Fusion of Global and Local Descriptors with Feature Calibration for Multi-Species Animal Re-Identification: AnimalCLEF 2025

Notebook for the AnimalCLEF 2025 Lab at CLEF 2025

Dongyeon Kim[1,*,†], Bohee Park[2,†], Hanjun Bae[1,†], Sua Lee[3,†] and Chaeyeon Lee[2,†]

[1]Myongji University, Seoul, South Korea
[2]Sookmyung Women's University, Seoul, South Korea
[3]Sogang University, Seoul, South Korea

Abstract

This paper presents a multi-matcher fusion pipeline developed for the AnimalCLEF 2025 individual animal re-identification challenge. The task involves identifying distinct individuals within the same species, under constraints such as one-shot learning and open-set recognition for previously unknown individuals. To address these challenges, we integrate three complementary matchers: MegaDescriptor for global visual features, ALIKED for local keypoint-based matching, and EVA02 for semantic-level similarity. These components are fused using WildFusion-based score calibration and a simple weighted averaging scheme. Additionally, species-specific preprocessing, such as orientation normalization for salamanders and 5-crop Test-Time Augmentation, is applied to enhance robustness. Our final pipeline achieved a public score of 0.50708 and a private score of 0.53185, representing a 23.2 percentage points relative improvement over a baseline solution (0.3002). According to the official leaderboard, our system ranks 44th out of 230 participating teams, placing in the top 19%. This outcome demonstrates the effectiveness of combining global, local, and semantic descriptors through calibrated fusion in a multi-species wildlife ReID context. The full implementation of our pipeline is available at https://github.com/dongyeon1031/AnimalCLEF2025.

Keywords

Animal Re-ID, Multi-Species, Fusion, Global Descriptor, Local Matching, CEUR-WS

1. Introduction

The global decline in biodiversity has intensified the demand for automated technologies capable of supporting wildlife monitoring, including population tracking, migration analysis, and behavioral studies [1]. Individual-level animal re-identification plays a crucial role in enabling fine-grained ecological analysis and supporting conservation strategies that go beyond species-level classification [2, 3]. However, most existing computer vision systems focus solely on species recognition and often fail to discriminate between individuals within the same species.

This study was conducted based on the individual identification task proposed in LifeCLEF 2025 [4], specifically following the problem definition and data configuration of the AnimalCLEF track [5]. The AnimalCLEF 2025 competition challenges [5] participants to design robust systems for identifying individual animals across multiple species [2, 6]. In particular, we followed the problem definition and data configuration provided in the AnimalCLEF track. The competition emphasizes open-set recognition, requiring models to generalize to unknown individuals captured under diverse conditions. To meet these demands, our approach prioritizes practical and modular design choices tailored for deployment in real-world scenarios.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
ehddus0087@naver.com (D. Kim); bboohhee06@gmail.com (B. Park); hbae0830@mju.ac.kr (H. Bae); alicelee49@naver.com (S. Lee); sarahlcy17@gmail.com (C. Lee)

In contrast to species classification, which aims to distinguish between different animal types (e.g., lion vs. giraffe), the central task of individual re-identification involves differentiating among distinct individuals within the same species (e.g., Lion A vs. Lion B). This presents a more complex recognition problem, particularly under conditions of limited training data and varying pose, lighting, and background. The key evaluation criterion is the system's ability to generalize beyond the training identities and accurately match unknown individuals based on visual cues alone.

The AnimalCLEF 2025 competition [5] focuses on individual animal re-identification and introduces several key technical challenges:

- **One-shot learning:** Many individuals in the reference set are represented by only one or two images, requiring models to generalize with minimal supervision.
- **Open-set recognition:** The test set contains individuals not seen during training, necessitating open-set recognition capabilities beyond standard closed-set classification.
- **Pose and illumination variation:** Images captured in unconstrained environments exhibit wide variations in pose, lighting, and resolution, demanding robust feature extraction and invariant representation learning.
- **Data imbalance:** The number of images per identity is highly imbalanced, which can introduce training bias and reduce generalization performance.

2. Related Work

2.1. Global Descriptor for Animal Re-Identification

Global descriptors are widely used in animal re-identification to compute visual similarity by embedding images into a feature space that captures overall appearance. MegaDescriptor [2] is a Swin Transformer [7]-based model trained on 29 wildlife datasets using metric learning with ArcFace loss [8]. It serves as a foundation model for animal re-identification and has demonstrated superior performance over other pretrained descriptors such as CLIP [9] and DINOv2 [10] across diverse species [2].

To further enhance semantic representation, EVA02 [11], a vision transformer pretrained with CLIP-style supervision, is known to capture high-level semantics that are often overlooked by conventional CNN-based descriptors.

WildFusion [12] performs calibrated fusion of similarity scores from heterogeneous descriptors, and has demonstrated effectiveness in improving robustness in open-set scenarios.

2.2. Local Matching-Based Complementary Methods

While global descriptors capture holistic visual appearance, they often struggle in scenarios involving occlusion, viewpoint variation, or partial visibility. To address these challenges, local matching techniques have been proposed to provide fine-grained correspondence information.

ALIKED [13] is a learning-based local feature extractor and matcher that balances matching accuracy with computational efficiency. It leverages patch-level correspondences to estimate image similarity and has shown strong performance in challenging visual re-identification tasks.

LoFTR [14, 15] represents a dense matching approach that enables pixel-wise correspondence across entire images without relying on keypoint detection. While LoFTR improves alignment robustness, its high computational overhead often limits its practical use in large-scale or real-time systems.

2.3. Practical Strategies for Visual Re-identification

Beyond descriptor design, several practical strategies have been explored to improve the robustness of visual re-identification pipelines. For instance, geometric normalization techniques have been used to standardize orientation in datasets with variable poses, and Test-Time Augmentation (TTA) has been employed to improve generalization by averaging predictions over multiple image views [16].

Furthermore, fusion strategies such as WildFusion [12] have demonstrated the effectiveness of combining global and local similarity scores via calibrated score-level fusion, enabling more flexible matching across representation levels.

These prior studies on descriptor fusion, local-global integration, and augmentation strategies collectively inspired the design of our pipeline, which integrates complementary modules to improve robustness in open-set and fine-grained recognition settings.

3. Dataset and Evaluation Metrics

The AnimalCLEF 2025 competition [5] addresses the task of individual-level recognition across multiple wildlife species. It evaluates models based on their ability to correctly match known individuals and to generalize to previously unknown ones, under an open-set recognition setting.

3.1. Dataset Overview

The AnimalCLEF 2025 challenge [5] provides image datasets [2, 6] for three species: sea turtles (SeaTurtleID2022), salamanders (SalamanderID2025), and lynxes (LynxID2025), derived from the CzechLynx dataset [17]. Each dataset is composed of a labeled database set, containing identity annotations for known individuals (e.g., LynxID2025_lynx_17), and a query set, for which the model must either retrieve the correct identity or predict *new_individual* if the target is not present in the database.

Metadata for all images is provided in a unified `metadata.csv` file, which includes image paths, individual identifiers, species labels, orientation information, and query/database status.

Each species presents unique visual challenges that may hinder reliable individual identification:

- **Sea turtles (SeaTurtleID2022):** Images are often low in resolution and suffer from underwater distortions.
- **Salamanders (SalamanderID2025):** Images may include human hands or inconsistent orientation, leading to feature misalignment.
- **Lynxes (LynxID2025) [17]:** Images are captured by camera traps, some taken at night. These often include back-facing individuals or cases where facial features are not clearly visible.

Beyond the basic provided data, a large auxiliary dataset, **WildlifeReID-10k**, is available. It includes images of 10,000 individuals from 36 animal species (marine mammals, birds, primates, livestock, etc.), totaling around 140,000 images. This auxiliary set can be used for model pre-training.

3.2. Evaluation Metrics

The competition evaluates model performance using two complementary metrics that account for both known and unknown individuals:

- **BAKS (Balanced Accuracy for Known Samples):** Measures the balanced classification accuracy for individuals in the reference set, adjusting for class imbalance across identities.
- **BAUS (Balanced Accuracy for Unknown Samples):** Measures the accuracy of detecting unknown individuals by evaluating whether *new_individual* is correctly assigned to novel queries.

The final score is computed as the geometric mean of the two metrics, as shown in Equation (1).

$$\text{Final Score} = \sqrt{\text{BAKS} \times \text{BAUS}}. \tag{1}$$

Equation 1 encourages a balanced treatment of both known identity classification and open-set novelty detection.

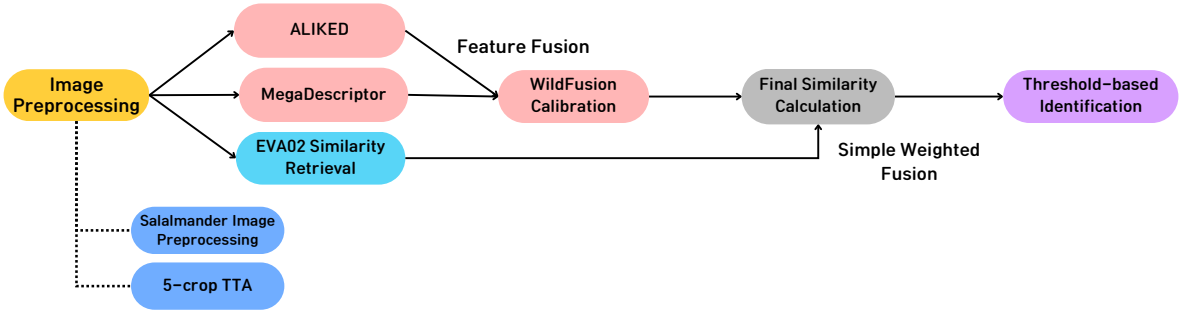

Figure 1: Overview of the proposed re-identification pipeline. Similarity scores are computed using three complementary matchers—MegaDescriptor (global) [2], ALIKED (local) [13], and EVA02 (semantic) [11]—and integrated via the WildFusion framework [12]. The final prediction is obtained through threshold-based classification on the fused similarity scores.

4. Methodology

The pipeline begins with image preprocessing (resizing and normalization), followed by feature extraction using three matchers: MegaDescriptor [2], EVA02 [11], and ALIKED [13]. Each matcher computes similarity scores based on global, semantic, or local representations, which are then calibrated and fused via the WildFusion module [12]. Final predictions are made through threshold-based classification on the fused scores. The overall pipeline operates through these stages, as summarized in Figure 1.

4.1. Preprocessing and Augmentation

We applied several preprocessing steps to improve feature consistency and robustness.

Image normalization and size processing ensures compatibility with the pretraining configuration of each model. Each image is resized and normalized to match the pretraining configuration of the respective model: MegaDescriptor [2] (384×384, ImageNet mean/std), EVA02 [11] (336×336, CLIP normalization).

Orientation Normalization for Salamander Dataset addresses inconsistent poses in the SalamanderID2025 data. Images are rotated based on orientation metadata: right views are rotated -90° and left views by $+90^\circ$ to standardize top-down alignment.

5-Crop Test-Time Augmentation (TTA) [16] is used to improve robustness during inference. Five crops include the center and four corners of the image, and their feature vectors are averaged, as shown in Equation (2).

$$\mathbf{v}_{\text{final}} = \frac{1}{5} \sum_{i=1}^{5} \mathbf{v}_i. \tag{2}$$

Metadata Utilization supports both preprocessing and matcher calibration. It guides salamander image orientation correction and helps distinguish query/database samples for calibration dataset construction.

4.2. Matching Strategy and Score Fusion

MegaDescriptor [2] extracts global embeddings using a Swin-Large backbone and computes cosine similarity, as shown in Equation (3).

$$\text{sim}(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \, \|\mathbf{d}\|}. \tag{3}$$

The similarity scores are normalized via isotonic regression to enable fusion with other matchers.

ALIKED [13] detects and matches keypoints to evaluate geometric consistency between image regions. It is particularly effective in challenging scenarios such as partial visibility, occlusion, and low-illumination, where global descriptors often fail to provide reliable similarity estimates. Furthermore, it offers a favorable trade-off between matching accuracy and computational cost, making it suitable for large-scale deployment.

EVA02 [11] uses a ViT-L/14-336 architecture to extract semantic-level embeddings. Cosine similarity is computed and calibrated for integration. This model is effective in species with low inter-individual variability.

WildFusion [12] calibrates raw similarity scores into probability-like outputs using isotonic regression, as shown in Equation (4).

$$\text{Cal}(s) = \hat{P}(y = 1 \mid s). \tag{4}$$

Final scores are first obtained through weighted fusion of global and local descriptors, as described in Equation (5).

$$\text{Sim}_{\text{WildFusion}} = \text{Fusion}(\text{Mega}, \text{ALIKED}). \tag{5}$$

Then, the final similarity score is computed by combining WildFusion and EVA02 scores, as shown in Equation (6). We initially set $\alpha = 0.5$ and subsequently conducted manual hyperparameter tuning to explore the effect of different α values.

$$\text{FinalScore} = \alpha \cdot \text{Sim}_{\text{WildFusion}} + (1 - \alpha) \cdot \text{Sim}_{\text{EVA02}}, \quad \alpha = 0.5. \tag{6}$$

The final decision logic is based on a threshold-based rule, as shown in Equation (7).

$$\hat{y} = \begin{cases} \text{Top-1 ID} & \text{if FinalScore} \geq \tau, \\ \text{Unknown Individual} & \text{otherwise.} \end{cases} \tag{7}$$

As seen in Equation (7), a threshold τ determines whether the prediction is assigned to an existing individual or labeled as unknown. We use an empirically tuned $\tau = 0.35$ with species-specific adjustments: -0.02 for sea turtles and $+0.02$ for salamanders.

We also experimented with LoFTR [14, 15]. While it showed robustness in aligning severely deformed instances, it was excluded from the final pipeline due to its high computational cost and limited performance gains relative to ALIKED [13] based on results from our validation experiments.

5. Experimental Setup

All experiments were conducted on a desktop PC equipped with an NVIDIA RTX 4080 Ti GPU running Windows 11. The implementation was based on Python 3.12.7 and PyTorch 2.5.1, with CUDA 11.8 for GPU acceleration.

The EVA02 matcher was implemented using the OpenCLIP library (v2.32.0) and employed the pretrained `merged2b_s6b_b61k` checkpoint for the ViT-L/14-336 model [18]. MegaDescriptor [2] and ALIKED [13] were implemented using the `wildlife-tools` library (v1.0.1), which provides standardized wrappers for global and local animal re-identification matchers.

To ensure reproducibility, we fixed the global random seed to 42 across Python `random`, NumPy, and PyTorch (including CUDA). No additional random splitting was performed, as the competition provided fixed database and query splits. The dataset was divided into a database set and a query set, following

Table 1
Comparison between the baseline solution, WildFusion, and our final pipeline submission, evaluated using the official public leaderboard score (i.e., final evaluation score). Each row summarizes the matchers and techniques used to compute similarity between query and database images.

Configuration	MegaDescriptor	ALIKED	EVA02	5-Crop TTA	Similarity Calculation	Private Score	Public Score
Baseline	O	X	X	X	Mega only (cosine), threshold=0.6	0.30898	0.3002
WildFusion	O	O	X	X	WildFusion (Mega + ALIKED)	0.44362	0.36555
Ours (Final)	O	O	O	O (all species)	0.5× WildFusion +0.5× EVA02	**0.53185**	**0.50708**

the official competition split. For score calibration, we additionally selected the first 1000 samples from both the query and database sets to form calibration subsets. These were used exclusively for training the isotonic regression model and were not involved in evaluation or matching.

Score calibration was performed using isotonic regression, trained on a balanced set of 1000 positive and 1000 negative image pairs generated from the calibration subset. Positive pairs consisted of images from the same individual, while negative pairs were drawn from different individuals within the same species. These matching pairs were automatically constructed during the score calibration procedure, and the internal pairing criteria were handled by the calibration module rather than explicitly defined in our implementation. Species balance was maintained during this sampling process.

The final similarity score was computed using late fusion of the calibrated scores from WildFusion [12] and cosine similarity scores from EVA02 [11]. Fusion was implemented manually outside the core pipeline. We tested multiple candidates for the fusion weight $\alpha \in \{0.3, 0.4, 0.5, 0.6\}$ and selected $\alpha = 0.5$ based on identification accuracy on the calibration set. A fixed decision threshold of 0.35 was used for determining identity matches in the final prediction step.

6. Results

To evaluate the effectiveness of our final re-identification pipeline, we compared it with the baseline solution, as summarized in Table 1.

The **Baseline** solution refers to the official starter notebook released for the AnimalCLEF2025 Competition [5]. It extracts global features using the MegaDescriptor [2] and computes cosine similarity between feature vectors. A fixed threshold of 0.6 is used to determine whether the match corresponds to a known or unknown individual. No local matcher or augmentation was applied [16]. This setup achieved a private score of **0.30898** and a public score of **0.3002**, highlighting the limitations of relying solely on global embeddings for individual animal identification, especially under challenging conditions like pose changes, occlusions, and background clutter.

The **WildFusion** configuration builds upon the baseline by incorporating the ALIKED [13] local matcher to enhance spatial alignment. Similarity scores from MegaDescriptor [2] and ALIKED [13] are independently computed and calibrated using isotonic regression, and subsequently fused via the WildFusion [12]. While semantic-level descriptors such as EVA02 [11] and Test-Time Augmentation [16] were not applied, the integration of local keypoint information significantly improved matching accuracy. This configuration achieved a private score of **0.44362** and a public score of **0.36555**, demonstrating the benefits of coarse-to-fine fusion using global and local features, particularly in scenarios with pose variation and partial visibility.

In contrast, the **Ours (Final)** configuration incorporates all proposed enhancements—ALIKED [13], EVA02 [11], and 5-Crop Test-Time Augmentation [16]—to improve re-identification performance. The EVA02 matcher [11] introduces semantic-level information, and its integration via weighted averaging with WildFusion [12] outputs led to notable improvements, as shown in Equation (6). In our experiments, we set $\alpha = 0.5$ and subsequently conducted manual hyperparameter tuning to explore the impact of different α values. Our final configuration achieved a private score of **0.53185** and a public score of **0.50708**, ranking 44th out of 230 teams (top 19%), and reflecting a 23.2 percentage points improvement over the baseline.

Table 2
Summary of alternative strategies and their performance. Each row describes an attempted modification to the re-identification pipeline, along with its private and public leaderboard scores from the AnimalCLEF 2025 challenge [5]. Despite theoretical motivations, most alternatives failed to outperform the final submitted configuration.

Strategy	Description	Private Score	Public Score	Remarks
Rerank Cascade	Re-matching with ALIKED for Top-K candidates	0.45588	0.42590	Accuracy degraded
LoFTR matcher	Replaced ALIKED with dense LoFTR matcher	0.44131	0.41803	Lower performance and slower speed
Segmentation Preprocessing	Foreground masking using Segment Anything model	0.49716	0.44136	Accuracy degraded
DINOv2 matcher added	Used DINOv2 matcher as auxiliary or replacement	0.51545	0.49805	No performance gain
Late Feature Fusion	Merged Mega and EVA02 features for classification	0.51124	0.49948	No significant improvement
Fusion MLP	MLP on Mega + EVA02 scores (submitted without training)	0.52501	0.49906	Training required

We applied Test-Time Augmentation [16] to enhance robustness against variations in pose and viewpoint. Five crops (center and four corners) were used to generate embeddings, which were then averaged. This method helped mitigate spatial misalignment and local occlusions leading to improved recognition accuracy compared to single-view inference.

7. Discussion

7.1. Failure Cases and Analysis

Several experimental strategies were tested to improve overall performance. However, some approaches resulted in suboptimal outcomes or failed to deliver the expected improvements. A summary of these strategies is presented in Table 2.

- **Rerank Cascade** (Private: 0.45588, Public: 0.42590): This method applied ALIKED matching to the top-K candidates generated by MegaDescriptor [2] and EVA02 [11] fusion. Although intended to refine the ranking and improve runtime efficiency, this approach resulted in a slight performance decline. One possible explanation is that the initial top-K candidates were already accurate in many cases, and the additional re-ranking introduced noise that disrupted correct top-1 predictions. This suggests that the benefits of re-ranking may be limited when the first-stage retrieval is already well-calibrated.
- **LoFTR Matcher [14]** (Private: 0.44131, Public: 0.41803): Replacing ALIKED [13] with the dense matcher LoFTR [14] resulted in degraded accuracy and slower inference. A plausible reason is that LoFTR's dense pixel-level correspondence may have captured irrelevant or background features, thereby weakening the distinctiveness of local matches. Furthermore, its computational overhead proved disadvantageous in a multi-matcher setting, making the method less practical under our pipeline constraints.
- **Segmentation Preprocessing** (Private: 0.49716, Public: 0.44136): We applied the Segment Anything Model (SAM) [19] to extract foreground-only regions for salamander and sea turtle images (see Figure 2). However, this approach did not improve performance and in some cases reduced it. This may be attributed to inconsistent mask quality and the unintended removal of contextual background cues that could aid in individual discrimination.
- **DINOv2 Matcher [10] Addition** (Private: 0.51545, Public: 0.49805): We added DINOv2 [10] as an additional matcher to MegaDescriptor [2] + ALIKED [13] pipeline. Although it yielded slightly higher scores than the baseline, the gains were not substantial. This suggests that simply

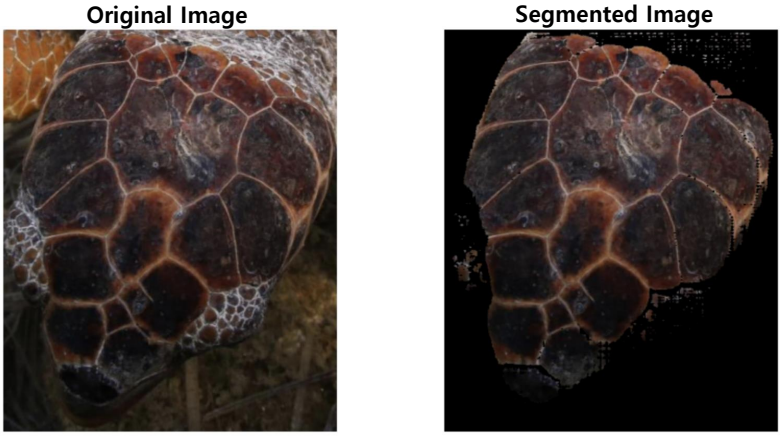

Figure 2: Example of SAM-based segmentation applied to a sea turtle image. While the background is removed, fine edge details may be distorted or lost.

increasing the number of matchers does not necessarily translate to consistent performance improvement.

- **Late Fusion [20]** (Private: 0.51124, Public: 0.49948): This approach simply averaged the similarity scores from MegaDescriptor [2] and EVA02 [11] without calibration or weighting. The resulting fusion underperformed relative to our WildFusion [12] + EVA02 [11] combination. A likely reason is that the lack of normalization of the scores prevented the model from effectively leveraging the complementarity between the two matchers.
- **Fusion MLP [21]** (Private: 0.52501, Public: 0.49906): A shallow MLP was introduced to learn a non-linear combination of scores from MegaDescriptor [2] and EVA02 [11]. Due to time constraints, the MLP was submitted without training, using random weights. Consequently, the untrained model underperformed. With proper training on positive and negative match pairs, this learning-based fusion approach remains a promising direction for future work.

7.2. Limitations and Computational Considerations

Although the proposed pipeline demonstrated strong performance in the AnimalCLEF 2025 challenge [5], several practical limitations remain.

First, the use of multiple independently operating matchers (MegaDescriptor [2], ALIKED [13], and EVA02 [11]) and 5-crop Test-Time Augmentation [16] increases inference cost. While this design contributes to robustness, it introduces computational overhead that may limit deployment in resource-constrained settings.

Second, the current system relies on manually tuned species-specific thresholds for decision-making. This reduces its adaptability to new domains or unseen categories. Incorporating more adaptive calibration techniques could help improve generalization in open-set scenarios.

8. Conclusion

In the context of the AnimalCLEF 2025 individual re-identification task [5], we developed a modular Global–Local–Semantic fusion pipeline. The core structure integrates global descriptors (MegaDescriptor [2]), local keypoint matchers (ALIKED [13]), and semantic-level representations (EVA02 [11]), further enhanced through 5-crop Test-Time Augmentation [16].

This combination enabled the system to maintain robust performance under challenging conditions, including variation in pose, lighting, resolution, and background. The final model achieved a private score of **0.53185** and a public score of **0.50708**, indicating consistent and improved performance through matcher-level fusion and spatial augmentation.

Future extensions of this research may include several directions aimed at improving generalization, efficiency, and real-world applicability.

- **Adaptive Open-set Handling**: The open-set nature of real-world re-identification demands further investigation into adaptive thresholding and novelty detection. Techniques such as confidence calibration or meta-recognition may offer more principled approaches than static thresholds.
- **Computational Efficiency [22]**: The current pipeline relies on multiple matchers and augmentation techniques, which increases computational cost. Optimizing for scalability via model pruning, knowledge distillation, or dynamic inference could enable deployment in large-scale or latency-sensitive settings.
- **Edge Deployment [23]**: Real-world applications often require mobile or embedded execution. Exploring quantization, lightweight model architectures, and platform-specific optimizations (e.g., TensorRT or mobile GPU support) would enhance portability for field use cases such as conservation monitoring or veterinary support.
- **Multimodal Integration [24]**: Incorporating auxiliary data such as timestamps, GPS locations, or environmental metadata could improve accuracy, particularly for visually ambiguous or low-variance species.
- **Cross-domain Transferability [25]**: The architecture's ability to discriminate fine-grained visual differences may be extended to other domains. For example, the framework could support assistive technologies like medication recognition for visually impaired users, where distinguishing similar visual instances is safety-critical.
- **Learning-based Fusion via MLP**: Due to time constraints, we were unable to train the fusion MLP and had to apply it with randomly initialized weights, which resulted in suboptimal performance. In future work, we plan to train the MLP to enable more effective model integration.

Acknowledgments

This work was carried out as an independent undergraduate team project, without external funding or institutional affiliation. The authors would like to thank all team members for their dedicated collaboration and contribution throughout the AnimalCLEF 2025 challenge.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o and Grammarly in order to assist with grammar, spelling, and clarity improvement. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Fergus, C. Chalmers, S. Longmore, S. Wich, Harnessing artificial intelligence for wildlife conservation, Conservation 4 (2024) 685–702.

[2] 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.

[3] P. C. Ravoor, T. Sudarshan, Deep learning methods for multi-species animal re-identification and tracking–a survey, Computer Science Review 38 (2020) 100289.

[4] L. Picek, S. Kahl, H. Goëau, L. Adam, T. Larcher, C. Leblanc, M. Servajean, K. Janoušková, J. Matas, V. Čermák, K. Papafitsoros, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, J. S. Cañas, G. Martellucci, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2025: Challenges on species presence

prediction and identification, and individual animal identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2025.
[5] L. Adam, K. Papafitsoros, R. Kovář, V. Čermák, L. Picek, Overview of AnimalCLEF 2025: Recognizing individual animals in images, Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum (2025).
[6] L. Adam, V. Čermák, K. Papafitsoros, L. Picek, Wildlifereid-10k: Wildlife re-identification dataset with 10k individual animals, arXiv preprint arXiv:2406.09211 (2024).
[7] 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.
[8] 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 (CVPR), 2019, pp. 4690–4699.
[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8748–8763.
[10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).
[11] 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.
[12] V. Cermak, L. Picek, L. Adam, L. Neumann, J. Matas, Wildfusion: Individual animal identification with calibrated similarity fusion, in: European Conference on Computer Vision, Springer, 2025, pp. 18–36.
[13] 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.
[14] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, Loftr: Detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8922–8931.
[15] Y. Wang, X. He, S. Peng, D. Tan, X. Zhou, Efficient loftr: Semi-dense local feature matching with sparse-like speed, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 21666–21675.
[16] D. Shanmugam, D. Blalock, G. Balakrishnan, J. Guttag, When and why test-time augmentation works, arXiv preprint arXiv:2011.11156 1 (2020) 4.
[17] L. Picek, E. Belotti, M. Bojda, L. Bufka, V. Cermak, M. Dula, R. Dvorak, L. Hrdy, M. Jirik, V. Kocourek, et al., Czechlynx: A dataset for individual identification and pose estimation of the eurasian lynx, arXiv preprint arXiv:2506.04931 (2025).
[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4015–4026.
[20] J. Kittler, M. Hatef, R. P. Duin, J. Matas, On combining classifiers, IEEE transactions on pattern analysis and machine intelligence 20 (1998) 226–239.
[21] N. Bodla, J. Zheng, H. Xu, J.-C. Chen, C. Castillo, R. Chellappa, Deep heterogeneous feature fusion for template-based face recognition, in: 2017 IEEE winter conference on applications of computer vision (WACV), IEEE, 2017, pp. 586–595.
[22] S. Han, H. Mao, W. J. Dally, Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, arXiv preprint arXiv:1510.00149 (2015).
[23] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko, Quantization

and training of neural networks for efficient integer-arithmetic-only inference, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2704–2713.

[24] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[25] K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning, Journal of Big data 3 (2016) 1–40.