Rapid Exploration of Foundation Models and Supervised Contrastive Learning for Open-Set Individual Animal Discovery

Notebook for the AnimalCLEF Lab at CLEF 2026

Xiuxi Fan[1,*]

[1]*Zhengzhou University of Light Industry, Zhengzhou, China*

Abstract

This paper presents our approach for the AnimalCLEF 2026 open-set individual animal discovery challenge at LifeCLEF/CLEF, where—unlike the re-identification formulation of AnimalCLEF 2025—test images must be clustered by individual identity (including individuals absent from training) and evaluated by the Adjusted Rand Index (ARI). Working under severe resource constraints (a single researcher, one RTX 5090 GPU, total budget under US$28), we iterated through approximately 20 experimental versions over 5 active days. We evaluate five vision backbones (DINOv3-7B, InternViT-6B, SigLIP2-Giant, EVA02-CLIP-E+, and MegaDescriptor-L-384) and find that, in our experiments, general-purpose foundation models yield near-zero individual-level discriminability in their raw feature spaces. Training supervised contrastive (SupCon) projection heads on concatenated multi-backbone features raises training-set ARI from $\sim$0.001 to $\sim$0.90, but a severe train–test gap persists. Our best submission (team **Dr_F**) achieves a private leaderboard ARI of 0.320 (public: 0.240), ranking 49th of 232 teams, without using the WildlifeReID-10k dataset. We share lessons from this rapid exploration, including the observed failure of pseudo-labeling in low-data regimes and the difficulty of zero-shot transfer for individual animal clustering in our setting.

Keywords

Animal Re-identification, Open-Set Clustering, Foundation Models, Supervised Contrastive Learning, Individual Discovery, AnimalCLEF

1. Introduction

Individual animal identification is a cornerstone of modern wildlife ecology, enabling population monitoring, migration tracking, mark–recapture abundance estimation, and behavioural studies at the individual level [1]. Camera-trap networks—now deployed at planetary scale across conservation programmes—produce image volumes far beyond manual cataloguing capacity, motivating automated pipelines that re-identify or discover individuals without invasive tagging. The AnimalCLEF 2026 challenge [2], organized as part of the LifeCLEF lab at CLEF [3],[1] introduces a fundamentally different formulation from its 2025 predecessor [4]: rather than classifying query images into known identities (re-identification), participants must *discover* groupings of individual animals through clustering of test images, evaluated by the Adjusted Rand Index (ARI) [5].

This reformulation presents three key challenges: (1) the test set contains novel individuals not present in training, requiring open-set generalization; (2) one of four species (*Phrynosoma cornutum*, the Texas Horned Lizard) has zero training data, demanding fully unsupervised discovery; and (3) extreme data scarcity—the Salamander dataset has a median of only one image per individual.

In this work, we investigate whether modern large-scale foundation models can provide sufficient individual-level discriminability for this clustering task. We deliberately exclude the officially provided WildlifeReID-10k dataset [1] (26.76 GB, 140K images covering partial target species) from our pipeline, isolating the contribution of general-purpose foundation models from domain-specific pre-training

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

*Corresponding author.

2141571568@qq.com (X. Fan)

[1] https://www.kaggle.com/competitions/animal-clef-2026

data. Working as a single researcher (team **Dr_F**) with a budget under US$28 on one RTX 5090 GPU, we adopted a rapid-iteration strategy: trying each approach briefly and moving on if initial results were unpromising, rather than exhaustively tuning every variant. Across approximately 20 versions over 5 active days, we evaluate five vision backbones, two metric learning strategies (ArcFace and SupCon), and four clustering algorithms. Our exploration reveals a critical gap between training-set performance and test-set generalization, and we share practical lessons from failed strategies including pseudo-labeling, lookup-based approaches, and species-level clustering methods. Our code is publicly available.[2]

2. Related Work

Animal Re-identification. MegaDescriptor [1], a Swin Transformer [6] trained with ArcFace loss [7] on the WildlifeReID-10k dataset comprising 29 wildlife re-identification datasets, represents the current state-of-the-art for animal re-ID. MiewID [8] provides an alternative EfficientNetV2-based backbone trained on 49 species. WildFusion [9] demonstrates that calibrated fusion of global descriptors (MegaDescriptor) and local keypoint matchers (ALIKED and LightGlue [10]) via isotonic regression significantly improves matching accuracy across 17 wildlife datasets. In the AnimalCLEF 2025 challenge [4], MegaDescriptor-based approaches dominated the leaderboard, with the top solutions combining MegaDescriptor features with calibrated similarity fusion [9]. However, AnimalCLEF 2025 was a re-identification (closed-set retrieval) task; the 2026 edition reformulates the problem as open-set clustering, requiring fundamentally different approaches.

Self-supervised Foundation Models. Recent self-supervised models such as DINOv2 [11] and DINOv3 [12] learn general-purpose visual representations through self-distillation at scale. Vision-language models including SigLIP2 [13] and EVA02-CLIP [14] learn aligned image-text embeddings from web-scale data. InternViT-6B [15] further scales the vision transformer to 6 billion parameters. While these models excel at species-level classification and retrieval, their effectiveness for *intra-species individual-level* discrimination remains largely unexplored.

Open-Set Animal Clustering. Traditional re-ID formulations assume a closed set of known identities. Open-set extensions require detecting novel individuals, typically through threshold-based novelty detection [16]. The AnimalCLEF 2026 formulation [2] goes a step further, requiring *simultaneous* assignment to known groups and discovery of entirely new identity groups—making it fundamentally harder than retrieval-based re-ID. Recent work by Markoff et al. [17] benchmarks vision transformers for zero-shot clustering of animal images, finding that t-SNE + HDBSCAN achieves near-perfect *species-level* clustering; however, individual-level discrimination within a species remains a largely unsolved problem, as our experiments confirm.

3. Dataset and Task

3.1. Dataset Overview

AnimalCLEF 2026 provides images of four species with widely varying characteristics (Table 1), sourced from camera-trap and field studies across three continents. The first three datasets—LynxID2025, SalamanderID2025, and SeaTurtleID2022—are inherited from AnimalCLEF 2025 [4], while TexasHornedLizards is newly introduced this year. **LynxID2025** is built on CzechLynx [18], a large-scale camera-trap dataset of the Eurasian lynx (*Lynx lynx*) collected in the western Czech Republic. **SalamanderID2025** captures fire salamanders (*Salamandra salamandra*) from Czech forests. **SeaTurtleID2022** [19] depicts loggerhead sea turtles (*Caretta caretta*) photographed at Zakynthos, Greece, where individuals are recognised by their head-scale patterns. **TexasHornedLizards** [20] covers the

[2] https://github.com/fan1344rwere/AnimalCLEF2026-Solution

Table 1
Dataset statistics for AnimalCLEF 2026. Median images per identity (Med. img/id) highlights extreme data scarcity, particularly for Salamander.

Species	Train	IDs	Med.	Test	Key Challenge
LynxID2025 [18]	2,957	77	17	946	Mixed orientations (L/R/F/B)
SalamanderID2025 [1]	1,388	587	**1**	689	310 singleton identities
SeaTurtleID2022 [19]	8,729	438	13	500	Head-scale identification
TexasHornedLizards [20]	0	0	–	274	**Zero training data**
Total	13,074	1,102	–	2,409	–

Texas horned lizard (*Phrynosoma cornutum*) from Texas, USA, and is the only species in this challenge with zero training data, requiring fully unsupervised discovery.

The provided `metadata.csv` follows the `wildlife-datasets` toolkit convention [1], exposing per-image fields `identity`, `orientation` (left/right/front/back), `date`, `species`, `split`, and `dataset`. The `orientation` field is particularly informative for LynxID2025, where the same individual produces visually dissimilar embeddings across orientations. The `Train` split is labelled and serves as the reference set for metric learning; the `Test` split must be clustered without identity labels.

The Salamander dataset is particularly challenging: 53% of identities (310 out of 587) appear only once in training, making supervised metric learning severely data-starved. The Texas Horned Lizard provides no training data whatsoever, requiring fully unsupervised discovery from 274 test images alone.

3.2. Task Definition

The AnimalCLEF 2026 task [2] requires clustering test images such that each cluster corresponds to one individual animal. Unlike standard re-identification (matching a query to a known catalogue), this formulation targets *discovery*: the test set contains both known individuals (with training data available) and new individuals absent from training, and participants do not know which images belong to which category. The submission assigns each test image a cluster label of the form `cluster_<species>_<id>`; images of the same individual must share the same cluster name. Performance is evaluated by ARI, which penalizes both splitting one individual into multiple clusters (over-clustering) and merging different individuals into one cluster (under-clustering).

For species with training data, participants may use the labeled identities for metric learning or as reference embeddings. For Texas Horned Lizards, no training data is available, requiring fully unsupervised clustering. Participants may optionally use the WildlifeReID-10k dataset [1] for additional pre-training; we deliberately exclude it to isolate the contribution of general-purpose foundation models.

3.3. Evaluation Metric

Performance is measured by the Adjusted Rand Index (ARI) [5], which quantifies agreement between predicted and ground-truth clusterings, adjusted for chance:

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]} \tag{1}$$

where RI is the Rand Index (the fraction of image pairs on which the predicted and ground-truth clusterings agree), $\mathbb{E}[\text{RI}]$ is its expected value under random assignment, and $\max(\text{RI})$ is the maximum achievable RI for the given number of clusters. ARI ranges from -1 to 1, with 1 indicating perfect clustering agreement and 0 corresponding to random assignment.

Table 2
Vision backbones evaluated, ordered by embedding dimensionality.

Model	Architecture	Dim	Pre-training Data
DINOv3-7B	ViT-7B	8,192	LVD-1.7B (self-distill.)
InternViT-6B	ViT-6B	6,400	InternVL-2.5 (contrastive)
SigLIP2-Giant	ViT-Giant	1,536	WebLI (sigmoid loss)
MegaDescriptor-L	Swin-L/384	1,536	WildlifeReID-10k (ArcFace)
EVA02-CLIP-E+	ViT-E/14	1,024	LAION-2B (CLIP)

4. Methodology

Our pipeline evolved through approximately 20 versions over the course of the competition. Below, we describe the components of our best-performing system (V22) and key architectural variants.

4.1. Feature Extraction

We evaluate five vision backbones, deliberately chosen to span domain-specific and general-purpose architectures across a wide range of model scales (Table 2).

For each image, we extract CLS token embeddings (or global average-pooled features for the Swin-based MegaDescriptor). For DINOv3-7B, we concatenate the CLS token with the spatial mean of patch tokens, yielding 8,192-dimensional feature vectors. All features are L2-normalized prior to downstream processing.

4.2. SAM-based Foreground Extraction

In versions V21–V25, we apply SAM 2.1 [21] (Hiera-Large checkpoint, FP16 precision) with a single center-point prompt at image coordinates $(w/2, h/2)$ to generate a foreground mask for each image. Among the predicted masks, we retain the one with the highest predicted IoU and crop each image to its bounding box before feature extraction; if segmentation fails, we fall back to the uncropped image. This step removes background clutter from camera-trap images, which is particularly prevalent in the Lynx nighttime captures.

4.3. Supervised Contrastive Learning

To enhance individual-level discriminability, we train a 2-layer MLP projection head on the concatenated backbone features. The head maps from the concatenated input dimension (up to 18,688 for all 5 backbones) to a 256-dimensional L2-normalized embedding space. Training is performed *independently per species* on training data only; no joint training across species is conducted. We employ the supervised contrastive (SupCon) loss [22]:

$$\mathcal{L}_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \tag{2}$$

where $P(i)$ denotes the set of samples sharing the same identity as anchor i (positives), $\mathbf{z}$ are the L2-normalized projected features, $\tau = 0.07$ is the temperature hyperparameter, and the denominator sums over all samples in the mini-batch except anchor i (i.e., normalization is at the batch level, not the class level). Training uses a PK sampler (P identities $\times$ K images per identity per mini-batch) over 50 epochs with learning rate 5×10^{-4} (AdamW) and cosine annealing. Only the projection head is trained; all backbone weights remain frozen.

Table 3
V22 SupCon configuration per species. The projection head (18,688 → 1,024 → 256) is trained independently for each species with training data only. Texas receives a transferred head from SeaTurtle. Singleton rates in the Notes column are computed on predicted test clusters.

Species	PK Sampler	α	θ	Notes
Lynx	$P=32, K=2$	1.0	0.310	SupCon only
Salamander	$P=32, K=2$	1.0	0.290	SupCon only; 72% singletons on test
SeaTurtle	$P=32, K=4$	0.8	0.450	80% SupCon + 20% MegaDesc raw
Texas	–	–	–	Transfer from SeaTurtle (failed: 1 cluster)

4.4. Clustering Strategies

We evaluate four clustering algorithms applied to the cosine similarity matrix: **(1) Hierarchical Agglomerative Clustering (HAC)** with average linkage and per-species distance thresholds optimized via grid search; **(2) HDBSCAN** [23] with cosine metric and per-species minimum cluster size tuning; **(3) Louvain community detection** on a k-nearest-neighbor graph with resolution parameter search; and **(4) DBSCAN** with cosine distance and ε tuning.

4.5. Per-Species Similarity Fusion

For multi-backbone configurations, we perform per-species grid search over backbone fusion weights:

$$S_{\text{fused}}^{(s)} = \sum_m w_m^{(s)} \cdot S_m, \quad \text{s.t.} \sum_m w_m^{(s)} = 1 \tag{3}$$

where $w_m^{(s)}$ is the weight assigned to backbone m for species s, optimized to maximize training-set ARI. The grid search uses weight steps of 0.2 and threshold values from 0.3 to 0.7 (V21), which is relatively coarse due to computational constraints.

4.6. Final System Configuration (V22)

Among our 11 Kaggle submissions (Table 6), V22 achieves the highest private leaderboard score (0.320), exceeding the next-best version V16 (0.280) by 0.040 ARI; we therefore select V22 as our final submitted system. V22 builds on V21, our foundation-model-only baseline (DINOv3, InternViT, SigLIP2, EVA02-CLIP; described in Section 5.1), by adding MegaDescriptor-L-384 as a fifth backbone and training a supervised projection head on the concatenated 18,688-dimensional features. Table 3 summarizes the complete configuration, extracted directly from our code and run logs.

Projection head architecture. A 2-layer MLP: Linear(18,688 → 1,024) → BatchNorm → GELU → Dropout(0.1) → Linear(1,024 → 256) → L2-normalize. All backbone weights are frozen; only the projection head is trained.

Training details. SupCon loss with $\tau = 0.07$, AdamW (lr $= 5 \times 10^{-4}$, weight decay 10^{-4}), cosine annealing over 50 epochs, gradient clipping at 1.0. The PK sampler samples P identities with K images each per mini-batch.

Ensemble with raw MegaDescriptor. After projecting all images through the trained head, we compute $\alpha \cdot S_{\text{SupCon}} + (1-\alpha) \cdot S_{\text{MegaDesc}}$ where α is optimized per species on training ARI (Table 3). For SeaTurtle, including 20% raw MegaDescriptor similarity slightly improves over pure SupCon, likely because MegaDescriptor already captures individual-discriminative features for this species (a MegaDescriptor-dominant 3-backbone fusion achieves SeaTurtle train ARI ≈ 0.86 in our earlier V16 configuration).

Texas Horned Lizard. With zero training data, we transfer the SeaTurtle projection head and apply it to Texas features. This produces a single cluster for all 274 images, indicating that cross-species transfer of the SupCon projection head is ineffective.

5. Experiments and Results

5.1. Raw Foundation Model Features (V21)

Our first experiment evaluates the raw features of four general-purpose foundation models for individual-level clustering, *without any fine-tuning or projection.* As shown in Table 4, even with per-species weight optimization and k-reciprocal re-ranking, the best training-set ARI remains near zero. For comparison, a MegaDescriptor-dominant 3-backbone fusion from our earlier V16 configuration achieves ARI of 0.86 on SeaTurtle—hundreds of times higher than any combination of the four general-purpose foundation models in V21. This confirms that the low ARI is primarily due to the foundation-model features lacking individual-level discriminability, not the clustering algorithm; however, both components contribute to the final result.

Closed-set retrieval vs. open-set clustering. Prior reports that DINO performs well on SeaTurtleID2022 in a *closed-set retrieval* setting—where the goal is to rank known training identities by similarity to a query image—are consistent with our findings. Open-set clustering places a strictly stronger demand on the feature space: there must exist a clear similarity gap between same-individual and different-individual pairs so that a single global clustering threshold can separate groups.

The strongest evidence that this gap is absent in our foundation-model features—and that the V21 bottleneck lies in the feature space rather than the clustering algorithm—comes from a controlled comparison on SeaTurtle: under identical HAC clustering with identical training data and identical grid-searched distance thresholds, a MegaDescriptor-dominant 3-backbone fusion (V16) achieves train ARI of 0.86, while the best 4-backbone fusion of general-purpose foundation models (V21, including DINOv3 at weight 0.2) yields only 0.004—a $\sim$200$\times$ difference. While these two fusions differ in backbone composition, the dominance of the Re-ID-trained MegaDescriptor in V16 and the near-zero ARI of every general-purpose backbone combination in V21 together indicate that the gap is driven by the feature extractor rather than by the clustering step. Had the bottleneck been the clustering algorithm, this gap would not appear, since the clustering algorithm, training data, and threshold-search procedure are held constant.

How low is DINOv3 alone under the V21 setup? The V21 per-species weight grid searches 56 convex weight combinations (4 backbones, step 0.2, sum=1.0), which explicitly includes the DINOv3-alone vector $(1, 0, 0, 0)$. The grid best ARI on SeaTurtle is 0.0009, attained by the (0.2/0.2/0.4/0.2) fusion—so DINOv3 alone under the V21 setup (which uses 518$\times$518 input and SAM 2.1 foreground crop) scores at most 0.0009 raw training ARI on this species. This directly addresses the concern that DINOv3 should not score so low: in our pipeline it does, even when allowed as the sole backbone. The corresponding grid best ARIs for Lynx (0.0048) and Salamander ($\approx$0) are similarly low. We also note that the V21 preprocessing choices may interact poorly with DINOv3 specifically: an independent re-run of DINOv3-7B alone on the same SeaTurtle train data at the model's default 224$\times$224 input resolution and *without* SAM 2.1 foreground crop yields a higher train ARI of 0.029 (best HAC threshold 0.10), suggesting that the SAM-cropped high-resolution V21 setup is partly responsible for the very low DINOv3 number. Even at this higher number, DINOv3 alone remains two orders of magnitude below the MegaDescriptor-dominant V16 fusion (0.86) and well below 0.1, so the central claim—that general-purpose foundation models lack individual-level discriminability for SeaTurtle—is preserved either way.

Furthermore, none of our four foundation backbones in any per-species weight combination clears 0.1 train ARI on any species with training data (Table 4), indicating that—*within our V21 preprocessing pipeline* (518$\times$518 input, SAM 2.1 foreground crop, cosine HAC)—the issue is structural rather than specific to a particular backbone. The t-SNE visualization in Figure 1(b) illustrates this failure mode concretely for MegaDescriptor features, where individual sea turtles of different identities are completely intermingled within the species cluster; the near-zero ARI of the 4-backbone fusion on the same species suggests a similar geometry for the foundation-model features.

Table 4
V21 results: training-set ARI using raw foundation model features (DINOv3, InternViT, SigLIP2, EVA02) with per-species weight optimization, k-reciprocal re-ranking, and HAC clustering. Weights are obtained by grid search on training ARI. The SeaTurtle result uses all four backbones.

Species	Best weight combination	Train ARI	Test clusters
Lynx	DINOv3 0.4 + EVA02 0.6	0.083	312
Salamander	EVA02 1.0 (others 0)	<0.002	189
SeaTurtle	4-backbone (0.2/0.2/0.4/0.2)	0.004	157
TexasHorned	Equal weights	–	1

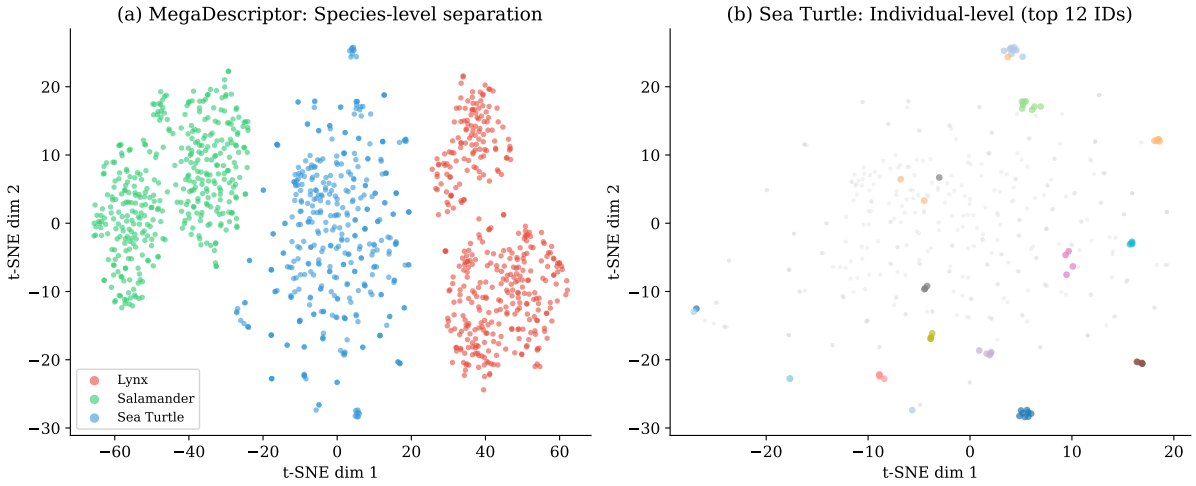

Figure 1: t-SNE visualization of MegaDescriptor-L-384 embeddings for 3 species with training data. (a) Species-level: Lynx (blue), Salamander (orange), and SeaTurtle (green) form clearly separated clusters, demonstrating strong species-level discriminability. (b) Individual-level within SeaTurtle: the top 12 identities (each colored differently) are completely intermingled, showing that MegaDescriptor features cannot distinguish individuals within a species. This motivates the SupCon projection approach in V22.

We therefore attribute the V21 bottleneck primarily to the feature space rather than the clustering step, while acknowledging that both components contribute to the final score. This motivates the supervised projection in Section 4.3, which explicitly reshapes the feature space to introduce such a gap.

Figure 1 visualizes MegaDescriptor embeddings via t-SNE: species form distinct clusters (a), yet individual sea turtles remain hopelessly intermingled within a species cluster (b).

5.2. Supervised Contrastive Projection (V22)

Adding MegaDescriptor as a fifth backbone and training a SupCon projection head on the concatenated 18,688-dimensional features produces dramatic improvements on the training set (Table 5). However, the private leaderboard score of **0.320** reveals a training–private gap of ∼0.60, confirming severe overfitting. Notably, Salamander achieves the highest training ARI (0.90) but produces 578 clusters for 689 test images with 72% singletons—suggesting the projection head memorizes individual training images rather than learning transferable dorsal pattern features.

5.3. Clustering Method Comparison

We evaluated four clustering algorithms on SeaTurtle training data using V22 SupCon features: Hierarchical Agglomerative Clustering (HAC) with average linkage, HDBSCAN [23], Louvain community detection, and DBSCAN. HAC produced the highest training ARI (**0.908**, 16% singletons) and was selected as the final clustering method for all species; the alternative methods were screened out during pilot experiments on cached features and were not retained in our production V22 logs. Qualitatively,

Table 5
Effect of SupCon projection (V22). Training ARI improves dramatically but does not generalize. Test cluster statistics are computed from our submitted V22 run; "Singletons" denotes the percentage of predicted test clusters that contain exactly one image (i.e., test images the system chose not to merge with any other).

Species	Raw ARI	SupCon ARI	Test clusters	Singletons
Lynx	0.083	**0.264**	69	23 (2%)
Salamander	<0.002	**0.901**	578	496 (72%)
SeaTurtle	0.004	**0.908**	203	79 (16%)
TexasHorned	–	–	1	0 (0%)

Table 6
Summary of key experimental versions with both public and private leaderboard scores. ↓ marks degradation below the V9 baseline. Many low-scoring versions reflect insufficient tuning under resource constraints rather than fundamental method failures—we adopted a rapid-iteration strategy, skipping approaches that did not show immediate promise.

V	Method	Public	Private	Key Finding
9	MegaDesc + HAC	0.177	0.209	Simplest, most stable
14	MegaDesc + MiewID + TTA	0.205	0.252	First time beating baseline
16	3-backbone ensemble	0.231	0.280	Per-species weights help
17	+ ArcFace fine-tuning	0.225	0.278	Salamander: 0% acc
18	Joint train+test clustering	$0.117^{\downarrow}$	0.154	Constraints overfit
19	Orientation-aware matching	0.179	0.178	Penalties too aggressive
20	Local-primary (Louvain)	$0.113^{\downarrow}$	0.145	Graph fragmentation
22	**SupCon 5-backbone**	**0.240**	**0.320**	**Best submission**
23	Pseudo-labels + HDBSCAN	$0.105^{\downarrow}$	0.126	Error propagation
24	t-SNE + HDBSCAN	$0.147^{\downarrow}$	0.145	Individual $\neq$ species clustering
25c	SupCon 3-backbone + HAC	0.172	0.240	Fewer backbones hurts

we observed that DBSCAN catastrophically fragments in high-dimensional cosine spaces, while Louvain over-fragments when similarity edges are sparse. HAC was preferred because its single distance threshold maps cleanly onto a per-species grid search on training ARI, providing the most reliable granularity control under our resource constraints.

5.4. Version History and Failed Strategies

Table 6 summarizes the full experimental trajectory. Several findings merit special attention.

Pseudo-label iteration is catastrophic (V23). We applied iterative self-training: cluster → pseudo-labels → SupCon retraining → recluster. This caused Lynx ARI to collapse from 0.74 (initial phase) to 0.10 (after 2 rounds), as clustering errors compounded across iterations in this extremely low-data regime.

Species-level $\neq$ individual-level clustering (V24). Following Markoff et al. [17], who report V-measure of 0.958 for *species-level* clustering with a t-SNE + HDBSCAN pipeline, we applied the same pipeline to individual discovery. In our hands the pipeline produced only 0.147 ARI for individual clustering—consistent with the two tasks requiring fundamentally different feature granularity, and confirming the per-species ceiling already reported in [17].

Lookup strategies are dangerous (V7, V13). Direct identity lookup (matching test images to training identities via nearest-neighbor with threshold) proved counterproductive. In V13, the lookup threshold was set low enough that the vast majority of Lynx test images were matched to known

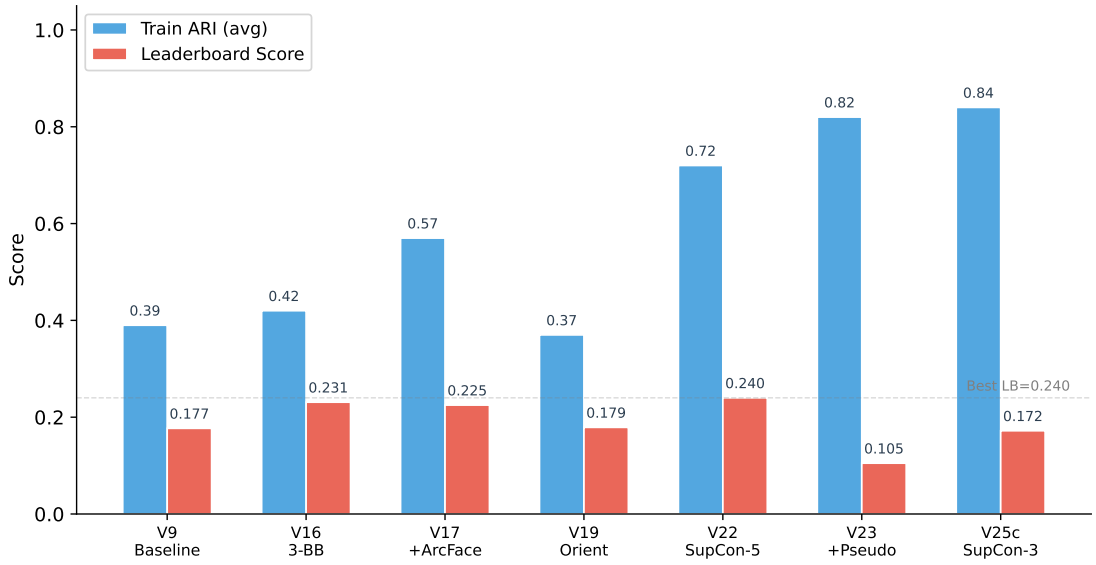

Figure 2: Training ARI vs. public leaderboard score across experimental versions. Higher training ARI does not translate to better test performance, demonstrating severe overfitting. V22's private score (0.320, green star) exceeds its public score (0.240), though this pattern is common across most teams.

Table 7
Overfitting across metric learning approaches. All methods show a substantial gap between training ARI and leaderboard performance.

Method	Train ARI (avg)	Public LB	Private LB
Raw MegaDescriptor (no learning)	0.39	0.177	0.209
ArcFace fine-tuning	0.57	0.225	0.278
SupCon (5 backbones)	0.72	0.240	**0.320**
SupCon (3 backbones)	0.84	0.172	0.240

training identities—catastrophically assigning novel individuals to incorrect training identities, yielding ARI of only 0.113.

6. Analysis

6.1. The Overfitting Problem

The central finding of our study is the persistent gap between training and test performance, illustrated in Figure 2. As training-set ARI increases through more sophisticated metric learning, the leaderboard score *does not improve proportionally*—and in many cases *decreases*.

Our best model (V22) achieves a private score of 0.320 versus public 0.240. While the private-public difference is positive, we note that many teams on the leaderboard exhibit a similar pattern (private > public), suggesting it may be partly a property of the competition's train/test split rather than purely evidence of method-specific generalization. The training–private gap of $\sim$0.60 remains substantial and represents the central challenge.

Table 7 confirms this pattern across multiple metric learning approaches. The overfitting is structural: with a median of 1–17 images per identity, projection heads inevitably learn identity-specific artifacts (camera angles, background textures, lighting conditions) rather than identity-invariant discriminative features. Additionally, all hyperparameters (fusion weights, clustering thresholds) are optimized on training ARI with no held-out validation set, further contributing to the gap.

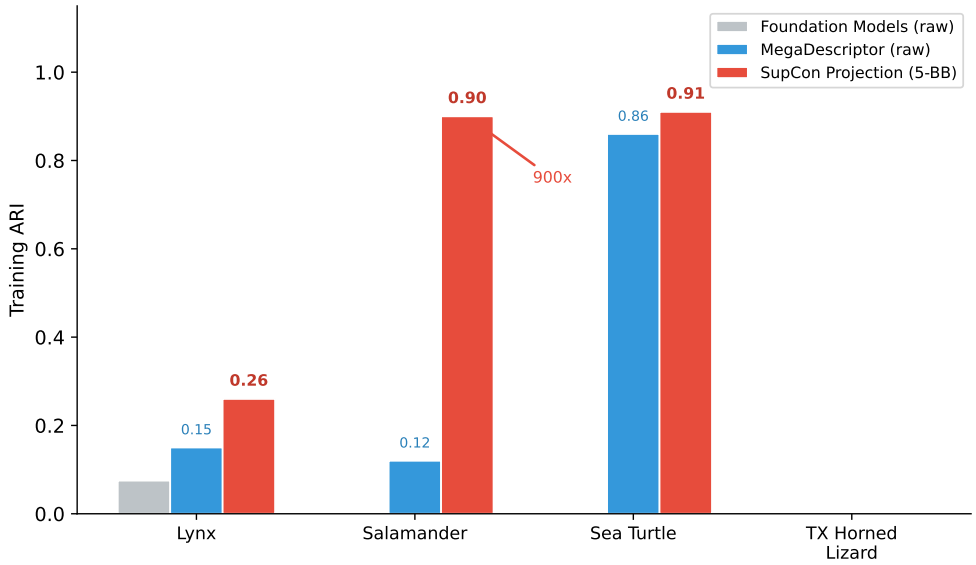

Figure 3: Per-species training ARI across three feature extraction approaches for LynxID2025, SalamanderID2025, and SeaTurtleID2022. Under our V21 preprocessing pipeline (518×518 input, SAM 2.1 foreground crop, cosine HAC), foundation-model raw features (gray) provide near-zero individual discriminability. SupCon projection (red) improves ARI by up to 900× on training data but the improvement does not generalize to the test set (Table 7). Texas Horned Lizard is excluded due to zero training data.

6.2. Per-Species Difficulty Analysis

Performance varies dramatically across species (Figure 3), driven by data availability and visual distinctiveness:

- **SeaTurtle** (train ARI 0.91): Head scale patterns are highly distinctive and consistent across views. With 8,729 training images, this species benefits most from metric learning. Test output: 203 clusters, 16% singletons.
- **Lynx** (train ARI 0.26): Camera-trap images with mixed orientations (left, right, front, back) produce vastly different feature vectors for the same individual, limiting discriminability even after SupCon projection.
- **Salamander** (train ARI 0.90, but 72% singletons on test): Despite high training ARI, the median of 1 image per individual causes the projection head to memorize individual training images rather than learning robust dorsal pattern features. The test set of 689 images is fragmented into 578 clusters, with 496 singletons.
- **Texas Horned Lizard** (1 cluster): With zero training data, all supervised approaches fail. Cross-species transfer from SeaTurtle's projection head yields only a single cluster for all 274 images.

6.3. Leaderboard Score Distribution

Figure 4 shows the score distribution of all 232 teams on both the public and private leaderboards. The two distributions overlap but are not identical: the top-ranked team on the public leaderboard (Ram Lab, ARI 0.819) drops to rank 2 on the private leaderboard, while M.I.A (public rank 4) rises to private rank 1. The top-5 private scores fall within a compact 4.2% relative range (0.674–0.704 ARI), and the median team scores 0.225 publicly and 0.228 privately. Our submission (team Dr_F, red star) is at private ARI 0.320 (rank 49 of 232) and public ARI 0.250 (rank 81), without utilizing the officially provided WildlifeReID-10k dataset.

We cannot definitively attribute the gap between our score and the top-ranked teams to any single factor. The 6th-place team's documented pipeline [24] illustrates the sophistication of top submissions:

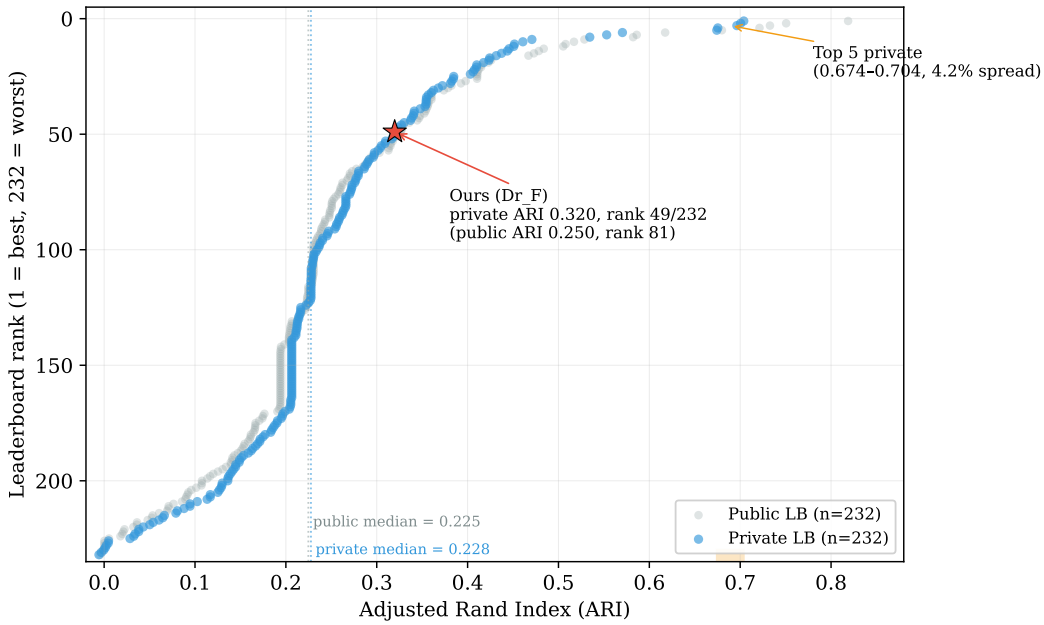

Figure 4: AnimalCLEF 2026 leaderboard score distribution (232 teams). Each point is one team; rank 1 is the best team. Public-leaderboard scores (gray, n=232) and private-leaderboard scores (blue, n=232) are shown together. Our submission (team Dr_F, red star) is at private ARI 0.320 (rank 49 of 232) and public ARI 0.250 (rank 81). The top-5 private scores (orange band) fall within a compact 4.2% relative range (0.674–0.704); medians are 0.225 (public) and 0.228 (private).

it combines ArcFace-fine-tuned MegaDescriptor checkpoints, LightGlue+SuperPoint local feature matching, per-species graph-based clustering with anchored identity propagation, and an XGBoost pair classifier that consumes LightGlue match statistics and descriptor cosines. The deliberate exclusion of WildlifeReID-10k from our pipeline is a design choice, not evidence that accounts for the leaderboard gap.

7. Discussion and Conclusion

Through rapid iteration under severe resource constraints, we evaluate the applicability of large-scale vision foundation models for open-set individual animal discovery. Our principal findings are:

1. **Raw foundation model features are insufficient for individual-level clustering in our experiments.** Despite scaling up to 7 billion parameters, general-purpose models (DINOv3, InternViT, SigLIP2, EVA02-CLIP) produce near-zero intra-species individual ARI *under the preprocessing pipelines we tested* (V21: 518×518 + SAM 2.1; default: 224×224 without SAM). We do not claim this is a fundamental property of the models themselves—a different preprocessing pipeline or a fine-tuned task head could plausibly yield different numbers—only that under both setups we evaluated, the gap to a Re-ID-trained backbone (MegaDescriptor) is two orders of magnitude.
2. **Supervised contrastive learning dramatically improves training ARI but overfits severely.** SupCon projection heads boost training ARI by up to $900\times$, yet exhibit a training–test gap of $\sim$0.60 (private LB). The Salamander case is particularly instructive: 0.90 training ARI but 72% singletons on test, suggesting image-level memorization rather than feature learning.
3. **HAC outperforms modern clustering alternatives.** Hierarchical agglomerative clustering with average linkage provides the most reliable results, likely due to its direct threshold-based granularity control.
4. **Multi-backbone diversity matters.** Five-backbone fusion (private LB = 0.320) outperforms three-backbone fusion (public LB = 0.172), indicating that architecturally diverse features provide complementary discriminative signals for the projection head.

5. **Cross-species transfer of projection heads fails.** Transferring the SeaTurtle SupCon head to Texas Horned Lizard produces a single cluster for all 274 images, indicating that individual-discriminative features are not transferable across distantly related species.
6. **Pseudo-labeling is catastrophic in low-data regimes.** Iterative self-training amplifies clustering errors, collapsing well-structured embeddings within two iterations.

Future work should explore *self-supervised* contrastive learning on the combined train+test set without identity labels, which could learn individual-discriminative features without the risk of identity memorization. Incorporating WildlifeReID-10k for domain-adaptive pre-training represents another natural extension.

8. Limitations

We acknowledge several limitations of our study:

1. **Resource constraints.** This work was conducted by a single researcher with a budget under US$28 on one GPU across approximately 5 active days. Many experimental versions were under-optimized due to time and cost pressure; low scores often reflect insufficient tuning rather than fundamental method failures.
2. **No validation set.** All hyperparameters (fusion weights, clustering thresholds, SupCon learning rate, number of epochs) are optimized directly on training-set ARI. This contributes significantly to the observed overfitting and inflates training performance.
3. **Coarse hyperparameter search.** The V21 weight grid uses steps of 0.2 with thresholds from 0.3 to 0.7—a very coarse search space. The V22 threshold search (steps of 0.02 on training ARI) is finer but still overfits to training data. A proper validation set would enable more reliable hyperparameter selection.
4. **Single hardware configuration.** All experiments run on a single RTX 5090 (32 GB). We do not assess sensitivity to different hardware, batch sizes, or numerical precision.
5. **No cross-validation.** Results are reported from single runs without cross-validation or multiple random seeds, making it difficult to distinguish signal from noise in the version comparison.
6. **Incomplete evaluation of backbones.** We do not report individual backbone ARI on the test set, making it difficult to isolate feature quality from clustering quality as the source of poor performance (though the near-zero training ARI with all backbones suggests the features themselves lack individual discriminability).

Acknowledgments

The computational experiments were performed on an NVIDIA RTX 5090 (32 GB) GPU hosted on the AutoDL cloud computing platform. The author thanks the AnimalCLEF organizers for providing the competition data and baseline code.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) for assistance with English grammar, LaTeX formatting, and figure generation. After using this tool, the author reviewed and edited all content and takes full responsibility for the publication.

References

[1] 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.

[2] L. Adam, K. Papafitsoros, D. A. Williams, D. Biffi, L. Picek, 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. Picek, 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.
[4] L. Adam, K. Papafitsoros, R. Kovář, V. Čermák, L. Picek, Overview of AnimalCLEF 2025: Recognizing individual animals in images, in: CLEF 2025 Working Notes, 2025.
[5] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218.
[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[7] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[8] L. Otarashvili, T. Subramanian, J. Holmberg, J. Levenson, C. V. Stewart, Multispecies animal Re-ID using a large community-curated dataset, arXiv preprint arXiv:2412.05602 (2024).
[9] V. Čermák, 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.
[10] X. Zhao, X. Wu, W. Chen, P. C. Chen, Q. Xu, Z. Li, Aliked: A lighter keypoint and descriptor extraction network via deformable transformation, IEEE Transactions on Instrumentation and Measurement 72 (2023) 1–16.
[11] M. Oquab, T. Darcet, T. Moutakanni, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).
[12] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, P. Bojanowski, DINOv3, arXiv preprint arXiv:2508.10104 (2025).
[13] M. Tschannen, et al., Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, arXiv preprint arXiv:2502.14786 (2025).
[14] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, Y. Cao, Eva-02: A visual representation for neon genesis, Image and Vision Computing 149 (2024) 105171.
[15] Z. Chen, et al., Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, arXiv preprint arXiv:2312.14238 (2024).
[16] A. Miyaguchi, C. Maruthaiyannan, C. R. Clark, Ds@gt animalclef: Triplet learning over vit manifolds with nearest neighbor classification for animal re-identification, in: CLEF 2025 Working Notes, 2025.
[17] H. Markoff, S. H. Bengtson, M. Ørsted, Vision transformers for zero-shot clustering of animal images: A comparative benchmarking study, arXiv preprint arXiv:2602.03894 (2026).
[18] L. Picek, E. Belotti, M. Bojda, L. Bufka, V. Čermák, M. Dula, R. Dvořák, L. Hrdý, M. Jiřík, V. Kocourek, J. Krausová, J. Labuda, J. Straka, L. Toman, V. Trulík, M. Vana, M. Kutal, CzechLynx: A dataset for individual identification and pose estimation of the eurasian lynx, Scientific Data (2026). doi:10.1038/s41597-026-06853-9.
[19] 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.
[20] 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.

[21] N. Ravi, et al., Sam 2: Segment anything in images and videos, arXiv preprint arXiv:2408.00714 (2024).
[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 18661–18673.
[23] L. McInnes, J. Healy, S. Astels, Accelerated hierarchical density based clustering, Journal of Open Source Software 2 (2017).
[24] G. Colling, AnimalCLEF 2026: Entry to the individual animal re-identification challenge (LifeCLEF lab, CLEF 2026), Working note preprint, 2026. URL: https://gillescolling.com/side-projects/animal-clef-2026. doi:10.5281/zenodo.20055000.