Meta-Algorithm for Open-Set Animal Re-ID: WildFusion, XGBoost, and Dual-Backbone ArcFace*

Nelly Semenova[1,*]

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

Abstract

This paper presents a three-stage meta-algorithm that addresses open-set individual animal re-identification. The cascade first employs WildFusion to fuse calibrated global–local similarity scores, then feeds concatenated MegaDescriptor-L and MIEW embeddings into an XGBoost classifier, and finally refines predictions with species-specific Dual-Backbone models fine-tuned using an ArcFace angular-margin loss. On the AnimalCLEF 2025 challenge, which includes loggerhead sea turtles, fire salamanders, and Eurasian lynxes and exhibits a pronounced long-tail imbalance, the proposed method achieved a private score of 67.42% and a public score of 65.11%, ranking 2nd out of 172 teams. Ablation analysis shows cumulative improvements of +21 percentage points (pp) from WildFusion over a MegaDescriptor baseline, +2.4 pp from XGBoost, and +3 pp from the Dual-backbone ArcFace stage. These results demonstrate that species-aware stacking of heterogeneous cues (global descriptors, calibrated local matches, tabular neighbor context, and metric fine-tuning) yields a robust and scalable solution for non-invasive wildlife monitoring.

Keywords

animal re-identification, wildfusion, megadescriptor, miew, vision transformer, efficientnet-v2, arcface

1. Introduction

Individual animal re-identification (Animal Re-ID) is the task of recognizing specific *individuals* in images. Accurate identity assignment is critical to ecology and wildlife conservation because it enables monitoring of population size, migration routes, and behavioral patterns of rare species in situ [1]. In Human Re-ID universal biometric cues such as faces or fingerprints are available, whereas these markers are not directly applicable to animals. Instead, recognition relies on unique natural markings—spot and stripe patterns, carapace mosaics, and similar traits—that vary markedly with viewpoint, pose, and illumination [2]. The problem is compounded by a shortage of labeled data: collecting and annotating photographs of individual animals is labor-intensive, so Animal Re-ID datasets are several orders of magnitude smaller than those used in Person Re-ID [3].

Traditional biological approaches—ringing, tagging, and DNA analysis—are reliable but invasive and unsuitable for large-scale monitoring [4]. Early computer-vision algorithms addressed one species at a time and relied on handcrafted features, which does not scale. With deep neural networks (first CNNs, later Vision Transformers [5]) Human Re-ID achieved a high level of accuracy, yet direct transfer to animals proved ineffective: the class set is open, inter-individual differences are subtle, and intra-species variability is high [6]. These factors motivated specialized methods for Animal Re-ID.

The **AnimalCLEF 2025** competition [7, 8] poses a multi-species challenge: identifying loggerhead sea turtles *Caretta caretta* (Greece), fire salamanders *Salamandra salamandra* (Czech Republic), and Eurasian lynxes *Lynx lynx* (Czech Republic) [9]. For each input image the system must decide whether the depicted animal belongs to a known *individual* in the training database (the *database set* or "gallery") or represents a *new individual*; if known, the correct ID must be returned. Consequently, the task combines classical Re-ID with an *open-set* component. Performance is evaluated by **BAKS** (balanced accuracy on known samples) and **BAUS** (balanced accuracy on unknown samples); the final score is the geometric mean of these two metrics [7].

CLEF 2025 Working Notes 9–12 September 2025, Madrid, Spain

*Corresponding author.
nelli.semenova@mail.ru (N. Semenova)
0000-0002-0190-8382 (N. Semenova)

Table 1
Detailed metadata coverage for each species in AnimalCLEF 2025

Species (Dataset)	Gallery	Query	Unique IDs	Date (DB / Q)	Orientation (DB / Q)
Eurasian lynx	2957	946	77	— / —	yes / yes
Fire salamander	1388	689	587	2017–2023 / 2024	yes / yes
Loggerhead sea turtle	8729	500	438	2010–2023 / 2024	yes / no

The meta-algorithm proposed in this study—combining WildFusion [1], XGBoost, and a Dual-backbone model with an ArcFace [10] head—ranked second among 172 teams, achieving a *private score* of 67.42% and a *public score* of 65.11% (team Webmaking). Subsequent sections describe the employed methods in detail (Sec. 3) and present a step-by-step analysis of the contribution made by each component to the final performance (Sec. 4).

2. AnimalCLEF 2025 challenge characteristics

2.1. Description and objectives

The primary goal of AnimalCLEF 2025 is to advance automated biodiversity monitoring, in particular the tracking of individual animals captured by camera traps and other imaging devices. Precise identification of *individuals* is pivotal in ecology: it enables reliable estimates of population size, migration routes, and behavioral profiles that underpin both scientific studies and conservation measures. Existing algorithms, however, tend to overfit to background or illumination cues and lose accuracy when applied to novel conditions. Consequently, the competition focuses on *universal* Re-ID approaches that can generalize across habitats and reliably recognize animals in a wide range of environments. Participants could either rely solely on the limited competition data or improve their models by pre-training on the large external dataset **WildlifeReID-10k** [11]. Overall, AnimalCLEF 2025 serves as a benchmark for state-of-the-art computer-vision methods and continues the LifeCLEF [8] series that expands the role of AI in wildlife monitoring.

2.2. Data and evaluation metric

For training and pre-training, participants were provided with the large WildlifeREID-10k dataset containing roughly 140,000 images of more than 10,000 individual animals across many species [11]. This external resource can be regarded as an extended training set. The competition data were collected specifically for AnimalCLEF 2025 and split into two parts: *Gallery*, which holds annotated images of known individuals and simultaneously serves as the training set and the gallery for matching, and *Query* (Tab. 1).

LynxID2025. This subset comprises 2957 training images of Eurasian lynx and 946 query images (3903 in total). The training split covers 77 unique individuals, with an average of 38 photographs per individual; the distribution is unbalanced—some animals have only a single image, whereas one individual appears in 353 shots. Image orientation is recorded for every picture (*left*, *right*, *front*, *back*, or *unknown*). Capture dates are not provided (the `date` field is empty).

SalamanderID2025. This subset contains 1388 training images of fire salamanders and 689 query images (2077 in total). The training split includes 587 unique individuals; the average is ~2.4 images per individual, the median is 1, and the maximum is 12. Orientation labels (*top*, *bottom*, *left*, *right*) are available for all images. Capture dates span 2017–2023 in the training set and extend to 2024 in the query set, enabling temporal analysis of the data collection period (for example, training pictures cover 2017–2023, while query images include shots made up to the end of 2024).

SeaTurtleID2022. This is the largest subset: 8729 training photographs of loggerhead sea turtles and 500 query images (9229 in total). The training split represents 438 unique sea turtles (mean 19.9 images per individual; median 13). Orientation labels include *left*, *right*, *front*, *top*, and composite directions such as *topleft* or *topright*; orientation is missing for all 500 query images. Capture dates are present for almost every photo, ranging from 2010 to 2024, reflecting the long-term nature of data collection.

The three subsets differ markedly. Lynx offers fewer individuals but more images per individual, whereas Salamander provides many individuals yet mostly single-image observations. Sea Turtle occupies an intermediate position in terms of individual count, but its total image volume is the largest. Such heterogeneity in size, orientation metadata, and temporal coverage underscores the need for adaptive identification strategies tailored to each species.

AnimalCLEF 2025 employs two metrics that jointly assess recognition quality on known and new individuals. BAKS is the per-class balanced accuracy over query images whose individuals are present in the gallery. BAUS is the balanced accuracy over query images belonging to new individuals absent from the gallery. The final ranking score is the geometric mean of BAKS and BAUS. The organizer split the query set into an open (*public*) portion comprising about 31% of the images and a hidden (*private*) portion comprising the remaining 69%. Only the private leaderboard determined the final standings, preventing overfitting to the public subset.

2.3. Related and preceding competitions

The task of individual animal identification had been explored before AnimalCLEF 2025. A notable predecessor is the Happywhale—Whale & DolphinID competition (Kaggle 2022), which required distinguishing thousands of individuals from 24 marine-mammal species using the mAP metric; the task suffered from a strong class imbalance but lacked an open-set component [12].

Between 2022 and 2024 several species-specific re-ID datasets were released together with mini-competitions, including Leopard ID and Hyena ID from WildMe & LILA Science [13, 14] and SeaTurtleID [15]. SeaTurtleID first introduced time-aware closed- and open-set splits later adopted by AnimalCLEF. An internal benchmark demonstrated 86.8% closed-set accuracy when using a Hybrid Task Cascade equipped with an ArcFace encoder, highlighting the challenges of long-term individual tracking even within a single species [15].

3. Methodology

3.1. State-of-the-art approaches and models

Modern Animal Re-ID methods rely on deep networks that extract image embeddings, i.e., compact feature vectors unique to each *individual*. Two principal categories of such features exist: **global descriptors** that summarize the entire image and **local matches** that align distinctive regions.

A prominent global approach is **MegaDescriptor** [16]. This supervised model is trained on a collection of many datasets (>10k individuals, ~140k images). Its backbone is a Swin-L Transformer with 384×384 input and about 229 M parameters. Essentially a Vision Transformer tuned for Animal Re-ID, it markedly outperforms generic models such as CLIP and DINOv2 [16].

An alternative global encoder is **MIEW** (Multi-species Individual Embeddings Wild, MiewID-msv3). This compact EfficientNet-V2 [17] CNN (about 51 M parameters) is trained with a contrastive Sub-center ArcFace loss on a dedicated dataset of 64 species (225k photos, 37k individuals). Unlike MegaDescriptor, which is trained per species, MIEW is optimized as a single multi-species model. Experiments show that this unified model surpasses species-specific training by an average of 12.5% top-1 and, more importantly, generalizes better to unseen species: on unknown taxa MIEW outperforms MegaDescriptor by 19.2% top-1 accuracy [3], demonstrating its ability to capture universal cues useful for any animal.

Global descriptors have limitations: they may miss fine-grained individual patterns. To compensate, **local methods** match image regions that carry unique markings. The modern **WildFusion** framework combines global and local information efficiently [1]. It fuses (i) cosine similarity of global embeddings (e.g., MegaDescriptor or DINOv2) and (ii) local keypoint correspondences obtained with matchers such as LoFTR [18] or LightGlue [19]. After isotonic calibration, the two similarity sources are merged into a single score. In a zero-shot setting WildFusion exceeds the pretrained MegaDescriptor-L, confirming that hybrid cues can substantially improve Re-ID performance [1].

A common path to higher accuracy is ensemble learning. In this work several ways of combining embeddings from MegaDescriptor and MIEW were explored. The best result was achieved by a meta-algorithm that blends predictions from (1) WildFusion (Sec. 3.3), (2a) XGBoost (Sec. 3.5), and (2b) a Dual-backbone network with an ArcFace head (Sec. 3.6). The reliable WildFusion, together with two strong embedding streams, proved highly effective: the transformer-based MegaDescriptor yields rich global representations, whereas the CNN-based MIEW provides features that remain robust on new species (Sec. 4).

3.2. MegaDescriptor-L-384 model

MegaDescriptor-L-384 is a foundation model for Animal Re-ID introduced in [16] and released on Hugging Face [20]. The backbone is *swin_large_patch4_window12_384* with 384×384 px input and 228.8 M parameters; it outputs a 1536-dimensional L2-normalized embedding suitable for cosine comparison.

The network was trained in a supervised manner with an ArcFace-style margin loss on the aggregated WildlifeDatasets corpus comprising 29 public datasets (~140k images, >10k individuals, 23 species). Merging such diverse sources exposes the model to wide variations in viewpoint, illumination, and marking patterns, thereby improving embedding generality. The authors report that MegaDescriptor-L-384 consistently outperforms CLIP (ViT-L/336) and DINOv2 (ViT-L/518) on all 29 benchmarks [16].

In practice, deployment requires only standard preprocessing: resize to 384×384, convert to a tensor, and normalize to means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. A single forward pass then produces the embedding [20]. The CC-BY-NC-4.0 license permits non-commercial use and modification, making MegaDescriptor-L-384 a strong out-of-the-box global descriptor within the pipeline.

3.3. WildFusion similarity fusion method

WildFusion [16] addresses a core limitation of purely global embeddings in Animal Re-ID—their sensitivity to background and illumination. By combining global image similarity with precise local keypoint verification, the method sharply reduces false matches between different *individuals* while recovering true correspondences under strong viewpoint changes. A detailed analysis of its impact on the final *private score* is presented in Sec. 4.

The algorithm comprises two stages. First, a fast cosine search in the MegaDescriptor-L-384 embedding space (Sec. 3.2) selects the $K = 300$ most similar gallery images (candidates). Each "query / candidate" pair is then evaluated by five independent local pipelines: *LoFTR*, *SuperPoint* [21], *ALIKED* [22], *DISK* [23], and *SIFT* [24]. For LoFTR, images are converted to 192×192 px grayscale, whereas the other pipelines operate on 512×512 px color inputs. The local scores and the normalized global similarity undergo isotonic calibration and are subsequently fused linearly into a single probabilistic score.

An open-source implementation is provided in the `wildlife-tools` package [25] in the `wildlife-datasets` repository [16]. Released under the GPL-3.0 license, the code requires no additional training, making WildFusion easy to integrate into an existing pipeline.

3.4. k-Reciprocal re-ranking strategy

On top of WildFusion probabilities, **k-reciprocal re-ranking** [26] is applied. Three similarity matrices S are available: "query × gallery", "gallery × query", and "gallery × gallery". For each image a list of the $K_1 = 20$ most similar neighbors is formed according to S. A gallery image x retains only those neighbors y that simultaneously place x within their first $K_2 = 6$ positions; the resulting mutual set is denoted $\mathrm{rc}(x)$. An analogous procedure yields $\mathrm{rc}(q)$ for every query, as the symmetric "gallery × query" matrix enables reciprocity checks in the reverse direction.

The final similarity between a query q and a gallery image x is calculated as

$$S_{\text{final}} = 0.9\, S_{\text{orig}} + 0.1\, S_{\text{Jaccard}}, \quad S_{\text{Jaccard}} = \frac{|\mathrm{rc}(q) \cap \mathrm{rc}(x)|}{|\mathrm{rc}(q) \cup \mathrm{rc}(x)|}.$$

This linear convex combination suppresses incidental matches caused by pose, masking, or background, while requiring no additional model training.

3.5. Gradient boosting on combined MegaDescriptor and MIEW embeddings

Score-CAM [27] heatmaps in Fig. 1 indicate that MegaDescriptor-L-384 focuses on compact texture regions, whereas MIEW-msv3 distributes attention across fine-grained spots and extended contours. The theoretical complementarity of these spatial patterns motivates a direct concatenation of the two embeddings ($\mathbb{R}^{3688}$), with each component pre-normalized by its L_2 norm [3, 16].

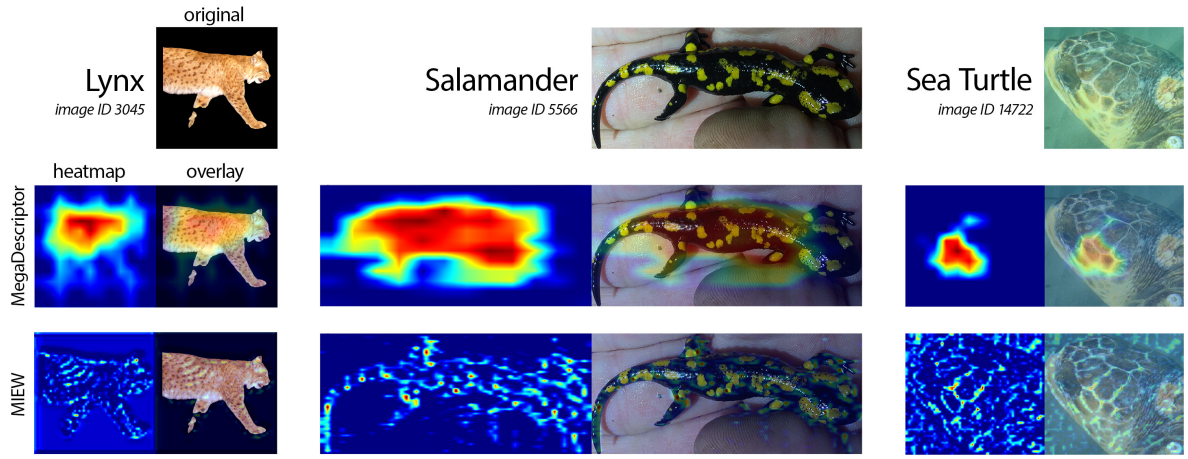

Figure 1: Score-CAM visualizations for two complementary identification models. The first row shows the original query images for three AnimalCLEF 2025 taxa. Rows 2–3 display *heatmaps* (left) and *overlay* views (right) obtained using Score-CAM for MegaDescriptor-L-384 (row 2) and MIEW-msv3 (row 3). For computational efficiency each map was computed with the top 40 activation channels of the target layer (last block for MD, block −4 for MIEW), a setting that retains over 97% of localization accuracy.

For every image, the feature vector includes (i) the concatenated global embedding, (ii) K=10 cosine distances to the ten nearest gallery images together with the corresponding *neighbor identifiers passed as categorical features*[1], and (iii) one-hot representations of view orientation and dataset membership (lynx, salamander, sea turtle). The key idea is that gradient boosting can non-linearly merge global descriptors, local density information in the embedding space, and categorical data on the closest *individuals*. The maximum depth was capped at 6, which prevents the model from memorizing category values via long split chains.

Incorporating both global descriptors enhances feature diversity: the transformer-based MegaDescriptor captures coarse texture patterns, whereas the CNN backbone of MIEW remains sensitive to

[1]The columns `nn_id_1…10` are cast to `pandas.Categorical` and processed by XGBoost's `enable_categorical` option, which learns optimal subset splits instead of numerical thresholds.

Table 2
Incremental impact of successive modules on private and public scores (metric = geometric mean **BAKS** × **BAUS** [%]; Δ denotes the change relative to the previous step).

Step	Modification introduced	Private	Δ	Public	Δ
0	**Baseline provided by the competition organizers:** MegaDescriptor-L threshold 0.6 for all	30.90	—	30.00	—
1	**Baseline:** MegaDescriptor-L cosine nearest neighbor with per-species thresholds	40.59	+9.69	37.92	7.92
2	WildFusion global + local similarity fusion with thresholds	61.72	+21.13	58.98	+21.06
3	*k*-reciprocal re-ranking (Lynx only) applied to WildFusion	62.09	+0.37	59.09	+0.11
4	XGBoost meta-classifier on MegaDescriptor + MIEW embeddings; WildFusion confidence adjustment	64.44	+2.35	61.89	+2.80
5	Dataset-specific meta-algorithm that combines WildFusion, XGBoost and Dual-backbone ArcFace	**67.42**	+2.98	**65.11**	+3.22

point-wise differences [3, 16]. XGBoost trained on this concatenation, augmented with local density features and metadata, produces a consistent improvement in the *private score* and *public score* relative to a single embedding baseline and to WildFusion alone (Sec. 4).

3.6. Dual-backbone model with an ArcFace head

The two previous ensemble components (WildFusion, Sec. 3.3, and XGBoost, Sec. 3.5) rely on *fixed* embeddings obtained without species-specific fine-tuning. Although this delivers high baseline accuracy, the capacity to adapt to species-specific visual patterns remains limited. The Dual-backbone model addresses the opposite need: it refines features *per species* and thus complements the rigid matching scheme of WildFusion and the tabular classifier XGBoost. Methodologically the model unites deep metric optimization via ArcFace [10] with the direct feature focus provided by two heterogeneous backbones.

The first stream employs **MegaDescriptor-L-384**, reliable at capturing global textures; the second employs MIEW-MSV3, sensitive to fine pointwise details. Both classification heads are removed, and their outputs after individual BatchNorm layers are concatenated into a 3688-dimensional vector.

A compact ArcFace head is placed on top of the joint space. ArcFace maximises inter-class angular margins in the embedding space, imposing a strict separability criterion. This margin-based approach is particularly effective under the small-sample conditions characteristic of AnimalCLEF 2025 [10].

WildFusion depends on a calibrated "global + local" heuristic, and XGBoost on tabular aggregation of fixed embeddings and metadata. Three independent Dual-backbone models, one trained for each species, supply descriptors tailored to their respective AnimalCLEF 2025 subsets and thereby improve the robustness of the ensemble.

4. Results and Discussion

Tab. 2 summarizes the step-by-step impact of every module on the final metric. The baseline cosine search with MEGADESCRIPTOR-L-384 already provides a reasonable baseline, yet each subsequent component steadily improves the score.

Fig. 2 plots every submission in the coordinates "fraction of *new individual* predictions / leaderboard score" for each species. The gap between public and private scores is small and relatively stable, indicating that the algorithm does not overfit the public portion of the test set. The five sequential development steps are discussed below.

Step 1. Threshold selection for the global descriptor MEGADESCRIPTOR-L-384 (HF hub: `BVRA/MegaDescriptor-L-384`; batch 32, input 384×384) was evaluated with both shared and thresh-

olds (Sec. 3.2). Each gallery image (n = 13,074) and each query image (n = 2135) was encoded as a 1536-dimensional vector. Cosine similarity was computed between every query vector and every gallery vector; the most similar gallery image provided the candidate identity for the query. If the similarity fell below the threshold, the label *new_individual* was assigned.

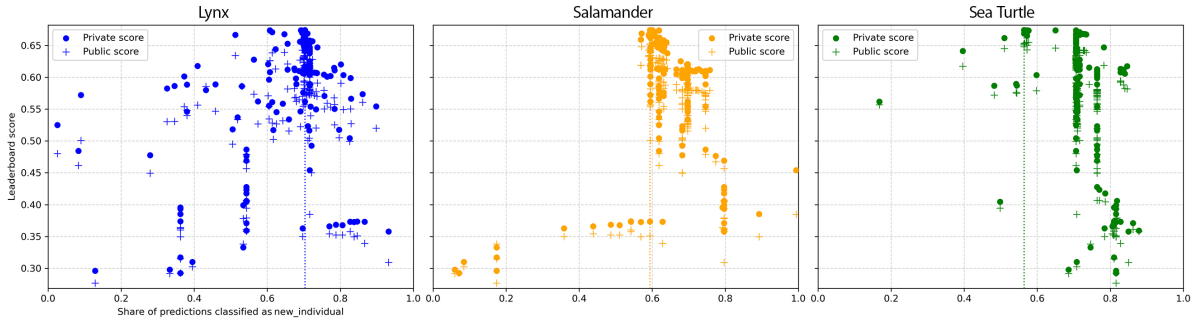

Figure 2: Private (●) and public (+) leaderboard scores versus the share of *new_individual* predictions for each AnimalCLEF 2025 species. Each symbol represents one submission; the abscissa shows the proportion of images labelled as *new_individual*, the ordinate shows its score on the respective leaderboard split. The dotted vertical line marks the best private submission (Y = 0.6742) together with the associated *new individual* shares: Lynx 70.33%, Salamander 59.36%, Sea turtle 56.40%. The horizontal axis reports the share of images that the model classifies as new individual, expressed in percentage of the entire query.

A grid search over a single global threshold in the range 59.0–74.5% yielded the best result at 74.0% (public 35.76%, private 37.32%). Separate thresholds were then explored for each taxon: Lynx 50–90%, Salamander 60–90%, Sea turtle 74–90%. The optimal triplet (Lynx 65.5%, Salamander 77.0%, Sea turtle 74.5%) produced a public score of 37.92% and a private score of 40.59%, establishing the **baseline** for subsequent steps.

Preliminary analysis of Fig. 2 revealed that submissions cluster vertically: runs with similar *fractions of new_individual predictions* tend to yield comparable public scores, and the highest–scoring points concentrate around species-specific shares of 70% (lynx), 60% (salamander) and 60% (sea turtle). Therefore, at all later steps (including the WildFusion, XGBoost confidence gate, and the final cascade) thresholds were selected so as to preserve these empirically favorable ratios. This policy explains the multiple vertical stripes visible in Fig. 2: each stripe marks a family of submissions that intentionally maintain the same *new_individual* quota while refining other components of the pipeline.

Possible future work includes replacing manual threshold search with probability–calibration techniques such as *Platt scaling* or *temperature scaling*, which learn a monotone mapping on the validation split and may further stabilize the species-specific *new_individual* ratios without exhaustive grid search.

Step 2. WildFusion: fusion of global and local features The open implementation of **WildFusion** [16, 25] is employed at this step (Sec. 3.3).

First, MegaDescriptor-L-384 retrieves the K = 300 gallery candidates with the highest cosine similarity; all 15,209 images (13,074 gallery + 2135 query) are encoded as 1536-dimensional vectors. Each ”query / candidate” pair is then evaluated by five independent local matchers: SuperPoint–LightGlue, ALIKE–LightGlue, DISK–LightGlue, SIFT–LightGlue (all at 512 × 512 px RGB) and LoFTR (192 × 192 px grayscale). Figure 3 shows that the detectors yield a comparatively small overlap of correspondences; this diversity underlies the gain obtained after isotonic calibration and score fusion. Calibrating the five matchers on the validation split required 3h on a single A100 GPU; the full *query* × *gallery* evaluation took a further 27h.

After combining global and local signals, thresholds for the *new_individual* label were tuned separately per species. The optimal values were Lynx 39.5%, Salamander 12.0%, and Sea turtle 16.0%.

This configuration achieved a **58.98%** public score and a **61.72%** private score, yielding a +21 pp improvement over the baseline in Tab. 2. Notably, the tuned WildFusion stage alone would already have secured an 8th place finish on the final leaderboard, even before adding the later cascade components.

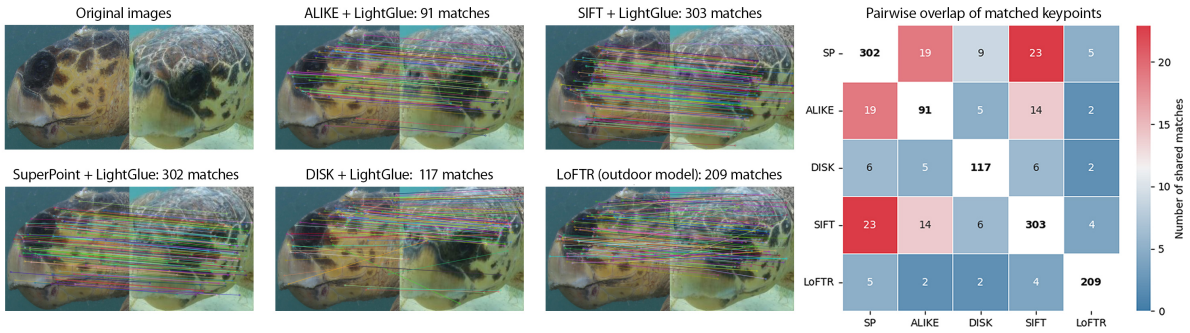

Figure 3: Visual comparison of local-matcher pipelines and their pair-wise keypoint overlap. *Left two columns*: the original query image (`ID 14793`, file `2b0a454de7cf2873_84.JPG`) and the nearest gallery candidate (`ID 10958`, file `tXNSzEXIfE_4979.JPG`, label `SeaTurtleID2022_t257`, both images depict the same individual). The next five panels show matches returned by the four LightGlue-based pipelines (SuperPoint, ALIKE, DISK, SIFT) and LoFTR (outdoor model). *Right*: heat-map of the *Pairwise overlap of matched keypoints*. Each cell reports how many distinct correspondences are shared by two pipelines after (i) rounding all coordinates to a grid of 1 px (`BIN` = 1) and (ii) merging matches whose endpoints fall within a tolerance of 5 px in both images (`THR` = 5). Diagonal values give the total number of matches produced by each method; off-diagonal values quantify complementarity, with warmer colors indicating a larger intersection.

The results confirm that merging global embeddings with complementary local keypoints is crucial for the substantial performance gain observed on AnimalCLEF 2025.

Step 3. *k*-reciprocal re-ranking over WildFusion outputs A full "gallery × gallery" similarity matrix was computed for the Lynx subset and re-ranking was applied according to the scheme of Z. Zhong [26]. The method parameters were fixed to the first neighborhood radius $K_1 = 20$, the reciprocity radius $K_2 = 6$, and the Jaccard weight $\lambda = 0.1$ in the linear combination with the original WildFusion score (Sec. 3.4).

Lynx was selected because its images share an artificially uniform black background, which increases the risk of pose-driven false matches; reciprocal filtering helps to attenuate this artifact. Building the *gallery × gallery* matrix required an additional 38 GPU-hours, so the procedure was not executed for Salamander or Sea turtle.

The gain, although modest, was positive: the public score rose from 58.98% to 59.09% (+0.11 pp), and the private score from 61.72% to 62.09% (+0.37 pp). This confirms the value of mutual neighbor filtering, yet the improvement did not justify the computational cost; all subsequent steps therefore relied on the original WildFusion scores (for Salamander or Sea turtle).

Step 4. Gradient-boosted ensemble of MegaDescriptor and MIEW embeddings For every image a dense feature vector of **3718** dimensions was assembled: the 3688-D concatenation of MegaDescriptor-L-384 and MIEW-msv3 embeddings, 10 cosine distances to the nearest gallery images, 10 categorial identifiers of those neighbors, and 10 one-hot categories (7 *orientation* flags + 3 *dataset* flags). A unified probability scale simplifies the tuning of the cascade.

An XGBoost model was trained with depth 6, $\eta = 0.15$, $\lambda = 2.0$, `tree_method=gpu_hist` and the `multi:softprob` objective; the best iteration was reached at round 296. Validation followed a "single image per *individual*" split (Sec. 3.5).

During inference the posterior probabilities of XGBoost acted as a confidence gate on top of WildFusion. Species-specific thresholds were tuned empirically: for Salamander, the WildFusion label was replaced when XGBoost confidence exceeded 20%; for Sea turtle, when it exceeded 95%. This cascaded refinement raised the *public score* to 61.89% and the *private score* to 64.44%, adding +2.80 pp and +2.35 pp, respectively, over the pure (re-ranked) WildFusion.

Step 5. Dual-backbone model with an ArcFace head and its integration into the meta-algorithm For each taxon an individual Dual-backbone network was trained that combines MegaDescriptor-L-384 with MIEW-MSV3. After separate `BatchNorm` layers the two vectors were concatenated into a 3688-dimensional feature, which was fed to a compact *ArcFace* head (Sec. 3.6). The head parameters were fixed per dataset: $(s, m) = (64, 0.5)$ for Lynx and Sea turtle, $(30, 0.35)$ for Salamander.

Augmentation pipelines were tailored to the visual specifics of each dataset. *Lynx*: background-mask removal followed by `RandomResizedCrop` with scale ≥ 0.9. *Salamander*: rotation according to the *orientation* field and moderate cropping that preserves key anatomical regions. *Sea turtle*: moderate cropping plus horizontal flip. All datasets additionally received `ColorJitter` and `CoarseDropout` (one mask $\leq 10\%$ of the image area).

Data were split in a stratified fashion: 90% of images for training, 10% for validation. Class imbalance over *individuals* was mitigated with a `WeightedRandomSampler`.

Optimization employed SGD in three stages: (1) a two-epoch initial training phase only the ArcFace head and the uppermost 25% of layers at learning rate $\eta = 10^{-2}$; (2) full backbone unfreezing with a base step $\eta_0 = 5{\times}10^{-3}$ under a cosine-annealing schedule; (3) a final fine-tuning stage of the last two epochs at $\eta = 10^{-4}$.

After fine-tuning, L_2-normalized *gallery* and *query* embeddings were indexed in a FAISS `IndexFlatIP` [28, 29]; for each query the 50 nearest neighbors were retrieved, and confidence was defined as $(\cos +1)/2$.

Descriptors from the three Dual-backbone models complemented WildFusion and XGBoost inside the final meta-algorithm, increasing ensemble robustness on challenging and rare cases and yielding an additional score gain (Tab. 2).

Final meta-algorithm and overall leaderboard performance The definitive submission followed a cascading scheme that invoked one to three models per species.

Eurasian lynx. (1) WildFusion predictions after k-reciprocal re-ranking ($\lambda = 0.1$); images with confidence below 39.7% assigned the label *new_individual*. (2) When XGBoost assigns a probability $\geq 99\%$, its class replaces the WildFusion label. (3) Dual-backbone embeddings serve as the final filter: similarity $< 64\%$ converts the label to *new_individual*, whereas similarity $> 89.3\%$ overwrites the class with the Dual-backbone prediction.

Fire salamander. (1) WildFusion with a confidence threshold of 13.0%. (2) Dual-backbone refines the outcome: similarity $< 62\%$ results in assigning the label *new_individual*, while similarity $> 80\%$ accepts the Dual-backbone label.

Loggerhead sea turtle. Dual-backbone operates as the exclusive source: similarity $< 70.3\%$ is interpreted as *new_individual*; otherwise the identifier proposed by the model is retained.

This species-specific combination of WildFusion, XGBoost, and Dual-backbone merges their errors that exhibit low mutual correlation. The ensemble achieved a **private score of 67.420%** and a **public score of 65.114%**, securing second place among 172 teams in AnimalCLEF 2025 (Tab. 2).

5. Practical significance and prospects for future work

High-precision Animal Re-ID methods open new opportunities in both scientific research and applied domains. These techniques facilitate automated census work and substantially facilitate field studies: instead of capturing and tagging animals, researchers can deploy camera traps, drones, or underwater cameras and then analyze the collected imagery algorithmically. Such approaches are already employed to monitor endangered species e.g., identifying snow leopards or whales allows estimating population size, migration routes, and individual longevity. Accurate and scalable Animal Re-ID thus constitutes a key enabling factor of non-invasive biodiversity monitoring.

Future improvements are envisioned along four complementary directions.

(i) Transductive graph-based models. AnimalCLEF 2025 data exhibit a pronounced long-tail distribution in the number of images per *individual*. Under these circumstances, graph re-ID strategies

such as GCN-based reranking may redistribute confidence from majority classes to minority classes and raise recall in the tail of the distribution [26, 30]. Initial GCN experiments reduced the public score; nevertheless, further exploration of neighborhood radii and regularization schemes remains promising.

(ii) Tiling and localized matching. Dividing images into distortion-free squares and performing pairwise tile matching leads to a quadratic growth in computational complexity and GPU memory consumption, yet can mitigate background influence and raise confidence for fine-scale spotted patterns.

(iii) Pseudo-labeling and self-training. Unlabeled images from GBIF (Global Biodiversity Information Facility) or surveillance video streams can augment the training set. A *confident set* may be formed with current models, followed by additional backbone fine-tuning while strictly controlling pseudo-label accuracy.

(iv) Automatic threshold tuning. Bayesian optimization or differentiable threshold tuning on the validation score would eliminate manual adjustment of the 39.7%, 13%, and 70.3% thresholds and adapt the meta-algorithm to new species without manual threshold tuning [31].

Acknowledgments

The author is grateful to all individuals and organizations involved in the collection and annotation of data that enable the training of models and the development of animal re-identification tools.

Declaration on Generative AI

During the preparation of this work, the following Generative AI tool was employed:

- **ChatGPT o3** (OpenAI, June 2025 model) — *Text Translation* of the paper from Russian to English, *Grammar and spelling check*, and *Improve writing style*.

All AI-generated suggestions were reviewed and edited manually; the authors assume full responsibility for the final content. No Generative AI system was used to create original scientific ideas, analyze data, or draw conclusions.

References

[1] V. Cermak, L. Picek, L. Adam, L. Neumann, J. Matas, Wildfusion: Individual animal identification with calibrated similarity fusion, in: European Conference on Computer Vision, Springer, 2025, pp. 18–36.

[2] Y. Lin, L. Liu, J. Shi, Categorical keypoint positional embedding for robust animal re-identification (2024). URL: https://arxiv.org/abs/2412.00818. arXiv:2412.00818.

[3] L. Otarashvili, T. Subramanian, J. Holmberg, J. J. Levenson, C. V. Stewart, Multispecies animal re-id using a large community-curated dataset, 2024. URL: https://arxiv.org/abs/2412.05602. arXiv:2412.05602.

[4] D. T. Bolger, T. A. Morrison, B. Vance, D. Lee, H. Farid, A computer-assisted system for photographic mark–recapture analysis, Methods in Ecology and Evolution 3 (2012) 813–822. URL: https://doi.org/10.1111/j.2041-210X.2012.00212.x. doi:10.1111/j.2041-210X.2012.00212.x.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16×16 words: Transformers for image recognition at scale, in: Proceedings of the 9th International Conference on Learning Representations (ICLR), ICLR, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[6] Y. Wu, D. Zhao, J. Zhang, Y. S. Koh, An individual identity-driven framework for animal re-identification, 2024. URL: https://arxiv.org/abs/2410.22927. arXiv:2410.22927.

[7] L. Adam, K. Papafitsoros, R. Kovář, V. Čermák, L. Picek, Overview of AnimalCLEF 2025: Recognizing individual animals in images, 2025.

[8] L. Picek, S. Kahl, H. Goëau, L. Adam, T. Larcher, C. Leblanc, M. Servajean, K. Janoušková, J. Matas, V. Čermák, K. Papafitsoros, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, J. S. Cañas, G. Martellucci, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2025.

[9] L. Picek, E. Belotti, M. Bojda, L. Bufka, V. Cermak, M. Dula, R. Dvorak, L. Hrdy, M. Jirik, V. Kocourek, et al., Czechlynx: A dataset for individual identification and pose estimation of the eurasian lynx, arXiv preprint arXiv:2506.04931 (2025).

[10] J. Deng, J. Guo, X. Niannan, 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.

[11] L. Adam, V. Cermak, K. Papafitsoros, L. Picek, Wildlifereid-10k: Wildlife re-identification dataset with 10k individual animals, in: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops, 2025, pp. 2099–2109.

[12] Happywhale – whale and dolphin identification, Kaggle Competition, 2022. URL: https://www.kaggle.com/competitions/happy-whale-and-dolphin.

[13] Wild Me, LILA Science, Leopard id dataset, Dataset, 2022. URL: https://lila.science/datasets/leopard-id.

[14] Wild Me, LILA Science, Hyena id dataset, Dataset, 2022. URL: https://lila.science/datasets/hyena-id.

[15] L. Adam, V. Čermák, K. Papafitsoros, L. Picek, 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.

[16] V. Čermák, L. Picek, 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.

[17] M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (ICML), volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10096–10106. URL: https://proceedings.mlr.press/v139/tan21a.html, long presentation.

[18] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, Loftr: Detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8928–8937. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Sun_LoFTR_Detector-Free_Local_Feature_Matching_With_Transformers_CVPR_2021_paper.html.

[19] P. Lindenberger, P.-E. Sarlin, M. Pollefeys, Lightglue: Local feature matching at light speed, arXiv preprint arXiv:2306.13643 (2023). URL: https://arxiv.org/abs/2306.13643.

[20] B. V. R. Alliance, Bvra/megadescriptor-l-384, https://huggingface.co/BVRA/MegaDescriptor-L-384, 2024. Model card on Hugging Face; accessed 2025-07-03.

[21] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, 2018. URL: https://arxiv.org/abs/1712.07629. arXiv:1712.07629.

[22] X. Zhao, X. Wu, W. Chen, P. C. Y. Chen, Q. Xu, Z. Li, Aliked: A lighter keypoint and descriptor extraction network via deformable transformation, 2023. URL: https://arxiv.org/abs/2304.03608. arXiv:2304.03608.

[23] M. J. Tyszkiewicz, P. Fua, E. Trulls, Disk: Learning local features with policy gradient, 2020. URL: https://arxiv.org/abs/2006.13566. arXiv:2006.13566.

[24] D. G. Lowe, Distinctive image features from scale–invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

[25] WildlifeDatasets, wildlife-tools: Reference implementation of wildfusion, https://github.com/WildlifeDatasets/wildlife-tools, 2024. Accessed 2025-07-03.

[26] 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 (CVPR), 2017, pp. 4656–4665. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Zhong_Re-Ranking_Person_Re-Identification_CVPR_2017_paper.html.

[27] H. Wang, Z. Du, M. Du, F. Yang, S. Hu, S. Liu, J. Zhou, 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 (CVPR W), IEEE, 2020, pp. 111–119. doi:10.1109/CVPRW50498.2020.00020.

[28] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library (2024). arXiv:2401.08281.

[29] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.

[30] Y. Zhang, Q. Qian, H. Wang, C. Liu, W. Chen, F. Wang, Graph convolution based efficient re-ranking for visual retrieval, 2023. URL: https://arxiv.org/abs/2306.08792. arXiv:2306.08792.

[31] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 2951–2959. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.