



Calibrated Similarity and Graph Clustering for Open-Set Animal Re-Identification

Mohamed Elbassat, Seifeldin Elkerdany, Mohamed Elbialy,
Gamal Abouelhamd, Jana Ghoneim, Assem Elkady, Mohamed
Elboraay and Nelly Semenova

EasyChair preprints are intended for rapid
dissemination of research results and are
integrated with the rest of EasyChair.

July 31, 2026

Calibrated Similarity and Graph Clustering for Open-Set Animal Re-Identification

Mohamed ElBassat^{1,3,*†}, Seifeldin Elkerdany^{1,2,*†}, Mohamed ElBialy^{1,5}, Gamal Abouelhamd^{1,2}, Jana Ghoneim^{1,4}, Assem Elkady^{1,3}, Mohamed Elboraay^{1,3} and Nelly Semenova^{6,‡}

¹Made In Alexandria Artificial Intelligence Team, Alexandria, Egypt

²Faculty of Computer Science and Engineering, Alamein International University, New Alamein City, 51718, Egypt

³Faculty of Computers and Data Science, Alexandria University, Alexandria, Egypt

⁴Faculty of Engineering, Alexandria University, Alexandria, Egypt

⁵Alexandria Higher Institute of Engineering and Technology, Alexandria, Egypt

⁶Moscow Pedagogical State University (MPGU University), 1/1 Malaya Pirogovskaya St., Moscow, 119435, Russian Federation

Abstract

AnimalCLEF26 addresses discovery-oriented animal re-identification, where systems must both attach query images to known individuals and discover unseen individuals by clustering them correctly. We present a similarity-to-clustering pipeline for this setting across Eurasian lynx, fire salamander, loggerhead sea turtle, and Texas horned lizard images. The method first isolates the target specimen using segmentation and then applies lightweight species-specific preprocessing for lynx, sea turtle, and salamander images to enhance identity-relevant visual cues, while Texas horned lizard images are used after segmentation only. Pairwise similarities are then estimated with WildFusion by calibrating and combining a MiewID global descriptor with two local matching branches, ALIKED + LightGlue and DISK + LightGlue. The resulting query-query similarities are refined and converted into identity clusters using graph-based clustering, while query-database similarities are used to attach confident samples to known identities. We evaluate training-free and fine-tuned MiewID variants, including Dynamic ArcFace and SphereFace2-Focal adaptations, and combine them in the final ensemble. Our selected ensemble substantially improves on the WildFusion baseline, achieving our best public ARI of 0.72124 and a private ARI of 0.70393, while a simpler preprocessing-before-calibration variant achieves our best private ARI of 0.71087. These results indicate that calibrated global-local fusion with species-aware preprocessing choices is effective for open-set wildlife re-identification under challenging field conditions and visual variation. The implementation code is available on GitHub.

Keywords

Animal Re-Identification, Open-Set Re-Identification, Individual Animal Identification, Computer Vision, Graph-Based Clustering, LifeCLEF 2026

1. Introduction

Animal re-identification supports wildlife monitoring by linking repeated observations to the same individual for tasks such as population estimation, movement tracking, and behavioral analysis [1, 2]. Manual identification and physical or electronic marking remain useful, but they are labor-intensive and scale poorly to the growing volume of images collected from camera traps, drones, and online sources [3, 4, 2]. This motivates the development of automated animal re-identification systems [1, 2]. In this work, we address this problem in the context of AnimalCLEF 2026, which was organized as part of the LifeCLEF 2026 lab and focuses on discovery and re-identification of individual animals in an open-set setting [5, 6].

CLEF 2026 Working Notes, September 21-24, Jena, Germany

*Corresponding authors.

†These authors contributed equally.

‡This author contributed by providing access to the previous year’s second-place solution and computational resources; they did not contribute to the development of the proposed method.

✉ cds.MohamedMohamed24510@alexu.edu.eg (M. ElBassat); seifeldean.hosny.2024@aiu.edu.eg (S. Elkerdany)

ORCID 0009-0006-3917-1319 (M. ElBassat); 0009-0005-7199-7391 (S. Elkerdany); 0009-0001-2281-7359 (M. ElBialy);

0009-0004-7223-5260 (G. Abouelhamd); 0009-0008-9943-295X (J. Ghoneim); 0009-0003-9847-9944 (A. Elkady);

0009-0002-1018-8150 (M. Elboraay); 0000-0002-0190-8382 (N. Semenova)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

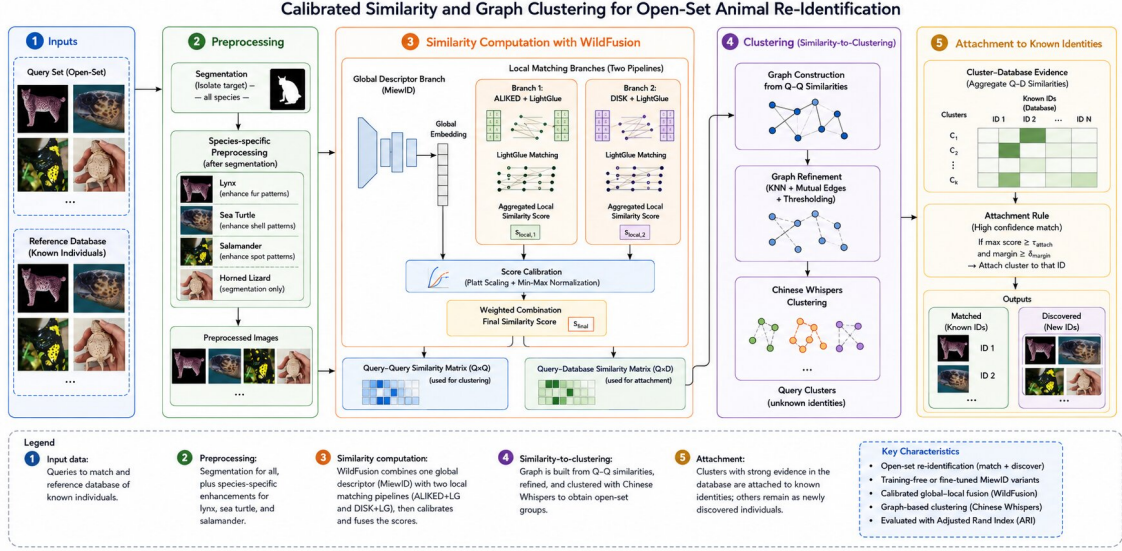


Figure 1: Overview of the proposed similarity-to-clustering pipeline.

In practice, this problem is not purely closed-set. A useful system must both match images of known individuals and discover previously unseen individuals by grouping their observations correctly [7, 2]. This open-set setting is challenging because the same animal may appear under different poses, viewpoints, lighting conditions, occlusions, ages, and backgrounds, while different individuals of the same species may differ only in subtle visual patterns [7, 2].

AnimalCLEF26 provides a benchmark for this setting across Eurasian lynx, fire salamander, loggerhead sea turtle, and Texas horned lizard images, and evaluates predictions with the Adjusted Rand Index (ARI) [7, 8, 9, 10, 11]. We address the task with a similarity-to-clustering pipeline in two variants: a training-free version and a version that adds task-adapted MiewID representations and ensembling. After segmentation, lynx, sea turtle, and salamander images receive additional species-specific preprocessing, whereas Texas horned lizard images are used after segmentation only. The resulting images are then compared with WildFusion, which combines one global descriptor with two local matching pipelines. Chinese Whispers then clusters the query images, and clusters with strong reference evidence are attached to known identities [12, 13].

2. Previous Work

Recent animal re-identification systems increasingly combine global embedding models, local visual matching, and open-set decision rules. WildFusion introduced a calibrated similarity-fusion framework that combines deep descriptor similarity with local keypoint matching scores, showing that global and local cues are complementary for individual animal identification [12]. This direction was further developed in AnimalCLEF 2025, where top-performing systems adapted global-local fusion to multi-species open-set recognition.

Pakhomov et al. proposed a hybrid global-local pipeline that first used global embeddings for candidate selection, then applied multiple local feature matchers, segmentation-aware filtering, weighted score aggregation, and novelty thresholding. Their system replaced the baseline MegaDescriptor branch with MiewID, added several local matcher combinations, and used segmentation to reduce background noise, leading to the first-place AnimalCLEF 2025 solution [14]. Semenova proposed a complementary meta-algorithm that combined WildFusion scores with an XGBoost classifier over MegaDescriptor and MiewID features, followed by species-specific Dual-Backbone ArcFace models. This showed that calibrated matching, tabular neighbor context, and metric fine-tuning can be stacked effectively for open-set animal re-identification [15].

Our method follows the same global-local philosophy, but adapts it to the AnimalCLEF26 discovery setting by combining segmentation, lightweight species-specific preprocessing, WildFusion-based calibrated similarities, fine-tuned MiewID descriptors, and graph-based clustering for open-set identity discovery.

3. Methodology

3.1. Pipeline Overview

Our method was a similarity-to-clustering pipeline designed for discovery-oriented animal re-identification. Given a set of query images for each species, the pipeline first isolated the target animal through specimen segmentation. It then applied additional species-specific preprocessing for lynx, sea turtle, and salamander images to enhance identity-relevant visual patterns, while Texas horned lizard images were used after segmentation only. Pairwise image similarities were computed using WildFusion, which combined one global descriptor branch with two local matching branches. The local branches were fixed across all official experiments and used ALIKED + LightGlue and DISK + LightGlue, while the global branch was instantiated with MiewID-based descriptors. The complete pipeline is summarized in Figure 1.

In the official pipeline, WildFusion calibration mappings were fitted after specimen segmentation and before any additional species-specific preprocessing. After the calibration mappings were fixed, final WildFusion inference was performed on the species-specific preprocessed images for lynx, sea turtle, and salamander, and on the segmentation-only images for Texas horned lizard. This produced two types of similarity matrices: a query–query matrix used for open-set identity discovery through graph-based clustering, and, when labeled database images were available, a query–database matrix used for known-identity attachment.

To improve robustness, we ran the official pipeline with three global descriptor variants: the original pretrained MiewID-msv3, a Dynamic ArcFace fine-tuned MiewID backbone, and a SphereFace2-Focal fine-tuned MiewID backbone. In all three runs, the local matching branches, WildFusion calibration and fusion stage, clustering algorithm, and known-identity attachment logic remained unchanged. The final system combined the outputs of these three descriptor-level pipeline variants.

3.2. Image Segmentation and Data Preprocessing

The first stage of the pipeline isolated the target specimen to reduce background noise before feature extraction. We used Segment Anything Model 3 (SAM 3) to generate foreground masks for the target animal, while the lynx images were used directly because the provided crops were already specimen-focused [16]. This design follows prior animal re-identification work showing that segmentation and part localization help models focus on individual-specific traits rather than environmental artifacts [10, 17].

After segmentation, we applied lightweight species-specific preprocessing only when it was expected to improve identity-relevant visual cues [18, 19]. For lynx, we enhanced fur texture, markings, and edges using sharpening, HSV-based CLAHE, and bilateral filtering. For sea turtles, we corrected underwater color degradation by boosting the red channel and applying gamma correction to improve shell, head, and scale visibility. For salamanders, we strengthened yellow–black pattern boundaries using Sobel edge enhancement followed by blending and gamma correction. We also explicitly controlled the composited background color after segmentation, since dark salamander regions against a dark background could reduce boundary contrast and make both global descriptors and local keypoint matching less reliable; lighter backgrounds preserved specimen contours more clearly. For Texas horned lizards, we applied no additional preprocessing beyond segmentation, preserving the original ventral spot patterns used for identification [9].

Representative segmentation and preprocessing outputs are shown in Figure 2.

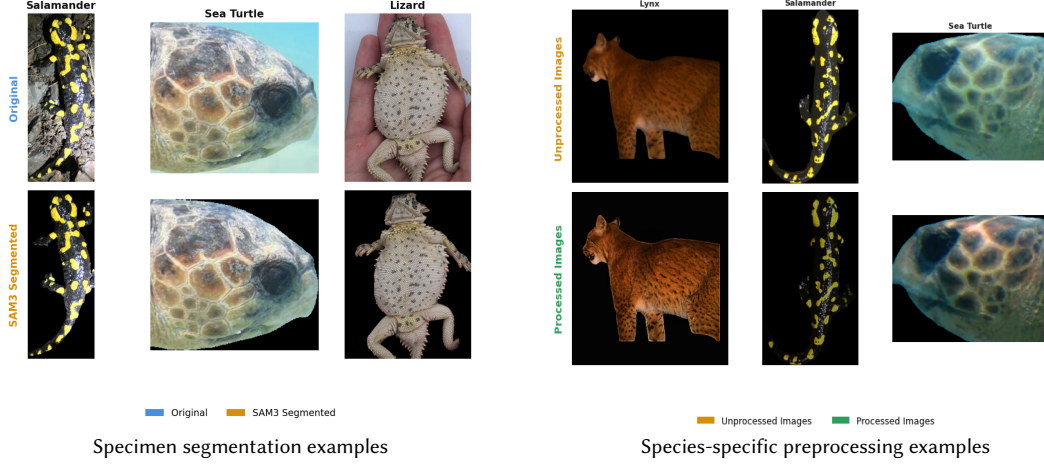


Figure 2: Representative specimen segmentation and species-specific preprocessing outputs.

3.3. Feature Extraction and Similarity Components

Our pipeline computed image similarity using two complementary sources of visual evidence: a global descriptor branch and two local matching branches. The global descriptor summarized the overall appearance of the segmented animal specimen, while the local matching branches compared fine-grained visual structures such as fur markings, salamander skin patterns, turtle shell texture, scales, spots, and scars. This combination was important in open-set animal re-identification because the same individual could appear under different poses, viewpoints, lighting conditions, and levels of occlusion, while different individuals of the same species could differ only in subtle local patterns.

The local matching branches were fixed across all evaluated pipeline configurations. We used ALIKED + LightGlue and DISK + LightGlue as two complementary local feature pipelines. These branches were used without task-specific fine-tuning and provided local similarity scores that were later calibrated and fused with the global descriptor similarity through WildFusion [12]. In contrast, the global descriptor branch was varied across the main pipeline instances, using either the original pretrained *MiewID*-msv3 descriptor or a fine-tuned *MiewID* backbone.

3.3.1. Global Descriptor: *MiewID*

The global branch was based on *MiewID*-msv3, which represents each image with a single embedding and uses pairwise cosine similarity as the global score before WildFusion calibration and fusion [20].

MiewID was designed for multi-species animal re-identification, making it suitable for the Animal-CLEF26 setting, where the model had to generalize across visually diverse species and support both known-individual matching and unseen-individual discovery [20]. In the training-free pipeline, we used the original pretrained *MiewID*-msv3 model directly. In the ensemble setting, the same pipeline was also evaluated with two task-adapted *MiewID* variants: one fine-tuned with Dynamic ArcFace and one fine-tuned with SphereFace2-Focal loss. In all cases, the global descriptor was the only component changed; the local matching branches, WildFusion calibration and fusion, clustering, and known-identity attachment stages remained unchanged.

We also considered MegaDescriptor, the foundation model released with the WildlifeDatasets ecosystem [1]. However, based on qualitative Score-CAM inspection after segmentation, *MiewID* appeared more suitable in our setting because its heatmaps more often focused on identity-relevant animal regions. In contrast, MegaDescriptor appeared to rely more on broader image-level cues. As shown in Figure 3, each row corresponds to one AnimalCLEF26 species and compares the original image with the *MiewID* and MegaDescriptor heatmaps and overlays [21, 20, 1, 15].

To better visualize identity separation, Panel (a) of Figure 5 shows a UMAP projection of the embedding space produced by the original *MiewID*-msv3 model without fine-tuning, generated with HyperView

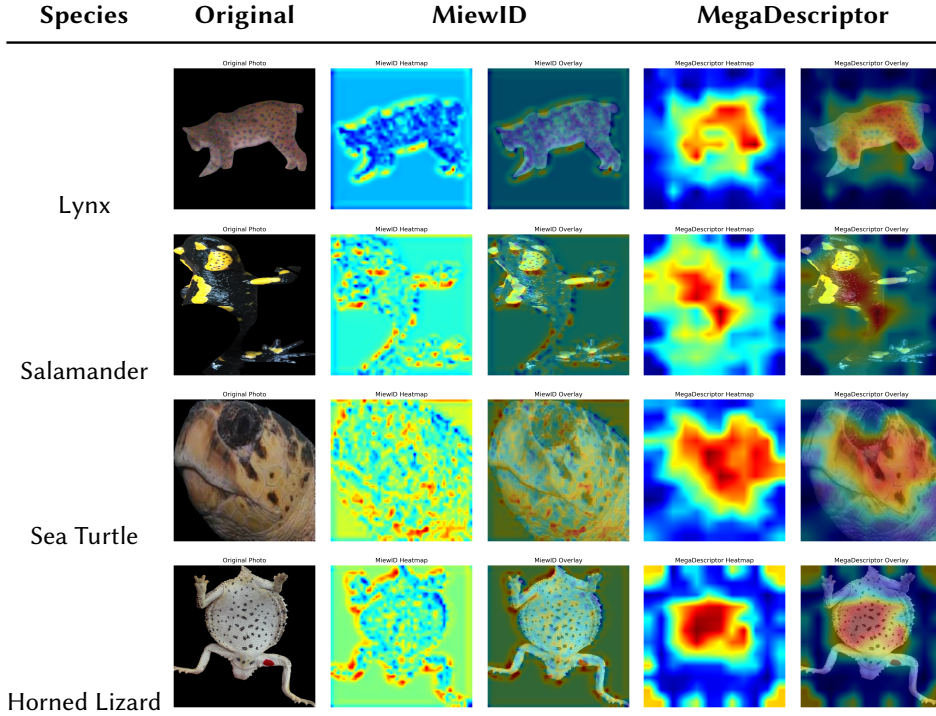


Figure 3: Score-CAM comparison of MiewID and MegaDescriptor across four species.

software [22, 23].

3.3.2. Local Matching Branches

Local feature matching complemented the MiewID global descriptor by comparing identity-specific anatomical patterns at the correspondence level. Each branch processed image pairs resized to 512×512 and produced a similarity signal from the number and confidence of matched local features.

ALIKED + LightGlue. The first branch used pretrained ALIKED features with LightGlue matching, without further fine-tuning. ALIKED provides efficient keypoint detection and deformable descriptor construction, while LightGlue establishes fast adaptive correspondences between extracted features [24, 25].

DISK + LightGlue. The second branch used pretrained DISK features with the same LightGlue matcher. DISK is optimized for reliable keypoint detection and description under geometric and photometric variation [26]. We retained only matches with confidence at least 0.9, reducing ambiguous correspondences and providing a complementary local descriptor family.

Qualitative Analysis of Local Matchers. Figure 4 compares ALIKED + LightGlue and DISK + LightGlue across lynx, salamander, and sea turtle examples. ALIKED tends to produce structured keypoints around salient contours and repeated markings, whereas DISK often yields denser correspondences over textured regions. This complementary behavior motivated using both branches in WildFusion rather than relying on a single local descriptor.

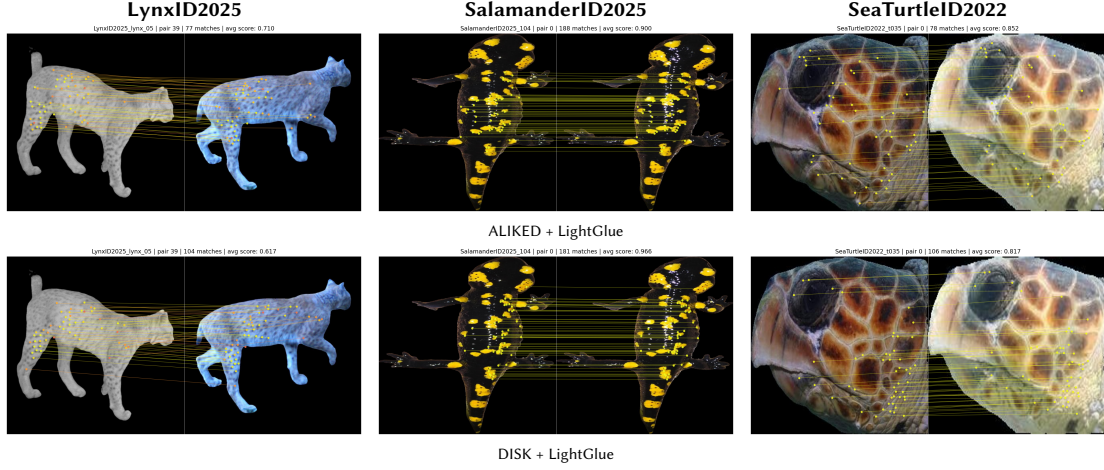


Figure 4: Local matcher comparison across three datasets

3.4. WildFusion Similarity Computation and Calibration

Using only global descriptors could miss fine local correspondences, while relying only on local matching was often computationally expensive and less stable at scale. To address this trade-off, we built on WildFusion, a calibrated similarity-fusion framework for individual animal identification [12]. WildFusion combines global descriptor similarity with local matching evidence into a unified score, allowing the system to benefit from both holistic appearance cues and precise local correspondences.

In our setup, WildFusion used three similarity branches: the MiewID global descriptor, ALIKED + LightGlue, and DISK + LightGlue. We used *IsotonicCalibration* for score calibration, with the WildFusion parameter B set to 256. The calibrated branch scores were then fused using the standard WildFusion averaging rule.

This design was well suited to our discovery-oriented setting. Global embeddings provided robust coarse identity structure, whereas local matches helped disambiguate visually similar individuals. By integrating both signals, WildFusion offered a stronger basis for downstream clustering and identity assignment than either component alone.

3.5. Clustering and Known-Identity Matching

After computing calibrated WildFusion similarities, our final stage converted pairwise image similarities into submission labels. The procedure was applied independently for each species, allowing the similarity refinement, graph construction, clustering, and known-identity attachment parameters to be selected separately for each dataset. For a given species, let $\mathcal{Q} = \{q_i\}_{i=1}^N$ denote the query images and let $\mathcal{G} = \{g_j\}_{j=1}^M$ denote the corresponding database images, when a database was available.

We computed two similarity matrices using WildFusion. The query-query matrix

$$S^{QQ} \in [0, 1]^{N \times N}$$

was used for clustering query images into individual animals. When database images were available, we also computed the query-database matrix

$$S^{QG} \in [0, 1]^{N \times M},$$

which was used only for known-identity attachment. Both matrices were clipped to the range $[0, 1]$ for numerical stability. Importantly, these two matrices were used for different purposes: S^{QQ} defined the graph structure for query clustering, whereas S^{QG} provided evidence for assigning clusters to known database identities.

3.5.1. K-Reciprocal Re-Ranking of Query–Query Similarities

When enabled for a given species, we applied k-reciprocal re-ranking to the query–query similarity matrix before graph clustering [27]. Re-ranking was restricted to S^{QQ} ; the query–database matrix S^{QG} used for known-identity attachment was left unchanged. This separation prevented the neighborhood refinement used for clustering from altering the calibrated query–database scores used to decide whether a query belonged to a known identity.

Let $S^{\text{orig}} = S^{QQ}$ denote the original WildFusion query–query similarity matrix, and define the corresponding distance matrix as

$$d_{ij}^{\text{orig}} = 1 - S_{ij}^{\text{orig}}.$$

For each query image q_i , we retrieved the k_1 nearest neighbors according to d^{orig} . A neighbor q_j was retained in the reciprocal set of q_i only if the relation was mutual: q_j had to be among the top- k_1 neighbors of q_i , and q_i also had to be among the top- k_1 neighbors of q_j . This gave the initial k-reciprocal set

$$\mathcal{R}_{k_1}(i) = \{j \mid j \in \mathcal{N}_{k_1}(i), i \in \mathcal{N}_{k_1}(j)\}.$$

The reciprocal set was then expanded to include reliable neighbors that might be missed by the strict mutual-neighbor rule. For each $j \in \mathcal{R}_{k_1}(i)$, we computed a smaller reciprocal set $\mathcal{R}_{\lceil k_1/2 \rceil}(j)$. This set was absorbed into the neighborhood of q_i when more than two-thirds of its members were already present in $\mathcal{R}_{k_1}(i)$:

$$\frac{|\mathcal{R}_{\lceil k_1/2 \rceil}(j) \cap \mathcal{R}_{k_1}(i)|}{|\mathcal{R}_{\lceil k_1/2 \rceil}(j)|} > \frac{2}{3}.$$

We denoted the expanded reciprocal set by $\mathcal{R}_{k_1}^*(i)$.

Each expanded set was encoded as a sparse affinity vector v_i , where closer neighbors received larger weights:

$$v_i(j) = \frac{\exp(-d_{ij}^{\text{orig}})}{\sum_{t \in \mathcal{R}_{k_1}^*(i)} \exp(-d_{it}^{\text{orig}})}, \quad j \in \mathcal{R}_{k_1}^*(i),$$

and $v_i(j) = 0$ otherwise. When $k_2 > 1$, this vector was smoothed by averaging it over the k_2 nearest neighbors in the original ranking. This reduced the effect of unstable individual neighbor assignments and produced a more robust local neighborhood representation.

A continuous Jaccard distance was then computed between the re-ranked neighborhood vectors:

$$d_{ij}^J = 1 - \frac{\sum_l \min(v_i(l), v_j(l))}{\sum_l \max(v_i(l), v_j(l))}.$$

The final distance combined this neighborhood-based distance with the original WildFusion distance:

$$d_{ij}^{\text{final}} = (1 - \lambda)d_{ij}^J + \lambda d_{ij}^{\text{orig}},$$

where $\lambda \in [0, 1]$ controlled the contribution of the original pairwise distance. The re-ranked similarity matrix was then obtained as

$$S_{ij}^{\text{rr}} = 1 - d_{ij}^{\text{final}}.$$

The matrix passed to clustering was therefore

$$S^{\text{CW}} = \begin{cases} S^{\text{rr}}, & \text{if re-ranking is enabled for the species,} \\ S^{QQ}, & \text{otherwise.} \end{cases}$$

The re-ranking parameters (k_1, k_2, λ) were selected as part of the species-specific clustering configuration, together with the graph-construction and label-propagation hyperparameters.

3.5.2. Graph Construction and Chinese Whispers Clustering

The selected query–query similarity matrix S^{CW} was converted into a sparse weighted graph. Each query image was represented as a node, and edge weights were derived from the pairwise similarities in S^{CW} . This representation was well suited to the open-set nature of the task because the number of identities in the query set was unknown in advance. Instead of requiring a predefined number of clusters, Chinese Whispers inferred clusters from dense connected regions in the graph [13].

Graph construction was controlled by a species-specific similarity threshold τ_{CW} , nearest-neighbor limit k_{NN} , and k-core degree k_{core} . Edges with similarity below τ_{CW} were removed. The graph was then sparsified using nearest-neighbor constraints so that clustering depended mainly on the strongest local relationships rather than weak long-range similarities. A k-core filtering step was used to preserve a reliable graph backbone by removing weakly connected nodes whose degree was below k_{core} . If this filtering removed all nodes, the algorithm fell back to the unfiltered graph to avoid degeneracy.

Chinese Whispers was then applied to the resulting weighted graph. Each node was initialized with a unique label. During each iteration, nodes were visited in randomized order, and each node adopted the label with the largest total incoming edge weight among its neighbors. The process was repeated until convergence or until the maximum number of iterations was reached. The final labels were remapped to contiguous cluster indices and used as the predicted query clusters.

In our implementation, Chinese Whispers received four species-specific hyperparameters:

$$(\tau_{CW}, k_{NN}, k_{core}, T),$$

where τ_{CW} was the edge threshold, k_{NN} was the top-neighbor limit, k_{core} was the minimum backbone degree, and T was the maximum number of label-propagation iterations.

3.5.3. Known-Identity Scoring

For species with a labeled database, we used the query–database matrix S^{QG} to estimate whether each query image matched a known identity. Let $\mathcal{G}_c$ denote the set of database images belonging to known identity c . For each query image q_i and each known identity c , we converted image-level similarities into an identity-level score by averaging the top- r similarities between q_i and the database images of identity c :

$$A_{i,c} = \frac{1}{\min(r, |\mathcal{G}_c|)} \sum_{s \in \text{Top}_r(\{S_{ij}^{QG} | g_j \in \mathcal{G}_c\})} s.$$

The top- r aggregation made the identity score less sensitive to low-quality or viewpoint-mismatched database images, while still allowing multiple strong matches to support a known-identity assignment.

For each query image, we then identified the best and second-best known identities:

$$\begin{aligned} c_i^{(1)} &= \arg \max_c A_{i,c}, \\ b_i &= A_{i,c_i^{(1)}}, \\ b_i^{(2)} &= \max_{c \neq c_i^{(1)}} A_{i,c}. \end{aligned}$$

The confidence margin was defined as

$$m_i = b_i - b_i^{(2)}.$$

A query image was considered a confident known-identity match only if

$$b_i \geq \tau_{\text{attach}} \quad \text{and} \quad m_i \geq \tau_{\text{margin}}.$$

The first condition ensured that the best known identity had sufficiently high similarity, while the second condition ensured that the assignment was not ambiguous with respect to the next-best identity.

3.5.4. Cluster-Level Known-Identity Attachment

Known-identity decisions were applied at the cluster level rather than directly at the image level. After Chinese Whispers produced query clusters, each confident query image cast a weighted vote for its best known identity. For a cluster z and known identity c , the accumulated vote was

$$V_{z,c} = \sum_{i: z_i=z} \mathbb{I} \left[b_i \geq \tau_{\text{attach}} \wedge m_i \geq \tau_{\text{margin}} \wedge c_i^{(1)} = c \right] b_i,$$

where z_i was the cluster assigned to query image q_i . If at least one known identity received votes in cluster z , the cluster was assigned to the identity with the highest accumulated vote:

$$\hat{c}_z = \arg \max_c V_{z,c}.$$

All images in that cluster were then labeled as the selected known identity.

If a cluster received no confident known-identity votes, it was treated as a newly discovered individual. The final label for query image q_i was therefore

$$\hat{y}_i = \begin{cases} \text{cluster_dataset_}\{\hat{c}_{z_i}\}, & \text{if cluster } z_i \text{ was attached to a known identity,} \\ \text{cluster_dataset_new_}\{z_i\}, & \text{otherwise.} \end{cases}$$

For species without a labeled database, the query–database matrix was unavailable and all clusters were treated as newly discovered identities.

This cluster-level voting strategy made the final prediction more stable than direct image-level attachment. A single uncertain query image could not attach to a known identity unless it passed both the score and margin thresholds, and when multiple confident images occurred in the same cluster, their evidence was accumulated before assigning the final known label.

3.6. MiewID Fine-Tuning

Although the out-of-the-box pretrained MiewID model already produced strong results, we further adapted it to the AnimalCLEF26 setting through species-aware fine-tuning. In this work, we explored two fine-tuning strategies. The first strategy fine-tuned MiewID with a dynamic ArcFace objective to improve embedding discrimination, while the second strategy explored a fine-tuned variant based on SphereFace2 combined with focal loss.

3.6.1. Dynamic ArcFace Fine-Tuning

We fine-tuned the pretrained MiewID-msv3 model to improve the discriminative quality of the global descriptor used in WildFusion. The goal of this stage was not closed-set identity classification, but learning a more compact and better-separated embedding space that remained useful for open-set identity discovery [7, 1].

To approximate the challenge setting, we constructed an open-set species-wise split in which a subset of validation identities was excluded from fine-tuning. The labeled data used for this stage contained 13,074 images from 1,102 identities across lynx, salamander, and sea turtle datasets, while Texas horned lizard contributed no labeled identities because it was available only in the challenge test set. Table 1 summarizes the resulting train/validation partition.

We initialized the model from the pretrained MiewID-msv3 checkpoint and fine-tuned it with a Dynamic ArcFace sub-center head using a two-phase schedule [28, 29, 30]. In the first phase, the backbone was frozen and only the classification head was trained as a short warm-up step. In the second phase, the full network was unfrozen and optimized end-to-end. In our configuration, the head used $k = 3$ sub-centers per class, embedding dimension 2048, scale $s = 64.0$, angular margin $m = 0.6$, and threshold $\theta_0 = 0.785$. After fine-tuning, the classification head was discarded and only the normalized backbone embeddings were retained for similarity computation in the WildFusion pipeline.

Table 1

Species-wise open-set split used for Dynamic ArcFace fine-tuning. The Train set (FT) contained individuals used for fine-tuning, while the Validation set was partitioned into “Known” and “Unknown” individuals to evaluate discovery performance.

Species	Original Data		Train Set (FT)		Validation Set			
	imgs.	IDs	imgs.	IDs	Known IDs	Known imgs.	Unk. IDs	Unk. imgs.
LynxID2025	2957	77	2519	75	69	339	2	99
SalamanderID2025	1388	587	1087	558	208	233	29	68
SeaTurtleID2022	8729	438	6265	419	263	2136	19	328
Total	13074	1102	9871	1052	540	2708	50	495

Qualitatively, the resulting embeddings showed tighter within-identity clusters and clearer separation between neighboring identities than the pretrained model, which was consistent with the objective of improving downstream clustering and known-identity attachment in the open-set setting. This effect is shown in panel (b) of Figure 5, using a UMAP projection generated with HyperView software [22, 23].

3.6.2. SphereFace2-Focal Fine-Tuning

We initialized the model from the pretrained `MiewID-msv3` checkpoint [20] and adapted it as a feature extractor with an identity-discriminative training head. All input images were resized to 440×440 pixels and normalized using ImageNet statistics. For the salamander subset, we used segmented images composited on a white background, motivated by the observation that background clutter can bias animal re-identification models away from individual-specific visual traits [7].

To evaluate the fine-tuned MiewID model under an open-set setting, we constructed a species-wise split in which known validation identities retained training images, while unknown validation identities were completely removed from the fine-tuning set. This allowed the model to be evaluated on both known-individual retrieval and unseen-individual rejection, while preserving most of the available labeled data for fine-tuning [7]. The resulting partition is reported in Table 2.

Table 2

Species-wise open-set split used for SphereFace2-Focal fine-tuning. The Train set (FT) contained individuals used for fine-tuning, while the Validation set was partitioned into “Known” and “Unknown” individuals to evaluate discovery performance.

Species	Original Data		Train Set (FT)		Validation Set			
	imgs.	IDs	imgs.	IDs	Known IDs	Known imgs.	Unk. IDs	Unk. imgs.
LynxID2025	2957	77	2816	55	19	38	22	103
SalamanderID2025	1388	587	1231	538	59	59	49	98
SeaTurtleID2022	8729	438	8467	388	50	100	50	162
Total	13074	1102	12514	981	128	197	121	363

SphereFace2-Focal Objective The fine-tuning head was based on the SphereFace2 binary-classification formulation [31]. Instead of training the model as a standard closed-set classifier, SphereFace2 treats each identity as an independent one-vs-all decision. For each training image, the correct identity is treated as the positive class, while all other identities are treated as negative classes. This design is better aligned with open-set re-identification, because the final task depends on pairwise similarity and unknown-individual discovery rather than choosing only from a fixed list of training identities.

Following SphereFace2, the model learned normalized embeddings and normalized identity prototypes on the hypersphere. The head used a similarity adjustment step to make positive and negative similarities more separable, together with a two-sided margin that made same-identity pairs satisfy a stricter similarity requirement while pushing different-identity pairs farther from the decision boundary. A

shared bias term was used as a universal threshold across identities, making the binary decisions more flexible and stable during training.

Because each image has one positive identity but many negative identities, we used positive-negative reweighting to avoid the loss being dominated by negative classes. We also combined the SphereFace2 objective with focal weighting [32]. This reduced the influence of easy decisions and encouraged the model to focus more on difficult cases, such as visually similar individuals from the same species.

In our experiments, we used a SphereFace2-Focal head with margin 0.3, scale 55, similarity-adjustment strength 3.0, positive-negative balance 0.99, global loss weight 20, and focal focusing parameter 1.0. After fine-tuning, the classification head was discarded and only the normalized backbone embeddings were used. These embeddings were then integrated into the WildFusion-based similarity pipeline as the global descriptor, where improved embedding compactness could benefit both known-identity attachment and unknown-individual clustering. Panel (c) of Figure 5 shows a similar tightening of the embedding space after fine-tuning, using a UMAP projection generated with HyperView software [22, 23].

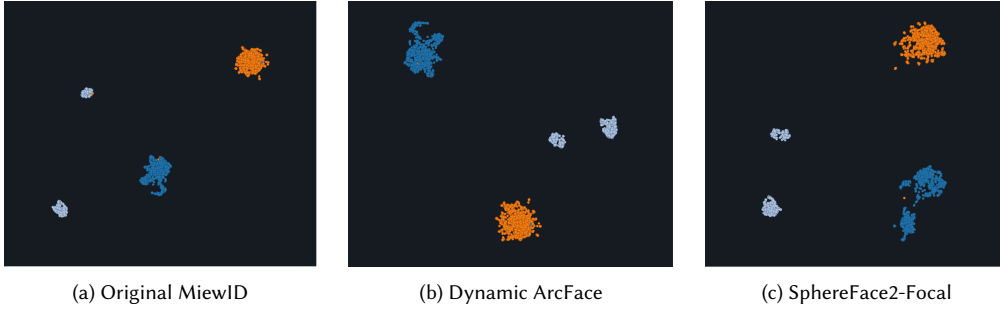


Figure 5: Embedding-space visualizations for the Original MiewID model, Dynamic ArcFace fine-tuning, and SphereFace2-Focal fine-tuning. Colors denote datasets: blue is LynxID2025, orange is SeaTurtleID2022, and gray is SalamanderID2025.

3.7. Hyperparameter Optimization with Optuna

We used Optuna [33] to select the species-specific decision parameters of the similarity-to-clustering pipeline. Validation splits followed the open-set competition setting: identities were divided into known and unknown sets, and known identities were further split into reference and query images (70% / 30%).

To keep the search efficient, pairwise similarity matrices were precomputed and reused across trials. Each trial selected the known-identity scoring parameters r , τ_{attach} , and τ_{margin} ; the re-ranking setting and parameters (k_1, k_2, λ) ; and the Chinese Whispers graph parameters $(\tau_{\text{CW}}, k_{\text{NN}}, k_{\text{core}}, T)$.

Each trial then followed the same inference pipeline described above. Query-database similarities were aggregated with top- r pooling for known-identity scoring. The selected query-query matrix S^{CW} was built from either the re-ranked similarity matrix or the original query-query matrix, depending on the species-specific configuration, and Chinese Whispers was applied to the resulting sparse weighted graph.

After clustering, known-identity attachment was performed at the cluster level. Confident query images cast weighted votes when $b_i \geq \tau_{\text{attach}}$ and $m_i \geq \tau_{\text{margin}}$; clusters without confident votes were treated as newly discovered individuals. The validation objective was

$$0.6 \times \text{ARI} + 0.4 \times \text{V-measure},$$

Here, ARI was the dominant term because it matched the competition metric. We included V-measure only as a secondary regularizing term to discourage degenerate clusterings, such as merging multiple identities into one large cluster, while all official leaderboard scores were computed using ARI alone [34]. The best configuration was used in the final submission pipeline.

3.8. Ensembling

The pretrained MiewID, Dynamic ArcFace, and SphereFace2-Focal variants often produced complementary errors, so the final system combines their predicted labels by majority voting. For salamanders, most variants used segmented images composited on a white background, whereas the Dynamic ArcFace variant used a black background. The ensemble therefore benefits from both descriptor diversity and modest preprocessing variation.

4. Evaluation Metric

AnimalCLEF26 requires methods to both assign query images to known individuals and discover previously unseen identities. Consequently, submissions are evaluated using the **Adjusted Rand Index (ARI)**, which measures the similarity between the predicted clustering and the ground-truth clustering while correcting for agreement expected by chance [7, 11].

Unlike the Rand Index (RI), which measures the proportion of image pairs on which the predicted and ground-truth clusterings agree, ARI normalizes this agreement so that random clusterings receive scores close to zero, while identical clusterings receive a score of one [11].

The ARI is defined as

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}, \quad (1)$$

where $E[RI]$ denotes the expected Rand Index under a random assignment model [11].

Because ARI penalizes both over-segmentation (splitting one individual into multiple clusters) and under-segmentation (merging different individuals into the same cluster), it is well suited to the open-set animal re-identification task addressed in AnimalCLEF26.

5. Results

The main pipeline segments each specimen, applies additional species-specific preprocessing for lynx, sea turtle, and salamander, while using segmentation-only images for Texas horned lizard, and computes calibrated WildFusion similarities from a MiewID global descriptor plus ALIKED + LightGlue and DISK + LightGlue local branches. These similarities drive Chinese Whispers clustering, cluster-level known-identity attachment when labeled references exist, and the final ensemble across three MiewID variants. Table 3 reports the official AnimalCLEF26 public and private leaderboard results using Adjusted Rand Index (ARI). We compare the competition baseline, which uses agglomerative clustering, individual pipeline variants, and our final ensemble submission.

Table 3

Comparison of ARI scores for the baseline, cumulative ablations, individual pipeline variants, and the final ensemble on the AnimalCLEF26 public and private leaderboards.

Method	Public ARI	Private ARI
Competition baseline (WildFusion)	0.20342	0.21221
Baseline using only MiewID as global feature extractor	0.60378	0.54901
+ SAM segmentation	0.63996	0.57880
+ Species-specific preprocessing	0.64784	0.57611
Main pipeline with Chinese Whispers clustering (out-of-the-box)	0.71919	0.66695
Dynamic ArcFace fine-tuned MiewID pipeline	0.70508	0.63749
SphereFace2-Focal fine-tuned MiewID pipeline	0.70584	0.65736
Final ensemble	0.72124	0.70393

Table 4 reports the species-specific hyperparameters selected by Optuna for each MiewID-based pipeline variant. The first three species use full query–query clustering followed by cluster-level

Table 4

Species-specific Optuna-selected hyperparameters for the three MiewID-based pipeline variants. The parameters control known-identity attachment, Chinese Whispers graph clustering, and k-reciprocal re-ranking.

Pipeline variant	Dataset	top- r	τ_{attach}	τ_{margin}	τ_{CW}	k_{NN}	T	k_{core}	Re-rank	(k_1, k_2, λ)
Original MiewID	LynxID2025	1	0.49	0.26	0.27	47	11	1	Yes	(11, 3, 0.835)
	SalamanderID2025	2	0.25	0.19	0.50	25	12	1	Yes	(5, 4, 0.758)
	SeaTurtleID2022	4	0.79	0.19	0.25	47	15	1	Yes	(39, 3, 0.449)
	TexasHornedLizards	—	—	—	0.49	15	20	1	No	—
SphereFace2-Focal MiewID	LynxID2025	1	0.28	0.42	0.29	42	10	1	Yes	(10, 1, 0.808)
	SalamanderID2025	2	0.35	0.23	0.40	60	20	5	Yes	(7, 4, 0.718)
	SeaTurtleID2022	5	0.82	0.45	0.23	47	25	1	Yes	(20, 5, 0.852)
	TexasHornedLizards	—	—	—	0.49	15	20	1	No	—
Dynamic ArcFace MiewID	LynxID2025	5	0.29	0.17	0.24	57	17	3	Yes	(5, 4, 0.810)
	SalamanderID2025	3	0.19	0.24	0.43	55	29	1	Yes	(5, 9, 0.662)
	SeaTurtleID2022	5	0.51	0.80	0.21	49	7	1	Yes	(9, 1, 0.824)
	TexasHornedLizards	—	—	—	0.49	15	20	1	No	—

known-identity attachment, while Texas horned lizard has no labeled database and therefore uses only the clustering-related parameters.

5.1. Additional Submitted Configuration: Preprocessing Before Calibration

We also evaluated an additional competition-time configuration to study the effect of applying the additional species-specific preprocessing steps before WildFusion calibration. Unlike the final ensemble, this configuration used only the original pretrained MiewID-msv3 model as the global descriptor, while keeping the same ALIKED + LightGlue and DISK + LightGlue local branches.

The main difference was the order of preprocessing and calibration. In this configuration, calibration was fitted on the same preprocessed image distribution used during final inference:

segmentation $\rightarrow$ species-specific preprocessing $\rightarrow$ WildFusion calibration $\rightarrow$ WildFusion inference.

Table 5

Comparison between the final selected ensemble submission and the additional submitted configuration on the official AnimalCLEF26 leaderboards.

Configuration	Public ARI	Private ARI
Final ensemble submission	0.72124	0.70393
Preprocessing before calibration	0.71439	0.71087

As shown in Table 5, the final ensemble achieved a higher public score, while the preprocessing-before-calibration configuration achieved a higher private score. This suggests that fitting WildFusion calibration on the final image distribution, including species-specific preprocessing where used, may improve hidden-split generalization. It is also computationally simpler because it avoids fine-tuned descriptor variants and multi-model ensembling.

6. Discussion

The results show that most of the performance gain comes from the overall similarity-to-clustering design rather than from fine-tuning alone. Even the training-free pipeline strongly outperforms the competition baseline, indicating that specimen segmentation, targeted species-specific preprocessing where beneficial, and calibrated fusion of global and local cues already provide a strong solution for discovery-oriented re-identification.

The fine-tuned MiewID variants remain useful because they add complementary information, even though they do not outperform the training-free model individually. This is reflected in the ensemble,

which yielded our best selected submission on the public leaderboard, while the preprocessing-before-calibration configuration achieved the strongest private leaderboard score. Together, these findings suggest that representation diversity improves robustness and that calibration strategy remains a key factor for generalization.

7. Limitations

Despite its strong performance, the proposed system has three main limitations. First, our pipeline uses Chinese Whispers to convert the query–query similarity graph into identity clusters. Although efficient and well suited to open-set discovery, Chinese Whispers updates node labels in randomized order, so repeated runs can produce slightly different cluster assignments, especially for weakly connected or visually ambiguous samples. This limits reproducibility relative to deterministic graph-clustering methods [13].

Second, the complete system is computationally expensive. Local matching and pairwise similarity computation require comparing many image pairs, and the final ensemble repeats the pipeline for three MiewID-based descriptor variants, increasing both inference time and memory usage.

Third, the calibration stage was fitted using only Lynx images rather than a more diverse multi-species calibration set. This may limit how well the calibrated similarities transfer across species with different visual statistics and imaging conditions.

These limitations suggest several directions for future work, including more deterministic graph clustering, lighter single-model alternatives that reduce the cost of pairwise matching and ensembling, and broader calibration strategies that better cover species diversity. The latter is especially promising because a simpler off-the-shelf configuration remained competitive on the private leaderboard.

8. Conclusion

We presented a discovery-oriented animal re-identification pipeline for AnimalCLEF26 that combines specimen segmentation, targeted species-specific preprocessing where beneficial, WildFusion-based similarity estimation, graph clustering, and known-identity attachment. The training-free pipeline already performs strongly, while combining training-free and fine-tuned MiewID variants yields our best public leaderboard submission and highlights the value of complementary global-local cues; however, a simpler preprocessing-before-calibration configuration achieved our strongest private leaderboard score. Future work will focus on more deterministic clustering, more efficient similarity computation, and improved calibration for stronger generalization.

Acknowledgments

We thank the LifeCLEF organizers for making the AnimalCLEF 2026 datasets available and for organizing the competition. We are especially grateful to Eslam Yacoub, PhD, Assistant Professor of English Curricula and Instruction at Alamein International University, for carefully reviewing the manuscript and for his guidance throughout the preparation of this working note. We also thank Antonio Rueda-Toicen from Hasso Plattner Institute for his support and thoughtful review of the paper, and Bahey Tharwat, PhD student at the University of Freiburg, for reviewing the paper and helping with figure preparation.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-5.5 and Prism GPT-5.2 for grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and assumed full responsibility for the final version of the work.

References

- [1] V. Čermák, L. Pícek, L. Adam, K. Papafitsoros, Wildlifedatasets: An open-source toolkit for animal re-identification, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5953–5963.
- [2] L. Adam, V. Čermák, K. Papafitsoros, L. Pícek, Wildlifereid-10k: Wildlife re-identification dataset with 10k individual animals, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2099–2109.
- [3] Q. Zhang, K. Ahmed, N. Sharda, H. Wang, A comprehensive survey of animal identification: Exploring data sources, ai advances, classification obstacles and the role of taxonomy, *International Journal of Intelligent Systems* 2024 (2024) 7033535. doi:<https://doi.org/10.1155/2024/7033535>.
- [4] A. M. Machado, M. Cantor, A simple tool for linking photo-identification with multimedia data to track mammal behaviour, *Mammalian Biology* 102 (2021). doi:[10.1007/s42991-021-00189-0](https://doi.org/10.1007/s42991-021-00189-0).
- [5] L. Pícek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Go"eau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqu'e, T. Denton, K. Barnard, C. N'edellec, L. Del'eger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: Ai challenges for biodiversity understanding and ecosystem management, in: *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*, Springer, 2026.
- [6] L. Adam, K. Papafitsoros, D. A. Williams, D. Biffi, L. Pícek, Overview of AnimalCLEF 2026: Discovery and re-identification of individual animals, in: *Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum*, 2026.
- [7] L. Adam, L. Pícek, K. Papafitsoros, D. Williams, D. Biffi, Animalclef26 @ cvpr & clef, <https://kaggle.com/competitions/animal-clef-2026>, 2026. Kaggle competition.
- [8] L. Pícek, J. Straka, M. Jirik, E. Belotti, M. Duľa, J. Krausová, M. Bojda, V. Čermak, L. Bufka, R. Dvořák, et al., Czechlynx: A dataset for individual identification and pose estimation of the eurasian lynx, *Scientific Data* 13 (2026) 511.
- [9] D. Biffi, M. R. Tucker, A. Ackel, D. A. Williams, Identification of individual texas horned lizards (*phrynosoma cornutum*) using genotypes and ventral spot patterns, *Ecology and Evolution* 15 (2025) e71167.
- [10] L. Adam, V. Čermák, K. Papafitsoros, L. Pícek, Seaturtleid2022: A long-span dataset for reliable sea turtle re-identification, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7146–7156.
- [11] L. Hubert, P. Arabie, Comparing partitions, *Journal of Classification* 2 (1985) 193–218. doi:[10.1007/BF01908075](https://doi.org/10.1007/BF01908075).
- [12] V. Čermak, L. Pícek, L. Adam, L. Neumann, J. Matas, Wildfusion: Individual animal identification with calibrated similarity fusion, in: *European Conference on Computer Vision*, Springer, 2024, pp. 18–36.
- [13] C. Biemann, Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems, in: *Proceedings of TextGraphs: the first workshop on graph based methods for natural language processing*, 2006, pp. 73–80.
- [14] R. Pakhomov, G. V. Demidov, K. Bogdan, S. Lanskich, D. Dinmuhametov, A. Khlopotnykh, Individual wildlife recognition via hybrid global-local matching and segmentation-aware filtering, in: *Conference and Labs of the Evaluation Forum*, 2025. URL: <https://api.semanticscholar.org/CorpusID:282253435>.
- [15] N. Semenova, Meta-algorithm for open-set animal re-id: Wildfusion, xgboost, and dual-backbone arcface, in: *CEUR Workshop Proceedings*, volume 4038, 2025, pp. 3154–3165.
- [16] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al., Sam 3: Segment anything with concepts, *arXiv preprint arXiv:2511.16719* (2025). doi:[10.48550/arXiv.2511.16719](https://doi.org/10.48550/arXiv.2511.16719).
- [17] E. Nepovinskykh, T. Eerola, V. Biard, P. Mutka, M. Niemi, M. Kunnasranta, H. Kälviäinen, Sealid: Saimaa ringed seal re-identification dataset, *Sensors* 22 (2022). URL: <https://www.mdpi.com/>

1424-8220/22/19/7602. doi:10.3390/s22197602.

- [18] R. Gonzalez, R. Woods, Digital Image Processing, Global Edition, Pearson Education, 2018. URL: <https://books.google.com.eg/books?id=p74oEAAAQBAJ>.
- [19] R. Szeliski, Computer Vision: Algorithms and Applications, Texts in Computer Science, 2 ed., Springer Cham, Cham, Switzerland, 2022. doi:10.1007/978-3-030-34372-9, hardcover ISBN: 978-3-030-34371-2; Softcover ISBN: 978-3-030-34374-3.
- [20] L. Otarashvili, T. Subramanian, J. Holmberg, J. Levenson, C. V. Stewart, Multispecies animal re-id using a large community-curated dataset, arXiv preprint arXiv:2412.05602 (2024).
- [21] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 24–25.
- [22] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018). doi:10.48550/arXiv.1802.03426.
- [23] Hyper3Labs, Hyperview: An interactive geometric workbench for embedding space analysis, 2026. URL: <https://github.com/Hyper3Labs/HyperView>.
- [24] X. Zhao, X. Wu, W. Chen, P. C. Chen, Q. Xu, Z. Li, Aliked: A lighter keypoint and descriptor extraction network via deformable transformation, IEEE Transactions on Instrumentation and Measurement 72 (2023) 1–16.
- [25] P. Lindenberger, P.-E. Sarlin, M. Pollefeys, Lightglue: Local feature matching at light speed, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 17627–17638.
- [26] M. Tyszkiewicz, P. Fua, E. Trulls, Disk: Learning local features with policy gradient, Advances in neural information processing systems 33 (2020) 14254–14265.
- [27] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1318–1327.
- [28] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699. doi:10.1109/CVPR.2019.00482.
- [29] J. Jiao, W. Liu, Y. Mo, J. Jiao, Z. Deng, X. Chen, Dyn-arcface: dynamic additive angular margin loss for deep face recognition, Multimedia Tools and Applications 80 (2021) 25741–25756.
- [30] M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 10096–10106.
- [31] Y. Wen, W. Liu, A. Weller, B. Raj, R. Singh, Sphereface2: Binary classification is all you need for deep face recognition, 2022. URL: <https://arxiv.org/abs/2108.01513>. arXiv:2108.01513.
- [32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988. doi:<https://doi.org/10.48550/arXiv.1708.02002>.
- [33] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631. doi:<https://doi.org/10.48550/arXiv.1907.10902>.
- [34] A. Rosenberg, J. Hirschberg, V-measure: A conditional entropy-based external cluster evaluation measure, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 410–420.