

A Prototype-Constrained Discovery Clustering System for AnimalCLEF 2026

Weiye Chen^{1,*†}, Dongrun Cao^{1,†}

¹Beihang University, Beijing, People's Republic of China

Abstract

This paper describes the 149_CV4Animal system for the AnimalCLEF 2026 individual animal re-identification and discovery task hosted on Kaggle. The task requires assigning every test image to an individual-level cluster. Unlike closed-set classification, the system must simultaneously recognize known individuals when a gallery is available and discover unseen individuals that appear only in the competition test data. The competition metric is Adjusted Rand Index (ARI), so the final objective is the quality of the global clustering structure rather than the accuracy of isolated nearest-neighbor matches. Our solution is built around a prototype-constrained discovery clustering pipeline. During training, known identities are represented by momentum-updated prototypes and optimized with a prototype contrastive objective. During inference, gallery prototypes and test images are placed into a shared similarity graph. The clustering process is constrained by cannot-link rules between different known identities, while samples without a compatible prototype are allowed to form novel discovery clusters. This paper focuses on the first two system formulations: TSC, a two-stage lookup-then-discovery cascade, and PCDC, the prototype-constrained clustering system. Later LGC/GQE heuristic refinements explored local matching, pair graphs, query expansion, and additional routing; these extensions are analyzed as part of the system development because they reveal why public-only tuning can hurt cluster-level generalization. We report two levels of evidence. First, our ProtoNCE/PCDC submission submission0314_01.csv obtained 0.29076 public ARI and 0.32758 private ARI, with a final rank of 58/230. Second, the full PCDC experiments for LynxID2025, SalamanderID2025, and SeaTurtleID2022 produced masked tuning ARI values of 0.7244, 0.8174, and 0.8815 respectively. These are internal proxy scores from a masked-prototype simulation, not public or private leaderboard ARI. We further analyze threshold sensitivity and submission-level cluster structure to illustrate how the proxy objective can affect the final partition. A short 10-epoch Lynx TSC/PCDC run was used only as a development diagnostic and is not treated as final performance evidence. Subsequent public-only tuning variants are discussed qualitatively as possible public-leaderboard overfitting cases rather than treated as robust improvements.

Keywords

Animal re-identification, prototype clustering, AnimalCLEF 2026

1. Introduction

This working note describes our participation in the AnimalCLEF 2026 task [1, 2]. Individual animal re-identification supports ecological monitoring by linking repeated observations of the same animal across time, location, and camera conditions. In real deployments, the identity catalog is incomplete. A system must therefore solve two problems at once: known-individual lookup and unknown-individual discovery. If a test image belongs to an animal already present in the training identity database, it should be assigned to that known identity cluster. If it shows a new animal, it should be grouped with other images of the same new individual.

AnimalCLEF 2026 evaluates this setting directly as a Kaggle clustering task. The submission is a two-column file containing `image_id` and `cluster`. The leaderboard score is ARI [3], which penalizes both splitting one individual across multiple clusters and merging different individuals into the same cluster. This changes the system design compared with a standard retrieval or classification problem. A method that improves top-k retrieval may still reduce ARI if it introduces false bridge edges between clusters.

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

*Corresponding author.

†These authors contributed equally.

✉ weiyichen200410@163.com (W. Chen); gugugua0@gmail.com (D. Cao)



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

Table 1

Dataset statistics and task setting.

Dataset	Train images	Train IDs	Test images	Setting
LynxID2025	2957	77	946	Lookup + discovery
SalamanderID2025	1388	587	689	Lookup + discovery
SeaTurtleID2022	8729	438	500	Lookup + discovery
TexasHornedLizards	0	0	274	Pure discovery

Our design emphasis is therefore on the graph structure used for clustering. We use supervised identities to learn an embedding space, but we avoid committing to a closed-set classifier at inference time. Instead, known identities become constraints inside a discovery graph. This design is intended to support both species with training identities and the pure-discovery case where no training identities are available.

The contributions of this work are:

- a unified Prototype-Constrained Discovery Clustering (PCDC) formulation for mixed lookup and discovery;
- a masked-validation strategy that simulates unseen identities while tuning clustering thresholds;
- a reproducible full PCDC training and threshold-selection summary for the three supervised species;
- an analysis of LGC/GQE heuristic refinements and subsequent public-only tuning variants, explaining why they may overfit public feedback despite plausible local improvements.

Following the structure commonly used in AnimalCLEF working notes, Section 2 describes the dataset and evaluation protocol, Section 3 presents the methodology, Section 4 details the experimental setup, Section 5 reports the results, and Section 6 discusses failure modes before the conclusion.

2. Task, Dataset, and Evaluation Protocol

The competition dataset [2, 4] aggregates wildlife re-identification images and related toolkit infrastructure from several sources, including CzechLynx [5], SeaTurtleID2022 [6], Texas Horned Lizard data [7], WildlifeReID-10k [8], and WildlifeDatasets [9]. We cite these resources to document the data provenance and task infrastructure used by the benchmark.

The dataset contains 15,483 images in total: 13,074 training images and 2,409 test images. Three species have labeled training identities. The LynxID2025 subset represents the Eurasian lynx identification part of the task and is derived from CzechLynx [5]. TexasHornedLizards has no training identities and is treated as pure discovery.

The predicted cluster labels use species prefixes, for example `cluster_LynxID2025_0`. The competition metric is ARI, which evaluates agreement between the predicted clusters and the private ground-truth partitions over the evaluated test images. Let n_{ij} be the number of samples that are in predicted cluster i and ground-truth cluster j , $a_i = \sum_j n_{ij}$, $b_j = \sum_i n_{ij}$, and n be the number of evaluated samples. ARI is

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}. \quad (1)$$

ARI is sensitive to both over-splitting and false merging, so we also monitor structural diagnostics during development: number of clusters, singleton ratio, largest-cluster ratio, and bridge-related graph statistics.

Method Evolution Emphasized in This Paper

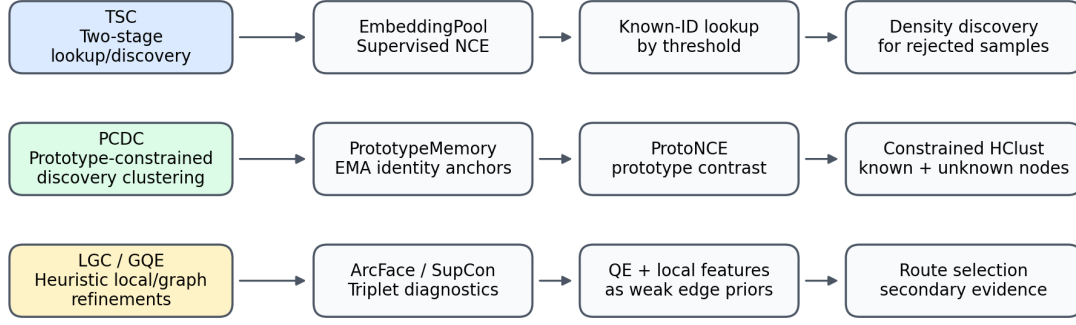


Figure 1: Method evolution emphasized in this paper. TSC is the two-stage baseline, PCDC is the primary reported system, and LGC/GQE summarizes local and graph refinements analyzed in the discussion.

3. Methodology

3.1. Pipeline Overview

The PCDC system is organized into four stages:

1. feature learning from labeled identities where available;
2. prototype memory construction for known identities;
3. prototype-constrained clustering of test samples and known prototypes;
4. conservative threshold selection under a masked open-set validation proxy.

The central idea is to avoid a hard separation between lookup and discovery. Known identities are represented by prototypes, but a test sample is not forced into the closest known identity if the graph structure does not support that merge. Unknown individuals can form clusters without any known prototype. Figure 1 summarizes the progression from TSC to PCDC and then to the later LGC/GQE probes.

3.2. Feature Encoder

The encoder maps an image x to a normalized embedding $f(x)$. PCDC uses two backbone families. Lynx uses the Hugging Face `conservationlabs/miwid-msv3` encoder [10] with a lazy linear projection to a 512-dimensional embedding and 512-pixel input. Salamander and sea turtle use `BVRA/MegaDescriptor-L-384` [11] from `timm` [12], followed by a linear projection to a 256-dimensional embedding and 384-pixel input. All embeddings are L2-normalized before prototype updates and similarity computation.

The full PCDC runs use Adam optimization, batch size 16, temperature 0.1, prototype EMA momentum 0.95, and a 10-epoch backbone freeze before unfreezing. Learning rate is selected by a species-specific grid and checkpoint selection uses the mean ARI across three masked-validation eps values.

3.3. Prototype Memory and ProtoNCE

The use of identity prototypes is related to prototypical contrastive learning [13] and momentum-updated cluster prototypes used in re-identification [14], although our prototypes are supervised identity anchors rather than unsupervised cluster assignments.

For each known identity y , the system maintains a normalized prototype p_y . Here, μ is the EMA momentum, B_y is the mini-batch subset whose labels equal y , and f_i is the embedding of sample i . During training, prototypes are updated by exponential moving average over batch features:

$$\mathbf{p}_y \leftarrow \text{norm} \left(\mu \mathbf{p}_y + (1 - \mu) \frac{1}{|B_y|} \sum_{i \in B_y} \mathbf{f}_i \right). \quad (2)$$

The primary training objective is ProtoNCE. For a sample embedding $\mathbf{f}_i$ with label y_i , the loss contrasts it against all known identity prototypes. In this equation, B is the batch size, C is the number of known identities, and τ is the temperature:

$$\mathcal{L}_{\text{ProtoNCE}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\mathbf{f}_i^\top \mathbf{p}_{y_i} / \tau)}{\sum_{c=1}^C \exp(\mathbf{f}_i^\top \mathbf{p}_c / \tau)}. \quad (3)$$

This avoids relying on positive pairs inside every mini-batch, which is important for long-tailed identities and single-image identities.

3.4. Unified Prototype-Constrained Clustering

At inference time, we build one similarity graph containing both known prototypes and test sample embeddings. Edges are sorted by decreasing similarity. A union operation is allowed unless the two components already contain different known identity prototypes. This creates a cannot-link constraint between different known identities while preserving the ability to discover new clusters.

The procedure is summarized below.

1. Build nodes for all known prototypes and all test samples.
2. Compute cosine similarities between nodes.
3. Sort candidate edges from high to low similarity.
4. For each edge, inspect the known prototype IDs already present in the two components.
5. Skip the edge if merging would combine two different known identities.
6. Otherwise merge the components.
7. Assign final clusters. Components with no known prototype become novel discovery clusters.

3.5. Methodological Evolution and Heuristic Extensions

Two-Stage Cascade Baseline (TSC). TSC uses a sequential lookup-and-discovery design. It first compares each test embedding with known identity prototypes. Samples above a lookup threshold are assigned to known identities; the remaining samples are clustered with DBSCAN [15] or HDBSCAN [16]. This design is simple and interpretable, but the hard lookup decision can remove samples from discovery too early, and the density-based stage cannot revise earlier known-identity assignments.

Prototype-Constrained Discovery Clustering (PCDC). PCDC replaces the hard cascade with a unified graph. Known prototypes and test samples are clustered together under cannot-link constraints. This lets the same graph decide whether a sample should attach to a known identity or remain in a new discovery component. PCDC is therefore the primary reported system in this paper.

Heuristic Extensions: LGC and GQE. We further explored additional structural signals split into two modules. Local-Global Edge Calibration (LGC) captures local feature matching and pairwise edge evidence. Graph-based Query Expansion (GQE) introduces species-specific routing and query expansion to refine connected components. These ideas are plausible for animal re-identification because many species are identified through local markings rather than object-level semantics. However, they also increase the number of thresholds, graph gates, and component-level decisions. Section 6 discusses why these changes did not provide sufficiently clear evidence of overall improvement in our development runs.

Prototype-Constrained Clustering

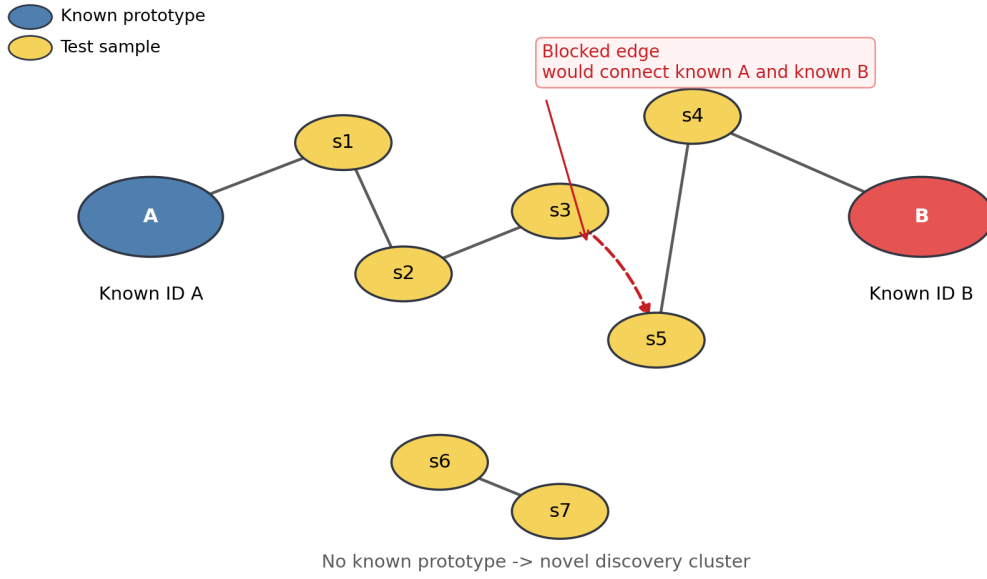


Figure 2: Prototype-constrained clustering. Edges that would merge two different known-identity prototype components are blocked, while components without known prototypes remain valid novel clusters.

3.6. Candidate Selection and Public-Feedback Control

Because repeated public leaderboard feedback may overfit test-specific structure, candidate selection is separated into two categories:

1. reproducible candidates selected by offline or masked validation;
2. public-aware probes generated by local test submission rewrites.

We use the first category for quantitative reporting. The second category is discussed as an overfitting risk because repeated local rewrites can adapt to the public leaderboard split without demonstrating improvement on the private leaderboard split.

4. Experimental Setup

4.1. Validation Protocol

The diagnostic Lynx split uses the LynxID2025 training data. The training code builds an open-set validation proxy as follows. A fixed random seed selects a fraction of identity labels as pseudo-unknown validation identities. All images of those identities are removed from training and used as `val_unknown`. For remaining closed-set identities with at least two images, one image is held out as `val_known`; the rest are used for training. Singleton closed-set identities remain in training. The diagnostic run used `val_frac = 0.15`.

For PCDC final threshold tuning, 20% of known prototypes are additionally masked to simulate discovery during test-like tuning. This masked prototype step is used to select an `eps` value under a mixed known/unknown condition.

4.2. Training Details

The full PCDC experiments use the species-specific training settings shown in Table 2. All three species use batch size 16, Adam [17], ProtoNCE temperature 0.1, prototype EMA momentum 0.95, and a

Table 2
Full PCDC species-specific training configuration.

Species	Backbone	Input	Emb. dim	Epoch grid	LR grid	QE candidates
LynxID2025	MiewID-MSv3	512	512	80	1e-4, 5e-4, 1e-3	0, 2, 3
SalamanderID2025	MegaDescriptor-L-384	384	256	80	1e-5, 5e-5, 1e-4	0, 2, 3, 5
SeaTurtleID2022	MegaDescriptor-L-384	384	256	40	1e-5, 5e-5, 1e-4	0

Table 3
Under-trained Lynx TSC/PCDC diagnostic configuration.

Parameter	Value
Dataset	LynxID2025
GPU	NVIDIA RTX 3090
Backbone route	MiewID-MSv3 wrapper
Input resolution	512
Embedding dimension	512
Epoch budget	10 epochs plus diagnostic retraining
Batch size	8
Optimizer	Adam
Learning rate	1e-4
Weight decay	1e-4
Backbone freeze	1 epoch, then unfreeze
Query expansion	disabled, qe_k = 0
TSC lookup thresholds	0.70, 0.75, 0.80, 0.85
TSC/PCDC eps grid	0.15, 0.20, 0.25
PCDC prototype EMA	0.95

10-epoch backbone freeze. The best checkpoint is selected by `ValGapMean@3eps` (ARI) on the masked validation split, then retrained on the full training set for the selected epoch count. The implementation was developed with PyTorch and `timm`; the main hardware used for the controlled Lynx diagnostic was an NVIDIA RTX 3090. Exact training and inference wall-clock times were not fully logged for every run, so we do not report incomplete timing numbers here.

During development, we also ran a controlled Lynx TSC/PCDC diagnostic under a shared 10-epoch budget. Because this budget is much smaller than the final PCDC schedule, it is used only to inspect formulation behavior and is not used as main evidence in the results section. The configuration is summarized in Table 3.

4.3. Inference and Submission Generation

The full PCDC inference uses prototype-constrained clustering for species with training identities. TexasHornedLizards has no labeled training identities, so it is handled by a conservative pure-discovery route outside the supervised prototype table.

5. Experimental Results

This section reports the full PCDC experiments for the three species with labeled training identities and relates them to public baseline references. The private-set labels are not released to participants, so the high ARI values in Table 6 are masked validation and masked tuning proxies only; they should not be interpreted as private-set ARI.

Table 4
Inference configuration.

Species / experiment	Route	Features and backend	Thresholds
LynxID2025 full PCDC	prototype-constrained	MiewID-MSv3 + PrototypeMemory; constrained hclust	eps 0.250, qe_k 0
SalamanderID2025 full PCDC	prototype-constrained	MegaDescriptor + PrototypeMemory; constrained hclust	eps 0.350, qe_k 0
SeaTurtleID2022 full PCDC	prototype-constrained	MegaDescriptor + PrototypeMemory; constrained hclust	eps 0.500, qe_k 0
TexasHornedLizards	pure discovery	no supervised prototypes; conservative clustering	species-specific
LynxID2025 diagnostic TSC	lookup + DBSCAN	MiewID-MSv3 wrapper; DBSCAN after rejects	lookup 0.70, eps 0.25
LynxID2025 diagnostic PCDC	prototype-constrained	MiewID-MSv3 + PrototypeMemory; constrained hclust	eps 0.25

Table 5
Official competition results and public baselines.

Team / system	Submitted run	Phase	Public ARI	Private ARI	Rank / teams	Notes
Organizer baseline	–	Public baseline	0.20646	–	–	MegaDescriptor + MiewID + DBSCAN
Organizer baseline	–	Public baseline	0.21221	–	–	WildFusion + HDBSCAN
149_CV4Animal	submission0314_01.csv	Final leaderboard	0.29076	0.32758	58 / 230	ProtoNCE/PCDC reference submission

Table 6
Full PCDC checkpoint selection, proxy ARI, and cluster structure.

Species	LR	Epoch	ValGap ARI	Masked ARI	eps	qe_k	Test	Clusters	Singleton
LynxID2025	1e-3	75	0.1596	0.7244	0.250	0	946	543	0.5222
SalamanderID2025	1e-4	70	0.1055	0.8174	0.350	0	689	601	0.8186
SeaTurtleID2022	1e-5	30	0.8140	0.8815	0.500	0	500	137	0.1500

5.1. Public Leaderboard and Official Baselines

The organizers reported two public-baseline references: MegaDescriptor + MiewID + DBSCAN at 0.20646 public ARI and WildFusion [18] + HDBSCAN [16] at 0.21221 public ARI. Our ProtoNCE/PCDC reference submission `submission0314_01.csv` obtained 0.29076 public ARI and 0.32758 private ARI with final rank 58 of 230 teams. This comparison is a public/private leaderboard reference, not a proxy score, and it is the main official result reported in this paper.

5.2. Full PCDC Proxy Evaluation

The full PCDC training selected one checkpoint per species, retrained that configuration on all training identities, and tuned the final constrained-clustering threshold by masking a fraction of known prototypes as pseudo-unknown identities.

These results characterize the supervised PCDC training and threshold-selection behavior. They do not imply that the method obtains 0.7–0.88 ARI on the public or private leaderboard. The official ProtoNCE/PCDC submission `submission0314_01.csv` is reported separately in Table 5.

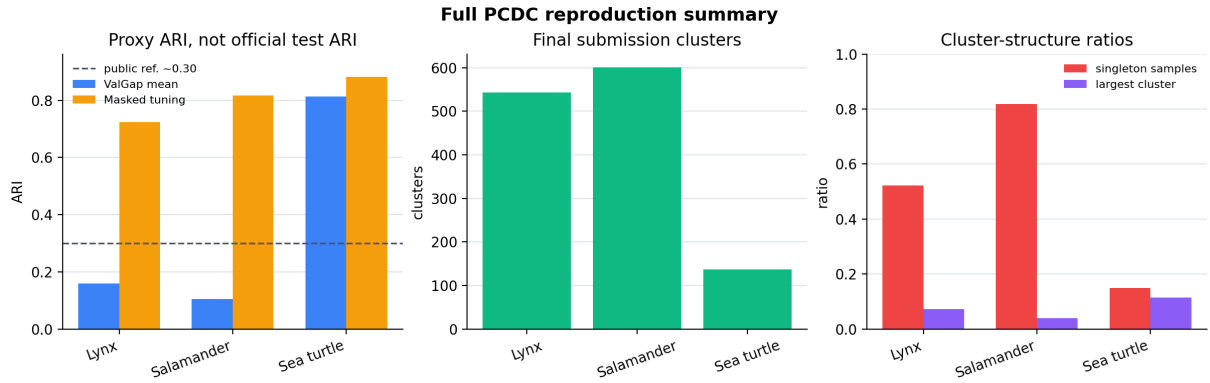


Figure 3: Full PCDC experiment summary. The left panel separates checkpoint-selection validation ARI from final masked tuning ARI. These values are proxy scores from known-label masking, not private-set ARI. The middle and right panels show the cluster structure of the generated species submissions.

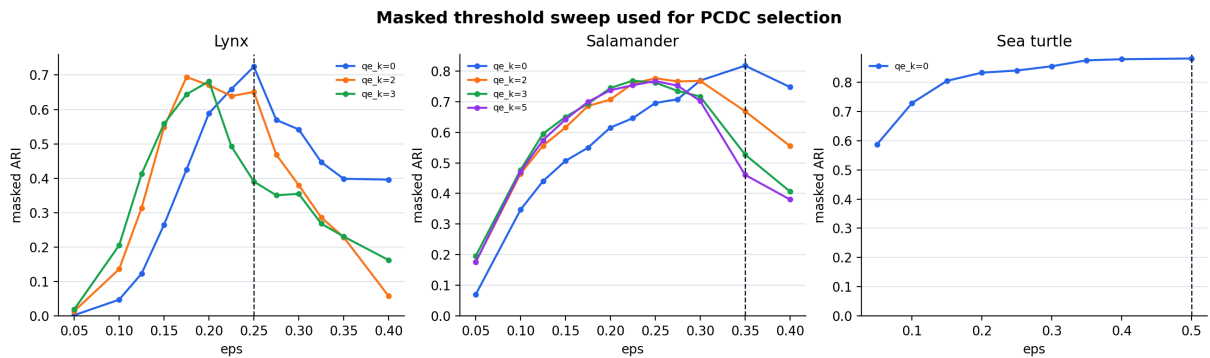


Figure 4: Full PCDC masked threshold sweep. The curves are internal masked-prototype proxy scores, not private-set ARI. Dashed vertical lines mark the selected eps values used for the final PCDC submissions.

5.3. Threshold Sensitivity and Cluster Structure

The masked tuning sweep is more directly tied to the reported PCDC submissions than the short Lynx diagnostic because it uses the full PCDC checkpoints. Figure 4 suggests that threshold choice has a substantial effect on the masked proxy: increasing eps first improves the masked proxy by allowing same-identity components to merge, but overly permissive thresholds can reduce ARI by creating false bridges. Lynx appears especially sensitive to this trade-off in the proxy setting. Salamander and sea turtle tolerate larger eps values in the masked proxy, but the generated submissions still have different cluster structures.

Figure 5 reports the resulting submission structure. The high singleton ratio for SalamanderID2025 is consistent with the many training identities and the conservative tendency of the constrained graph under fine-grained local texture. LynxID2025 also produces many small components, while SeaTurtleID2022 forms fewer and larger clusters. These diagnostics are useful because ARI can be affected either by over-splitting true individuals or by a small number of false merges that bridge separate individuals.

The controlled 10-epoch Lynx TSC/PCDC run is therefore excluded from the primary quantitative evidence. It was useful as a sanity check during implementation, but the fixed short budget appears to under-train both methods and produced near-zero test-like ARI. Reporting it as a main result would likely overstate what the experiment can support.

5.4. Online Feedback and Public-Overfitting Analysis

Table 7 separates internal proxy metrics, submission-structure diagnostics, and online feedback. For LGC/GQE and rewrite-pipeline variants, the project evidence is consistent with public-leaderboard

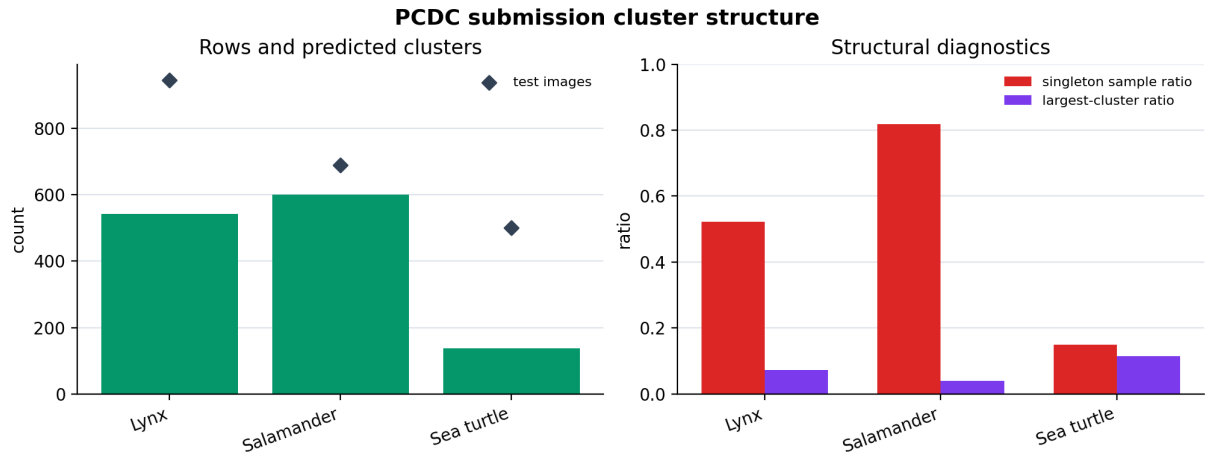


Figure 5: Submission-level cluster structure for the full PCDC runs. The plot reports row counts, predicted cluster counts, singleton ratios, and largest-cluster ratios.

Table 7

Interpretation of proxy metrics and online feedback.

Evidence	Use in this paper	Comment
Full PCDC masked tuning proxy	Threshold selection	0.7244 / 0.8174 / 0.8815
PCDC cluster structure	Structural diagnostic	singleton and largest-cluster ratios
ProtoNCE/PCDC reference submission	Official result	0.29076 public / 0.32758 private
LGC/GQE and public-only variants	Overfitting analysis	qualitative analysis

Table 8

Submission-level cluster structure used in the analysis. Singleton ratio is the fraction of test images assigned to singleton clusters.

Species	Rows	Clusters	Singleton ratio	Largest ratio	Route
LynxID2025	946	543	0.5222	0.0719	Prototype-constrained clustering
SalamanderID2025	689	601	0.8186	0.0392	Prototype-constrained clustering
SeaTurtleID2022	500	137	0.1500	0.1140	Prototype-constrained clustering
TexasHornedLizards	274	235	0.9021	0.0219	Pure-discovery route

overfitting risk and does not establish a clear overall gain, so they are analyzed qualitatively rather than used as quantitative leaderboard improvements.

5.5. Cluster Structure Diagnostics

The PCDC species submissions used for structural analysis are summarized in Table 8. The official ProtoNCE/PCDC result, 0.29076 public ARI and 0.32758 private ARI, is lower than the masked proxy values. This gap motivates treating the proxy as a selection tool rather than as a leaderboard claim.

6. Discussion

The experiments suggest that a key difficulty is not feature extraction alone but the conversion of similarities into a stable clustering graph. The full PCDC experiments provide the primary quantitative description of this process in our study, while the earlier 10-epoch Lynx diagnostic is too under-trained to support a final method comparison. TSC illustrates a possible limitation of a hard lookup-then-discovery cascade: early lookup decisions cannot be corrected by later clustering. PCDC is intended to reduce this limitation by allowing known identities and unknown test samples to compete inside one

constrained graph.

6.1. Heuristic Extensions: LGC and GQE

The LGC/GQE codebase extended PCDC in several directions. It added ArcFace-style classification heads [19], Triplet loss [20], supervised contrastive losses, query expansion, horizontal-flip test-time augmentation, local reranking, pairwise graph construction, and route selection. The later graph-oriented code added edge evidence, graph gating, large-component splitting, small-cluster rescue, micro-merging, and prototype-hybrid attachment. These changes were motivated by genuine species-specific issues: lynx require robust flank matching under pose changes, salamanders require caution with repeated local texture, and sea turtles benefit from stable global descriptors.

6.2. Overfitting Mechanisms

The same extensions also explain the overfitting risk. First, local matchers can reward visually similar texture patches without proving identity, which may create false bridge edges. Second, pair classifiers can obtain acceptable edge-level AUC while still being poorly calibrated at the component level; ARI penalizes a wrong merge much more severely than an isolated retrieval error. Third, query expansion, threshold sweeps, route selection, and component rewrites create many degrees of freedom. With repeated public feedback, these knobs can adapt to the public leaderboard split rather than the private leaderboard distribution. Fourth, public-only surrogate scoring uses a small number of observed submissions and is therefore vulnerable to fitting structural coincidences such as singleton ratio, largest-component size, or species-specific cluster counts.

Our interpretation is that LGC/GQE-style evidence is better treated as auxiliary edge evidence than as an automatic replacement for the PCDC graph. Local reranking, pair graphs, and component repair should be supported by cluster-level validation that directly measures merge and split behavior. In our experiments, these additions did not provide clear evidence of an overall gain over the original PCDC line, so the final report emphasizes TSC and PCDC while using LGC/GQE to explain what types of refinements were attempted and why they may overfit.

6.3. Limitations

Several limitations follow from the competition setting. TexasHornedLizards has no supervised training identities, so it requires a pure-discovery route rather than the supervised prototype mechanism used for the other species. The reported 0.7244, 0.8174, and 0.8815 ARI values are masked validation or masked tuning proxies rather than private-set ARI. The controlled TSC/PCDC diagnostic uses a limited 10-epoch budget and is excluded from the primary quantitative results because it is not representative of the full PCDC schedule. Finally, the LGC/GQE graph refinements need stronger cluster-level validation before they can be presented as robust improvements.

7. Conclusion

We present a prototype-constrained discovery clustering system for AnimalCLEF 2026. The system treats known identities as graph constraints and leaves room for unknown individuals to form new clusters. This formulation is designed for the ARI-based clustering objective rather than for closed-set classification alone. The official ProtoNCE/PCDC submission `submission0314_01.csv` obtained 0.29076 public ARI and 0.32758 private ARI at rank 58/230. The masked PCDC experiments produced high internal proxy scores, but they remain separate from leaderboard results. LGC/GQE explored local and graph-based refinements, but these attempts introduced additional public-overfitting risk and did not provide clear overall gains in our development evidence. We therefore describe the system mainly as a progression from TSC to PCDC, with LGC/GQE analyzed as an informative but unstable extension. Code is available at https://github.com/szjcw/149_CV4AnimalCLEF/.

Acknowledgments

We thank the organizers of the AnimalCLEF 2026 competition [2, 4] for providing the benchmark and dataset infrastructure. We acknowledge the authors of SeaTurtleID2022 [6], WildlifeReID-10k [8], WildlifeDatasets [9], and the Texas Horned Lizard dataset [7] for their foundational efforts in wildlife re-identification data collection and curation.

Declaration on Generative AI

During the preparation of this working-note draft, the authors used OpenAI ChatGPT/Codex for paraphrase and rewording, improving writing style, grammar and spelling checks, formatting assistance, citation-format assistance, and figure-layout refinement. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] L. Pícek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, 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.
- [2] 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.
- [3] L. Hubert, P. Arabie, Comparing partitions, *Journal of Classification* 2 (1985) 193–218. URL: <https://doi.org/10.1007/BF01908075>. doi:10.1007/BF01908075.
- [4] L. Adam, L. Pícek, K. Papafitsoros, D. Williams, D. Biffi, AnimalCLEF26 @ CVPR & CLEF, <https://kaggle.com/competitions/animal-clef-2026>, 2026. Kaggle.
- [5] L. Pícek, J. Straka, M. Jirik, E. Belotti, M. Duřa, J. Krausová, M. Bojda, V. Cermak, L. Bufka, R. Dvořák, et al., CzechLynx: A dataset for individual identification and pose estimation of the Eurasian lynx, *Scientific Data* 13 (2026) 511. doi:10.1038/s41597-026-06853-9.
- [6] 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.
- [7] 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. doi:10.1002/ece3.71167.
- [8] L. Adam, V. Cermak, K. Papafitsoros, L. Pícek, WildlifeReID-10k: Wildlife re-identification dataset with 10k individual animals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2025, pp. 2124–2134.
- [9] 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.
- [10] L. Otashashvili, 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).
- [11] Bohemian Visual Recognition Alliance, BVRA/MegaDescriptor-L-384 model card, Hugging Face, 2024. URL: <https://huggingface.co/BVRA/MegaDescriptor-L-384>, accessed July 3, 2026.
- [12] R. Wightman, N. Raw, A. Soare, A. Arora, C. Ha, C. Reich, F. Guan, J. Kaczmarzyk, M. Rizin, H. Kim, et al., *rweightman/pytorch-image-models: v0.8.10dev0 release*, Zenodo (2023). doi:10.5281/zenodo.7618837.

- [13] J. Li, P. Zhou, C. Xiong, S. C. Hoi, Prototypical contrastive learning of unsupervised representations, arXiv preprint arXiv:2005.04966 (2020).
- [14] Z. Dai, G. Wang, W. Yuan, S. Zhu, P. Tan, Cluster contrast for unsupervised person re-identification, in: Proceedings of the Asian conference on computer vision, 2022, pp. 1142–1160.
- [15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. 226–231.
- [16] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, The Journal of Open Source Software 2 (2017) 205. URL: <https://doi.org/10.21105/joss.00205>. doi:10.21105/joss.00205.
- [17] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2015. URL: <https://arxiv.org/abs/1412.6980>, International Conference on Learning Representations.
- [18] 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. URL: https://ceur-ws.org/Vol-4038/paper_253.pdf.
- [19] 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, IEEE, 2019, pp. 4685–4694. URL: <https://doi.org/10.1109/CVPR.2019.00482>. doi:10.1109/CVPR.2019.00482.
- [20] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 815–823. URL: <https://doi.org/10.1109/CVPR.2015.7298682>. doi:10.1109/CVPR.2015.7298682.