Anchored Graph Clustering with Local Feature Matching for AnimalCLEF 2026

Gilles Colling[1,2]

[1]Division of BioInvasions, Global Change & Macroecology, Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria
[2]Vienna Doctoral School of Ecology and Evolution, University of Vienna, Vienna, Austria

Abstract

AnimalCLEF 2026 evaluates individual animal discovery and re-identification across four species: Eurasian lynx, fire salamander, loggerhead sea turtle, and Texas horned lizard. Submissions assign test images to identity clusters and are evaluated using the Adjusted Rand Index (ARI) against hidden ground-truth identities. We present a per-species anchored graph-clustering framework combining pretrained re-identification descriptors (MiewID and MegaDescriptor), LightGlue local feature matching, and a tabular pairwise classifier for refinement edges. The system anchors the clustering graph to the labelled reference set: when a test image's blended similarity to a known training individual exceeds a species-specific threshold, it inherits that identity anchor. Test images sharing the same anchor are grouped together, while non-anchored pairs are merged only when supported by a high-confidence pairwise classifier score. Our selected submission achieved 0.61741 ARI on the public leaderboard and 0.57038 ARI on the private leaderboard.

Keywords

animal re-identification, individual identification, LightGlue, local feature matching, graph clustering, Adjusted Rand Index, AnimalCLEF, MiewID, MegaDescriptor

1. Introduction

Animal re-identification links a photograph to an individual animal. AnimalCLEF 2026 frames this as a *discovery* and clustering task: every test image is assigned to a cluster meant to represent one individual, and the test set may include individuals never seen during training, so the predicted cluster labels are themselves the discovery output rather than retrieval keys drawn from a closed training vocabulary [1, 2]. The official metric is the Adjusted Rand Index [8, 12], which compares the predicted clustering to the hidden identity ground truth after correcting for chance agreement. ARI penalises both directions of error: a false merge joins two animals and produces many wrong same-individual pairs, while a missed merge splits one animal across several clusters.

The competition spans four species with different individual-marking modalities: Eurasian lynx (coat patterns), fire salamander (dorsal spot patterns), loggerhead sea turtle (head and flipper scutes), and Texas horned lizard (ventral spot patterns) [3, 4]. Each species places different demands on the matching system: some are well-separated by global appearance descriptors, others require local feature matching to distinguish individuals on small markings.

We submitted a per-species graph-clustering pipeline. Pretrained re-identification backbones produce candidate similarity scores; LightGlue local matching produces a complementary score; species-specific thresholds turn high-confidence scores into edges in an undirected graph. Connected components of this graph become the predicted clusters. The pipeline anchors the graph through known training identities: for species with a labelled reference set, a test image whose nearest training neighbour exceeds the species threshold inherits that training individual's identity, and two test images mapped to the same training individual are merged automatically. This bridge produced most of the safe merges in the final submission.

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany
gilles.colling051@gmail.com (G. Colling)
0000-0003-3070-6066 (G. Colling)

The contributions of this paper are: (i) a description of a catalogue-anchored graph-clustering pipeline that scored 0.61741 ARI on the public leaderboard and 0.57038 ARI on the private leaderboard, with 833 clusters across 2 409 test images; (ii) an analysis showing that test-test-only redesigns that drop the anchor consistently lose public-LB ARI, even when their pair-level diagnostics improve; (iii) reproducible per-species similarity blends and thresholds, with submitted-CSV SHA-256 hashes for traceability; and (iv) a documented set of failed experiments (head-crop matching, anatomy-based shape features, hierarchical agglomerative clustering, self-supervised backbone fine-tuning) to inform future submissions.

2. Related Work

2.1. Animal re-identification

Animal re-identification has shifted from species-specific hand-crafted pipelines to general-purpose deep metric learning [5, 7]. The WildlifeDatasets toolkit [5] aggregates many existing per-species datasets and provides a uniform evaluation framework; the associated MegaDescriptor model is a Swin-based backbone trained on this corpus across many species. WildlifeReID-10k [7] extends the same line with over 10 000 individuals across additional species, primarily as a pretraining corpus for general-purpose re-identification backbones. MiewID [6] is a multispecies re-identification model trained on a different community-curated corpus; it tends to generalise better than MegaDescriptor to unseen species, and the two models are complementary in practice.

2.2. Local feature matching

Local feature matching answers a different question than embedding similarity: whether visual points in two images can be put into geometric correspondence. LightGlue [11] is an attention-based matcher that operates on deep keypoint descriptors and that runs roughly an order of magnitude faster than its predecessor SuperGlue. Local matching is well-suited to patterned animals where individual identity is encoded in spots, scales, scutes, or markings rather than in overall body shape, and several recent re-identification systems combine local and global signals through calibrated fusion [5].

2.3. Open-set and clustering formulations

In a closed-set re-identification formulation, the system returns the most likely training identity for each query image. AnimalCLEF 2025 used an open-set variant in which each test image had to be classified either as a known training identity or as "new" [7], scored by a balanced known/unknown accuracy. AnimalCLEF 2026 changes the framing further: every test image must be assigned to a cluster, scored by ARI between the predicted and ground-truth partitions. This converts re-identification into a constrained graph-clustering problem in which global decisions about merges and splits matter as much as local pair decisions.

2.4. The Adjusted Rand Index

The Rand Index counts the fraction of item pairs that are placed consistently between two partitions of the same set, where consistent means grouped together in both partitions or kept apart in both. The Adjusted Rand Index [8] rescales this fraction so that a random labelling has expected score 0 and a perfect match has score 1. We use the scikit-learn implementation [12] for local cross-validation.

3. Task and Data

The official AnimalCLEF page links to the Kaggle competition that hosted the released data, sample submission, and leaderboard [1, 2]. The lab page lists two headline species (Eurasian lynx and loggerhead sea turtle); the Kaggle release extends this to four species namespaces:

- `LynxID2025` (946 test images),
- `SalamanderID2025` (689 test images),
- `SeaTurtleID2022` (500 test images), derived from the SeaTurtleID2022 study [3],
- `TexasHornedLizards` (274 test images), linked to the Texas horned lizard study [4].

Each namespace ships with its own train split labelled with individual identities; the test split contains individuals that may or may not appear in the train split. Submissions are CSV files with columns `image_id` and `cluster`, where `cluster` is a label string scoped within the species namespace. Two test rows that share a cluster label are predicted to show the same individual.

The official metric is ARI, computed against the hidden ground-truth partition. Public-leaderboard scores are computed on a fraction of the test set during the competition; private (final) scores are computed on the complement and revealed at the end. Throughout this paper we report both: public-leaderboard ARI for development comparisons (where private scores were unavailable while we ran the experiments), and private-leaderboard ARI for the selected final submission.

Per-species visual cues. The four species place different demands on the matching system. Lynx individual identity lives in coat-spot patterns on the flanks and limbs; pose and lighting variation is large but coat patterns are reasonably stable across visits. Salamander individuals are separated by yellow dorsal spot patterns on a black background; the patterns are highly individual but image-to-image alignment is poor (zoom, body coverage, head-to-dorsal axis vary). Sea turtle individuals are identified from facial scute patterns visible in head shots [3]; aquatic backgrounds add water reflections and motion blur. Texas horned lizard individuals are identified from ventral spot patterns [4], with limited overlap between training and test photographs in our package, which makes test-test pair evidence the primary signal for this species.

4. System Overview

Figure 1 sketches the pipeline. The system runs independently per species: a lynx image is never compared to a salamander image. Within each species, three similarity signals are computed: global descriptor similarity, LightGlue local-matching similarity, and a learned pair-classifier score. These are combined into per-species similarity matrices, gated by species-specific thresholds, and resolved into connected components.

The submission build proceeds in seven steps:

1. Load metadata, image paths, the train/test split, and the labelled train identities.
2. Extract or load global descriptors from MiewID and MegaDescriptor for every image.
3. Extract or load LightGlue match scores for the train-test and test-test pairs that pass an initial global-similarity prefilter.
4. For Lynx, Salamander, and Sea Turtle, score every test image against every train identity using s_{final} and assign a train identity when the blended score exceeds τ_{anchor}.
5. For Horned Lizards, where the train-anchor signal is weak, build a test-test edge set directly from pair evidence.
6. Add conservative refinement edges from the learned pair classifier on top of the anchored graph for Salamander and Horned Lizard.
7. Resolve connected components with union-find and write the final CSV.

Table 1 gives the cluster shape of the best submitted run. Singleton fractions make the hard species visible: 52.2 % of Salamander and 85.0 % of Horned Lizard clusters are unmerged single images, against one of 36 for Lynx.

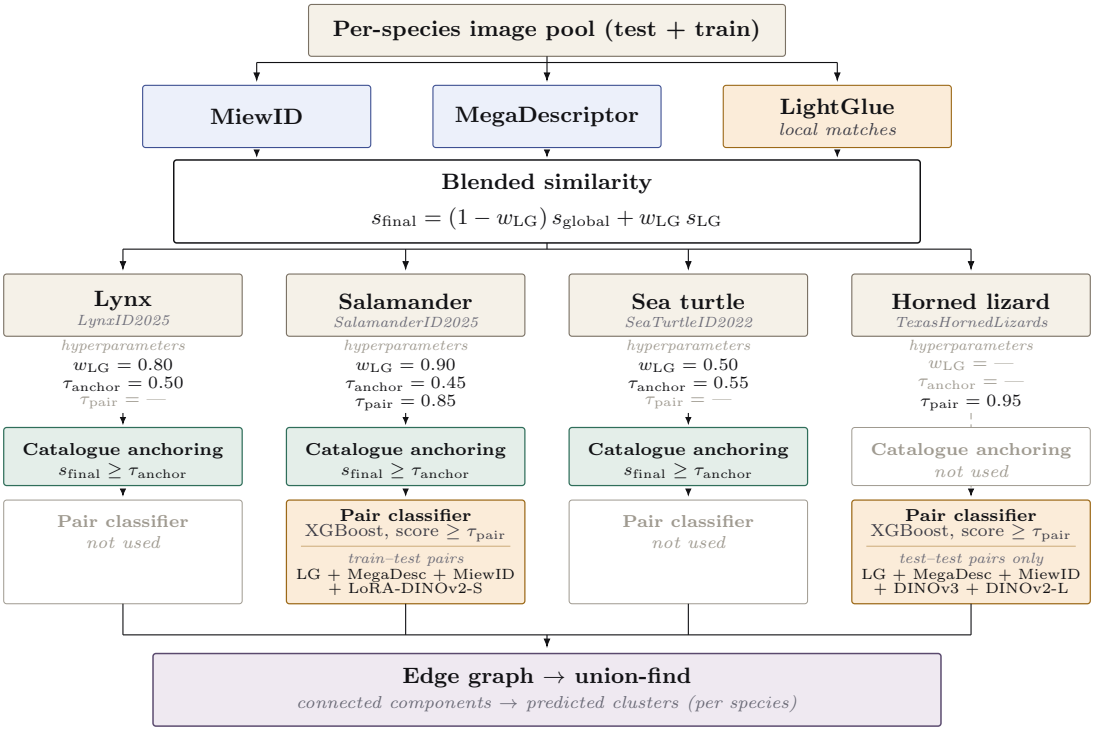

Figure 1: System overview. Per-species image pools are encoded with two pretrained global descriptors (MiewID, MegaDescriptor) and matched locally with LightGlue. The blended similarity s_{final} drives both catalogue anchoring (when a test image's nearest training neighbour passes the species threshold τ_{anchor}) and the inputs to an XGBoost pair classifier that supplies refinement edges when its score passes τ_{pair}. Connected components of the resulting per-species graph become the predicted clusters.

Table 1
Cluster shape of the selected final submission. Singletons are clusters of size 1; multi-image clusters are size ≥ 2; max is the largest cluster's size.

Species	Rows	Clusters	Singletons	Multi-image	Max
LynxID2025	946	36	1 (2.8%)	35	156
SalamanderID2025	689	402	210 (52.2%)	192	25
SeaTurtleID2022	500	169	48 (28.4%)	121	11
TexasHornedLizards	274	226	192 (85.0%)	34	7
Total	2409	833	451 (54.1%)	382	—

5. Visual Similarity Signals

5.1. Global descriptors

MiewID [6] and MegaDescriptor [5] produce L_2-normalised vector embeddings per image. For two images, cosine similarity between their vectors gives a whole-image identity score. The submitted system uses a fixed weighted blend

$$s_{\text{global}} = 0.7\, s_{\text{MiewID}} + 0.3\, s_{\text{MegaDescriptor}},$$

tuned on local cross-validation. The global signal is fast enough to score every train-test pair across all species, and serves as the prefilter that decides which pairs are scored with LightGlue.

5.2. LightGlue local matching

LightGlue [11] matches deep keypoints between two images and returns a count of accepted matches after geometric verification. Counts are normalised per species against the empirical distribution of match counts on labelled positive and negative pairs to produce a similarity score $s_{\mathrm{LG}} \in [0, 1]$. The blended similarity used in catalogue anchoring is

$$s_{\mathrm{final}} = (1 - w_{\mathrm{LG}})\, s_{\mathrm{global}} + w_{\mathrm{LG}}\, s_{\mathrm{LG}},$$

with the per-species LightGlue weight w_{LG} and anchoring threshold τ_{anchor} given in Table 2.

5.3. Learned pair scores

Salamander and Horned Lizard each use a separate gradient-boosted tree model (XGBoost [9]) trained on tabular features per image pair. Both classifiers share a common base of features: raw, rank-normalised, and percentile-ranked LightGlue counts, MegaDescriptor and MiewID cosine similarities, and a co-clustering prior tuned on the public leaderboard. The species-specific features differ in what global descriptor evidence they add on top of this base: the Salamander classifier adds a LoRA-fine-tuned [10] DINOv2-S [13] cosine that targets dorsal spot identity, while the Horned Lizard classifier adds DINOv3 [14] (CLS, patch-max, patch-mean, attention-pooled) and DINOv2-L cosines, which are more useful when the local-matching signal is weak. Each classifier is trained on labelled pairs derived from its species' train split (positives are same-individual within species, negatives are different-individual within species) and applied to test-test and train-test pairs at inference time. We use the classifier only as a refinement gate: only pairs whose classifier score exceeds the species-specific gate τ_{pair} contribute edges to the graph.

5.4. Hyperparameters

Table 2 collects the per-species hyperparameters of the submitted system. The LightGlue weight w_{LG} controls how aggressively local matching dominates the blended similarity; τ_{anchor} is the catalogue-anchor threshold; τ_{pair} gates the pair classifier output. Where τ_{pair} is not listed, the species path does not use the pair classifier as a refinement gate.

Table 2
Per-species hyperparameters of the submitted system.

Species	w_{LG}	τ_{anchor}	τ_{pair}
LynxID2025	0.80	0.50	—
SalamanderID2025	0.90	0.45	0.85
SeaTurtleID2022	0.50	0.55	—
TexasHornedLizards	—	—	0.95

6. Species-Specific Pipelines

6.1. LynxID2025

Lynx contributes the largest test set (946 images). Coat patterns are stable across visits, so the global descriptor blend already separates individuals well; LightGlue mainly helps when the global descriptor cannot distinguish two visually similar individuals. With $w_{\mathrm{LG}} = 0.80$ and $\tau_{\mathrm{anchor}} = 0.50$, the submitted clustering anchors most test images to a small set of training identities (Table 1). Single-species experiments, including triplet fine-tuning of a DINOv2 [13] backbone on the Lynx train split, did not improve leaderboard ARI, so we kept the descriptor-based path that the leaderboard had already validated.

6.2. SalamanderID2025

Salamander was the hardest species. Dorsal spot patterns are highly individual, but image-to-image alignment is poor: the same individual can appear at very different zoom levels, with different parts of the body in frame, against different backgrounds. Local matching produces many candidate edges; many visually plausible ones are false positives near the clustering boundary.

The submitted Salamander path uses a LightGlue-heavy blend ($w_{\mathrm{LG}} = 0.90$, $\tau_{\mathrm{anchor}} = 0.45$) for catalogue anchoring, then adds Salamander test-test merges from the learned pair classifier gated by τ_{sal}. The submitted setting $\tau_{\mathrm{sal}} = 0.85$ was selected by local cross-validation against train-as-test holdout merges. Lower values added too many unsafe edges; higher values missed useful merges.

6.3. SeaTurtleID2022

Sea turtle scute patterns are well-captured by the global descriptor blend, so this species uses a balanced $w_{\mathrm{LG}} = 0.50$ with $\tau_{\mathrm{anchor}} = 0.55$ and no learned-pair refinement gate. Early leaderboard runs showed that broad Sea Turtle changes which improved local pair metrics often reduced overall ARI, so the species path was kept conservative for the rest of the competition.

6.4. TexasHornedLizards

The Horned Lizard split has limited train-test overlap: many test images depict individuals with no clear nearest training neighbour. The submitted Horned Lizard path therefore relies on test-test pair evidence. The learned pair classifier is applied to all test-test pairs and edges are kept only when the score exceeds the high gate $\tau_{\mathrm{hl}} = 0.95$, leaving an intentionally conservative clustering with many singleton components (Table 1). A hierarchical agglomerative clustering (HAC) variant on the same pair-score matrix scored 0.57706 on the public leaderboard (Table 3) and was rejected in favour of this graph-merging baseline.

7. Catalogue Anchoring

Catalogue anchoring connects the clustering graph to the labelled reference set. For a test image x, the system searches train images of the same species, computes s_{final} against each, and selects the best train neighbour. If s_{final} exceeds the species threshold τ_{anchor}, x inherits that train individual's identity. Two test images that inherit the same train identity are automatically merged.

This changes the graph problem in two ways (Figure 2). First, a test-test-only method has to prove that two test images match each other directly; anchoring lets the system route them through a shared reference identity instead, which is often easier when poses, crops, or lighting differ but both images match the same known animal. Second, the anchor labels are deterministic across submissions, so iterative refinement can preserve the anchored backbone and only experiment with test-test merges on top of it.

The follow-up rows in Table 3 make this concrete: every variant that added or rescored test-test edges without preserving the anchored backbone scored below the anchored run on the public leaderboard.

8. Failed Experiments

We document four directions that improved local pair-level diagnostics without translating into leaderboard ARI gains, to inform future submissions.

Head-crop matching for Salamander. We trained a Salamander head detector and ran LightGlue matching only inside the detected head crop, hypothesising that head spots are more discriminative

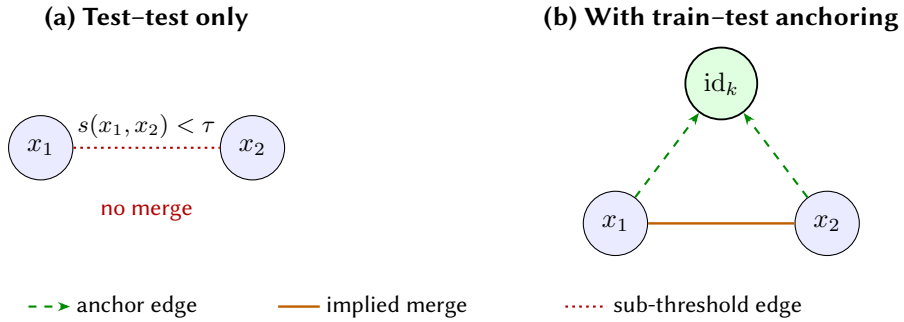

Figure 2: Train–test anchoring. (a) A direct test–test method must score the pair (x_1, x_2) above the species threshold; when poses, crops, or lighting differ, the direct score $s(x_1, x_2)$ often falls short. (b) If both x_1 and x_2 separately match a known training individual id_k above τ_{anchor}, they inherit its label and are merged automatically, recovering pairs that direct test–test scoring misses.

than full-dorsal spots. Local pair AUC on labelled head-crop pairs was high, but the resulting head-crop similarity matrix did not change the public leaderboard score: the head-crop edges were largely redundant with edges already supplied by full-image LightGlue.

Anatomy-based shape features for Salamander. We computed dorsal-spine curvature and body-silhouette descriptors on segmented Salamander masks. These features had non-trivial pair AUC but contributed almost no new high-confidence edges to the graph.

Self-supervised Salamander backbone. We fine-tuned a DINOv2 [13] backbone with self-supervised pretraining on the Salamander train split, expecting more discriminative embeddings than MiewID for this species. Train-train pair AUC improved, but the resulting embeddings did not produce safe new catalogue-anchor merges at any threshold tested.

Test-test Salamander redesign. A late-stage redesign rescored Salamander test-test edges from scratch and dropped the catalogue anchor. It scored 0.59207 on the public leaderboard, well below the anchored 0.61741.

All four experiments improved pair-level diagnostics (AUC, top-1% precision, hard-edge accuracy) without changing leaderboard ARI. Pair-level gains only translate into ARI if the new edges survive the operating threshold the merge graph actually uses and add coverage on top of the anchored backbone instead of replacing it.

9. Submitted Runs

Table 3 lists the main leaderboard submissions, identified by the SHA-256 prefix (8 hex characters) of the submitted CSV. Private-leaderboard scores are reported only for the selected final submission, because Kaggle reveals private scores only for entries the team selects as final.

Full SHA-256 hashes for the submitted CSVs:

```
3CC406B7C75A82C8122395875A4040AD7D0B9265F70514F522EE92169067F1D4
58F410BD9C0E374D2A57FDAC1B6A3D95006EBF2746BF01B192278CE3B5033A3A
C772B87C7989B0C23D360EC895934236ED48D1D9C055C2DE0C010833E6598DF6
69C8754D983C43FDCA811F792D3F68B58875D93E20FA1335A8669B1964FEA92E
DF2ADD2C16FE48327CE76A3677EA223654419784C6002B22B4E395D66A57AE8B
E4C31AB87E89A90027B26B607ABD7594D271153F665147C394592B926F56DEF9
C2048236EF0C929885F17D20C7D2503101D715D019DB45329945154301 21A6E1
```

Table 3
Leaderboard submissions, grouped by method. The "Anchored graph (final)" row was selected as the team's final submission and is the only row with a recorded private-LB score.

Method family	Description	Public LB	Private LB	SHA-8
LightGlue baseline	Species-specific local-matching and descriptor baseline	0.56792	—	3CC406B7
Pair-refined baseline	Baseline with learned pair refinements	0.59644	—	58F410BD
Anchored graph (final)	Catalogue anchoring with conservative Salamander refinement	0.61741	0.57038	C772B87C
Alternative pair-gate setting	Pair-gate variant with a different threshold	0.61259	—	69C8754D
Additive Salamander edge model	Extra Salamander test-test edges added to the anchored graph	0.60156	—	DF2ADD2C
Test-test Salamander redesign	Rebuilt Salamander edges without preserving the catalogue anchor	0.59207	—	E4C31AB8
Horned Lizard HAC probe	Agglomerative-clustering probe for Horned Lizard	0.57706	—	C2048236

9.1. Public and Private Scores

The selected submission scored 0.61741 ARI on the public split and 0.57038 ARI on the private split, a difference of 0.046. The per-species thresholds in Table 2 were chosen by maximising public-LB ARI directly: each entry in Table 3 corresponds to a public-LB probe used to update one or more thresholds. Tuning the operating point against the public split is itself a form of overfitting to that split, and the 0.046 drop is the gap between the two ARI estimates this procedure produces. We have no separate held-out estimate of generalisation error.

10. Limitations

The submitted system has four known limitations. The catalogue anchor relies on test images of training individuals existing in the test set; for species or splits where this overlap is small (Horned Lizard in our package), the anchor signal is weak and the system falls back to less reliable test-test pair evidence. The per-species thresholds were tuned on a combination of public-leaderboard feedback and local train-as-test cross-validation; they may not transfer to the private leaderboard or to a different release. The learned pair classifier operates on descriptor cosine similarities and LightGlue match counts only; it has no access to image content beyond these summaries, which limits its ability to disambiguate individuals when both signals are similar. Finally, our evidence that catalogue anchoring is the load-bearing component is observational: we compare submitted variants in Table 3 that differ in more than the anchor (similarity blend, refinement gating, threshold), rather than running a controlled anchor-on/off ablation that holds all other components fixed. Such an ablation would isolate the anchor's contribution more cleanly and is left to future work.

11. Code and Data Availability

The released AnimalCLEF 2026 dataset, sample submission, and leaderboard are available through the official Kaggle competition page [1]. The submitted system has the following trained components. (i) MegaDescriptor [5] (`BVRA/MegaDescriptor-L-384`) was fine-tuned with ArcFace [15] heads on the AnimalCLEF 2026 train split for Lynx and Sea Turtle on Apple M4 Pro (Metal); these checkpoints supply the species-specific MegaDescriptor embeddings used in the similarity step. (ii) MiewID [6] (`conservationxlabs/miewid-msv3`) was used unchanged as a pretrained encoder; the `cos_miew` feature seen by the pair classifiers comes from the upstream MiewID embeddings. (iii) For the Salamander branch, a LoRA adapter (rank 8, $\alpha = 16$, on query and value projections) was trained on DINOv2-S [13]

over a hand-labelled Salamander spot-correspondence dataset of approximately 7 800 colour-coded keypoint correspondences across image pairs of the same individual (3 217 train–train points and 4 580 verified test–test/test–train points, collected over two rounds of manual annotation). The objective is a CLIP-style symmetric InfoNCE loss that pulls together 128×128 patch crops at matching keypoints and pushes apart patches at different keypoints on the same image pair, so that the encoder discriminates spot-level identity rather than generic "yellow-blob" texture. The resulting pair-similarity matrix is blended with LightGlue in the Salamander clustering step, and the LoRA-DINOv2-S cosine is one of the input features to the Salamander pair classifier. (iv) Two XGBoost pair classifiers (one for Salamander, one for Horned Lizard) were trained leave-image-out ($K=10$) on hand-confirmed labelled pairs – pooled test-test and test-train for Salamander, and test-test only for Horned Lizard – and score every test-test pair at inference time; pairs scoring at or above the species τ_{pair} contribute edges to the clustering graph. (v) LightGlue [11] with the SuperPoint [16] front-end was used unchanged. Submitted CSVs are uniquely identified by the SHA-256 hashes listed in Section 9.

Pipeline source code, per-species YAML configs encoding the hyperparameters of Table 2, the submission SHA-256 manifest, and a verifier script are released at https://github.com/gcol33/animal-clef-2026 under an MIT licence. The trained weights for the leaderboard run – two MegaDescriptor fine-tunes (Lynx, Sea Turtle), the LoRA-DINOv2-S adapter and its spot detector for Salamander, and the two XGBoost pair classifiers (Salamander and Horned Lizard) – are released separately on the Hugging Face Hub at https://huggingface.co/gcol33/animal-clef-2026 under CC-BY-4.0; the `MANIFEST.md` in that repository documents each file's role and source, and the GitHub README pins the commit revision used for the submitted runs. Image data is not redistributed and must be obtained from the Kaggle competition page.

Licences and dataset acknowledgments. Our code release is MIT-licensed and our trained weights are released under CC-BY-4.0, subject to the upstream restrictions noted here. The Sea Turtle MegaDescriptor fine-tune is a derivative of training on the AnimalCLEF 2026 Sea Turtle split, which is sourced from SeaTurtleID2022 [3]; SeaTurtleID2022 is released for non-commercial research use only and may not be re-uploaded, so the released Sea Turtle weights may not be used for commercial purposes. Backbones used unchanged or as fine-tuning starting points (`BVRA/MegaDescriptor-L-384`, `conservationxlabs/miewid-msv3`, `facebook/dinov2-small`, and the DINOv3 checkpoints used in the Horned Lizard branch) retain their upstream licences. The AnimalCLEF 2026 train/test split itself is governed by the Kaggle competition terms [1]. No raw images, JPEG bytes, or dataset paths from any upstream source are redistributed in either the GitHub repository or its Hugging Face mirror.

12. Conclusion

Our best AnimalCLEF 2026 submission scored 0.61741 ARI on the public leaderboard and 0.57038 ARI on the private leaderboard, with a per-species graph-clustering pipeline. Pretrained MiewID and MegaDescriptor descriptors supplied broad identity similarity, LightGlue local matching supplied complementary pair evidence, and species-specific thresholds controlled when each signal contributed graph edges. The largest single design difference between the best run and the alternatives we submitted was catalogue anchoring: routing test images through known training identities, where they exist, produces stable cluster bridges that test-test-only redesigns did not recover.

Among the alternatives we submitted, catalogue anchoring was the most reliable foundation, and we suggest future iterations build on top of it rather than replace it; new neural or anatomy-based signals earn their place by adding safe edges to the full all-species graph, evaluated at the clustering level. Pair-level diagnostics (AUC, top-1% precision, hard-edge accuracy) were useful for debugging in this work; the four runs in Section 8 show that strong pair-level numbers can leave leaderboard ARI unchanged.

Acknowledgments

We thank the AnimalCLEF organisers for releasing the data and running the leaderboard, and the maintainers of MiewID, MegaDescriptor, and LightGlue for making the pretrained models openly available.

Declaration on Generative AI

During the development of the system, the author used a local REAP-48B model for coding assistance (typing speedup and routine code transformations); REAP-48B is a 48-billion-parameter sparse mixture-of-experts checkpoint derived from Qwen3-Next-80B [18] via the REAP expert-pruning method [17], served on an Apple M4 Pro. No proprietary or remote LLM was used. The machine-learning design and the prose of this manuscript are the author's own work; the author takes full responsibility for the publication's content.

References

[1] L. Adam, L. Picek, K. Papafitsoros, D. Williams, and D. Biffi. AnimalCLEF26 @ CVPR & CLEF. Kaggle, 2026. https://www.kaggle.com/competitions/animal-clef-2026

[2] ImageCLEF. AnimalCLEF2026: Discovery and Re-Identification of Individual Animals. https://www.imageclef.org/AnimalCLEF2026

[3] L. Adam, V. Čermák, K. Papafitsoros, and L. Picek. SeaTurtleID2022: A long-span dataset for reliable sea turtle re-identification. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7131–7141, 2024. https://doi.org/10.1109/WACV57701.2024.00699

[4] D. Biffi, M. R. Tucker, A. Ackel, and D. A. Williams. Identification of individual Texas Horned Lizards (*Phrynosoma cornutum*) using genotypes and ventral spot patterns. *Ecology and Evolution*, 15(3):e71167, 2025. https://doi.org/10.1002/ece3.71167

[5] V. Čermák, L. Picek, L. Adam, and K. Papafitsoros. WildlifeDatasets: An open-source toolkit for animal re-identification. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5941–5951, 2024. https://doi.org/10.1109/WACV57701.2024.00585

[6] L. Otarashvili, T. Subramanian, J. Holmberg, J. J. Levenson, and C. V. Stewart. Multispecies animal re-ID using a large community-curated dataset. *arXiv preprint arXiv:2412.05602*, 2024. https://doi.org/10.48550/arXiv.2412.05602

[7] L. Adam, V. Čermák, K. Papafitsoros, and L. Picek. WildlifeReID-10k: Wildlife re-identification dataset with 10k individual animals. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2124–2134, 2025. https://arxiv.org/abs/2406.09211

[8] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985. https://doi.org/10.1007/BF01908075

[9] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. https://doi.org/10.1145/2939672.2939785

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[11] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys. LightGlue: Local feature matching at light speed. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 17581–17592, 2023. https://doi.org/10.1109/ICCV51070.2023.01616

[12] scikit-learn developers. `sklearn.metrics.adjusted_rand_score`. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

[13] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[14] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. https://doi.org/10.48550/arXiv.2508.10104

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. https://doi.org/10.1109/CVPR.2019.00482

[16] D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. https://doi.org/10.1109/CVPRW.2018.00060

[17] M. Lasby, I. Lazarevich, N. Sinnadurai, S. Lie, Y. Ioannou, and V. Thangarasa. REAP the experts: Why pruning prevails for one-shot MoE compression. *arXiv preprint arXiv:2510.13999*, 2025. https://doi.org/10.48550/arXiv.2510.13999

[18] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. https://doi.org/10.48550/arXiv.2505.09388