liu, W. Li, ReDeformTR: wildlife re-identification based on light-weight deformable transformer with multi-image feature fusion, IEEE Access 12 (2024) 106321–106332.
[65] E. Nepovinnykh, T. Eerola, H. Kalviainen, I. Chelak, NORPPA: novel ringed seal re-identification by pelage pattern aggregation, in: IEEE WACVw, 2024.
[66] E. Nepovinnykh, V. Immonen, T. Eerola, C.V. Stewart, H. Kälviäinen, Re-identification of patterned animals by multi-image feature aggregation and geometric similarity, IET Comput. Vision 19 (1) (2025) e12337.
[67] P. Borlinghaus, F. Tausch, L. Rettenberger, A purely visual re-id approach for bumblebees (bombus terrestris), Smart Agric. Technol. 3 (2023) 100135.

[68] N. Dlamini, T.L. van Zyl, Comparing class-aware and pairwise loss functions for deep metric learning in wildlife re-identification, Sensors 21 (18) (2021) 6109.
[69] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: ICML, 2020, pp. 1597–1607.
[70] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Neurips 28 (2015) 91–99.
[71] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in: IEEE CVPR, 2018, pp. 6154–6162.
[72] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, in: IEEE/CVF ICCV, 2019, pp. 3702–3712.
[73] X. Zhang, X. Zhou, M. Lin, J. Sun, ShuffleNet: an extremely efficient convolutional neural network for mobile devices, in: IEEE CVPR, 2018, pp. 6848–6856.
[74] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: CVPRw, 2019.
[75] W. Zheng, F.Y. Wang, Do the best of all together: hierarchical spatial-frequency fusion transformers for animal re-identification, Inf. Fusion 113 (2025) 102612.
[76] S. Kim, D. Kim, M. Cho, S. Kwak, Proxy anchor loss for deep metric learning, in: IEEE/CVF CVPR, 2020, pp. 3238–3247.
[77] A. Radford, J.W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021, pp. 8748–8763.
[78] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: IEEE ICCV, 2017, pp. 2961–2969.
[79] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: MICCAI, Springer, 2015, pp. 234–241.
[80] D. Mishkin, F. Radenovic, J. Matas, Repeatability is not enough: learning affine regions via discriminability, in: ECCV, 2018, pp. 284–300.
[81] A. Mishchuk, D. Mishkin, F. Radenovic, J. Matas, Working hard to know your neighbor's margins: local descriptor learning loss, Neurips 30 (2017) 4826–4837.
[82] H. Weideman, C. Stewart, J. Parham, et al., Extracting identifying contours for African elephants and humpback whales using a learned appearance model, in: IEEE/CVF WACV, 2020, pp. 1276–1285.
[83] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: ECCV, 2018, pp. 801–818.
[84] Y. Sato, S. Nakajima, N. Shiraga, H. Atsumi, S. Yoshida, T. Koller, G. Gerig, R. Kikinis, Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images, Med. Image Anal. 2 (2) (1998) 143–168.
[85] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 274–282.
[86] H. Wang, J. Shen, Y. Liu, Y. Gao, E. Gavves, NFormer: robust person re-identification with neighbor transformer, in: IEEE/CVF CVPR, 2022, pp. 7297–7307.
[87] Q. Hu, H. Li, Z. Hu, F. Nie, Diverse semantic information fusion for unsupervised person re-identification, Information Fusion 107 (2024) 102319.
[88] Y. Cho, W.J. Kim, S. Hong, S.-E. Yoon, Part-based pseudo label refinement for unsupervised person re-identification, in: IEEE/CVF CVPR, 2022, pp. 7308–7318.
[89] H. Chen, B. Lagadec, F. Bremond, ICE: inter-instance contrastive encoding for unsupervised person re-identification, in: IEEE/CVF ICCV, 2021, pp. 14960–14969.
[90] Z. Dai, G. Wang, W. Yuan, S. Zhu, P. Tan, Cluster contrast for unsupervised person re-identification, in: ACCV, 2022, pp. 1142–1160.
[91] J. Wall, G. Wittemyer, B. Klinkenberg, I. Douglas-Hamilton, Novel opportunities for wildlife conservation and research with real-time monitoring, Ecol. Appl. 24 (4) (2014) 593–601.
[92] Y. Ge, D. Chen, H. Li, Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification, in: ICLR, 2019.
[93] M. Fruhner, H. Tapken, From persons to animals: transferring person re-Identification methods to a multi-Species animal domain, in: ACM International Conference Proceeding Series, 2024.
[94] S. Li, L. Fu, Y. Sun, Y. Mu, L. Chen, J. Li, H. Gong, Individual dairy cow identification based on lightweight convolutional neural network, PLoS ONE 16 (11) (2021) e0260510.
[95] L. Zheng, Y. Yang, Q. Tian, SIFT Meets CNN: a decade survey of instance retrieval, IEEE TPAMI 40 (5) (2017) 1224–1244.
[96] D. DeTone, T. Malisiewicz, A. Rabinovich, SuperPoint: self-supervised interest point detection and description, in: CVPRw, 2018, pp. 224–236.
[97] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ARCFace: additive angular margin loss for deep face recognition, in: CVPR, 2019, pp. 4690–4699.
[98] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: IEEE CVPR, 2015, pp. 815–823.
[99] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al., Hybrid task cascade for instance segmentation, in: IEEE/CVF CVPR, 2019, pp. 4974–4983.
[100] B. Cheng, A. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, Neurips 34 (2021) 17864–17875.
[101] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Learning generalisable omni-scale representations for person re-identification, IEEE TPAMI 44 (9) (2021) 5056–5069.
[102] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, et al., Dinov2: learning robust visual features without supervision, (2023). arXiv preprint arXiv:2304.07193
[103] L. Huang, X. Zhao, K. Huang, Got-10k: a large high-diversity benchmark for generic object tracking in the wild, TPAMI 43 (5) (2019) 1562–1577.
[104] N. Huang, B. Xing, Q. Zhang, J. Han, J. Huang, Co-segmentation assisted cross-modality person re-identification, Inf. Fusion 104 (2024) 102194.
[105] M. Yu, Y. Ge, Z. Chen, R. You, L. Zhu, M. Lin, Z. Xu, No escape: towards suggestive clues guidance for cross-modality person re-identification, Inf. Fusion (2025) 103185.