

Cross-Species Animal Re-Identification with Semantic Consistency Learning

Shuoyi Chen^{*}, Yuejia Li^{*}, and Mang Ye[†]

School of Computer Science, Wuhan University, Wuhan, China
{chenshuoyi,liyuejia,yemang}@whu.edu.cn

Abstract. Generalizable animal Re-Identification (ReID) aims to recognize individual animals across species with diverse morphologies and ecological contexts. Unlike person ReID, where different domains share similar body structures, animal species often exhibit drastically different anatomical structures and visual patterns, making it difficult to establish shared visual correspondences. As a result, representations learned across species tend to form fragmented embedding spaces, which severely limits cross-species generalization. To address this challenge, we propose Semantic Consistency Learning (SCL), a framework designed to learn representations that remain stable across appearance variations while preserving semantic structures shared across species. SCL consists of two complementary components. Foreground-Background Decoupled Spectral Normalization (FDSNorm) stabilizes feature statistics by suppressing environment-induced style variations in a region-aware manner, while Cross-species Neighborhood Modeling (CNM) captures transferable relational structures across species through dynamic feature neighborhoods. Extensive experiments on 11 public animal ReID datasets demonstrate that SCL consistently outperforms state-of-the-art methods under multiple cross-species evaluation protocols and generalizes effectively to previously unseen species and ecological domains. Code is available at <https://github.com/Kemalau/ECCV-26-SCL>.

Keywords: Animal Re-Identification · Domain Generalization

1 Introduction

Animal re-identification (ReID) aims to recognize the same animal individual across different times, viewpoints, and environmental conditions [20, 48]. It plays a critical role in ecological monitoring, wildlife conservation, and long-term population analysis. While remarkable progress has been achieved in person and vehicle ReID [9, 10, 31], these tasks are typically studied in relatively structured environments where objects share consistent geometric layouts and appearance statistics. In contrast, animal ReID operates in open-world scenarios that involve diverse species, complex habitats, and highly varying visual characteristics. Animals from different species exhibit substantial variations in body morphology,

^{*} Equal contribution. [†] Corresponding author.

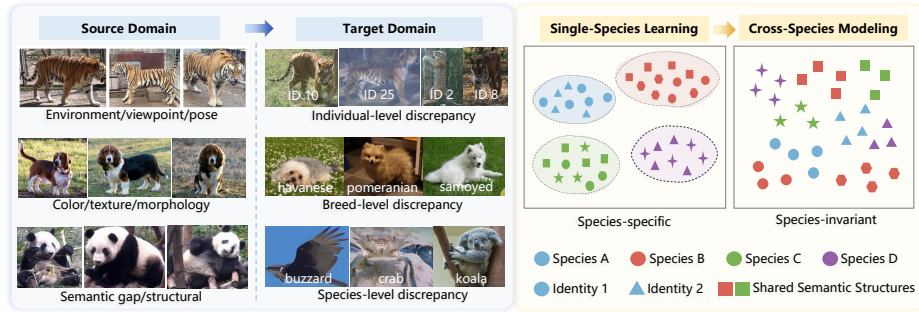


Fig. 1: Multi-level domain discrepancies in animal ReID, including individual, breed, and species level variations beyond viewpoint and pose changes.

texture patterns, and environmental context. These factors introduce severe distribution shifts that go far beyond the viewpoint or illumination changes commonly addressed in human-centered ReID tasks. As a result, developing models that can generalize across species remains a fundamental challenge.

Most existing animal ReID studies focus on a single-species setting, where models are trained and evaluated using data from one specific species [22, 26, 34, 40]. Under this formulation, models can learn discriminative representations tailored to the appearance statistics of that species and achieve strong performance on curated benchmarks. However, such a paradigm limits model reusability in real-world applications. Each new species often requires collecting additional data, annotating identities, and retraining models. To address this limitation, recent work has begun to explore multi-species animal ReID, which aims to train a unified model using data from multiple species and generalize to unseen species or new ecological domains. Some approaches attempt to improve robustness by constructing larger and more diverse datasets [7, 18, 20], while others focus on extracting richer visual cues such as local discriminative patterns [33] or high-frequency information [25]. Although these methods improve recognition accuracy on known species, they mainly enhance instance-level discrimination within species and do not explicitly model how representations should be shared across species. Consequently, the learned embeddings often fail to generalize to unseen species or ecological domains.

In the person ReID community, domain generalization (DG) has been widely studied to address distribution shifts between source and target domains [19, 23, 54]. Existing DG approaches typically aim to learn domain-invariant representations in order to improve cross-domain robustness. Representative techniques include feature disentanglement [29, 57], style normalization [21, 38], and meta-learning [12, 37]. These methods are effective in mitigating variations caused by viewpoint, pose, and illumination changes. However, they implicitly rely on the assumption that objects across domains share consistent semantic structures. For example, human bodies exhibit stable part layouts and similar geometric configurations across domains. This assumption rarely holds in animal ReID.

As illustrated in Figure 1, animals present substantial discrepancies at the individual, breed, and species levels. Cross-species variations in body morphology and semantic structure are significantly larger than those encountered in person ReID. Therefore, existing DG methods struggle to learn transferable representations for multi-species animal ReID.

These limitations indicate that effective cross-species animal ReID requires representations that remain robust to appearance variations while preserving semantic structures shared across species. To this end, we propose a Semantic Consistency Learning (SCL) framework. The key idea is to stabilize appearance statistics while maintaining structural information in the learned representations. Specifically, to mitigate representation instability caused by environmental and species variations, we introduce Foreground–Background Decoupled Spectral Normalization (FDSNorm), a frequency-domain normalization mechanism that decouples foreground and background regions for adaptive spectral modulation. Unlike existing normalization strategies that suppress style variations globally in spatial or spectral spaces, our approach explicitly accounts for semantic differences across regions. We preserve the original phase information while adaptively modulating amplitude spectra in different semantic regions, enabling structure-preserving style control under cross-species domain shifts.

However, stabilizing feature statistics alone does not explicitly model cross-species relationships in the embedding space. To address this limitation, we introduce Cross-species Neighborhood Modeling (CNM), which captures relational structures across species through dynamic neighborhood construction. Specifically, CNM discovers mutual neighbors to form both intra-species and inter-species neighborhoods, enabling the model to learn from relational topology rather than relying solely on appearance similarity. This mechanism encourages the learning of structural regularities shared across species while preserving discriminative capability within each species. In summary, our main contributions are as follows:

- We introduce Foreground–Background Decoupled Spectral Normalization (FDSNorm), a frequency-domain normalization mechanism that decouples foreground and background regions for adaptive spectral modulation. By preserving phase information while modulating amplitude spectra, FDSNorm suppresses appearance variations and stabilizes feature representations under cross-species domain shifts.
- We propose Cross-species Neighborhood Modeling (CNM), which explicitly captures relational structures across species through mutual neighbor discovery. CNM dynamically constructs intra-species and inter-species neighborhoods, enabling the model to learn shared semantic regularities across species while preserving discriminative capability within each species.
- We introduce two complementary evaluation protocols for cross-species animal ReID to enable a comprehensive evaluation of generalization to unseen species. Extensive experiments on 11 public datasets demonstrate consistent improvements over competitive state-of-the-art methods.

2 Related Work

Object ReID. Re-Identification (ReID) has made substantial progress, with a large body of work focused on person and vehicle identification [8, 47, 51, 56]. This has produced many powerful methods, from strong convolutional baselines [32, 44] to more recent transformer-based [15] and vision-language [27] architectures. While these general-purpose frameworks are versatile, their standard implementation requires training a separate, species-specific model for each animal category when adapted for animal ReID. Concurrently, a distinct line of research has emerged that focuses specifically on the challenges of animal ReID. These methods are tailored to specific intra-species challenges, such as identifying livestock by coat patterns [3], re-identifying tigers by their unique stripe patterns [26], using high-frequency supervision for fine-grained details [25], or addressing pose variation with 3D models [55]. These domain-generalizable ReID methods rely on consistent body structures and shared semantic correspondences across domains. In cross-species animal ReID, drastic differences in anatomy and visual patterns break this assumption, limiting their ability to learn transferable representations.

Domain Generalized ReID. DG ReID has a substantial literature, particularly in person ReID [19, 35], aiming to improve generalization under changes in scene, viewpoint, and illumination. Various approaches have been explored to achieve this. One line of work focuses on normalization-based methods [12, 21], which suppress camera/style statistics to preserve identity cues. Other strategies include employing Mixture-of-Experts (MoE) [13, 46] to structure domain variability. These models typically require predefined experts for known domains, making them ill-suited for unseen species. Methods using memory banks [28, 43] stabilize matching across domains, but their efficacy diminishes in animal ReID due to high inter-species similarity and vast intra-species diversity, which can pollute the memory bank. Meta-learning frameworks [4, 54] have also been proposed to simulate test-time shifts. Their limitation lies in the simulation: the meta-tasks often simulate variations in viewpoint or illumination, failing to prepare the model for the drastic object-type shift encountered when generalizing to a new animal species. More recently, data-driven routes like BAU [11] emphasize augmentations, and CLIP-based methods [52, 53] leverage vision-language priors. While powerful, standard CLIP priors often capture species-level semantics rather than fine-grained individual identity, requiring significant adaptation. Training-free approaches such as Pose2ID [50] leverage pose priors at test time, but their reliance on structured human pose limits applicability to animal ReID.

Multi-species ReID. More recently, large-scale animal ReID foundation models such as MegaDescriptor [7] and MiewID [39] have been proposed, leveraging massive community-curated datasets to train unified embedding networks. UniReID [20] specifically adapted CLIP-based architectures to tackle domain generalized animal ReID and contributed the large-scale Wildlife71 dataset. Community benchmarks such as AnimalCLEF [2] further promote evaluation of individual animal recognition at scale. However, these studies mainly focus on constructing large-scale datasets and adopt conventional instance-level met-

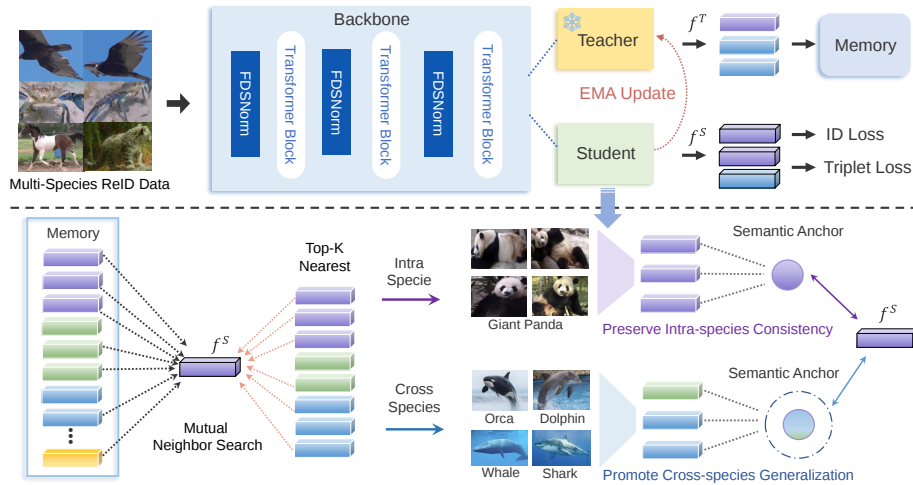


Fig. 2: Overview of the proposed Semantic Consistency Learning framework. The upper part illustrates the Foreground–Background Decoupled Spectral Normalization, while the lower part presents Cross-species Neighborhood Modeling.

ric learning objectives commonly used in ReID. They do not explicitly design mechanisms for generalization to completely unseen species.

3 Method

3.1 Overview

The objective of multi-species animal ReID is to learn a unified representation that preserves individual-level discriminability while generalizing across species with diverse morphologies and ecological environments. However, jointly learning representations from heterogeneous species introduces substantial distributional discrepancies. Differences in texture patterns, body structures, and environmental contexts lead to unstable feature statistics and hinder the formation of a coherent embedding space. As a result, representations learned from different species tend to cluster around species-specific appearance statistics rather than capturing transferable semantic structures.

Furthermore, many existing ReID approaches rely on alignment-based learning strategies that exploit shared visual cues or explicit correspondences across samples. While effective in person ReID or single-species settings, such assumptions rarely hold across species with drastically different anatomies and visual characteristics. Consequently, these methods struggle to establish consistent cross-species representations and often produce fragmented embedding spaces with limited generalization to unseen species. Domain generalization methods developed for person ReID attempt to mitigate distribution shifts by learning

domain-invariant representations through techniques such as feature disentanglement, style normalization, or meta-learning [11, 38]. However, these approaches implicitly assume comparable semantic structures across domains. In the multi-species setting, where anatomical structures and visual semantics differ substantially, this assumption becomes invalid, limiting their ability to capture transferable representations across species.

To address these challenges, we propose the Semantic Consistency Learning (SCL) framework, which promotes stable and transferable feature learning through two complementary components. As illustrated in Figure 2, (1) Foreground–Background Decoupled Spectral Normalization (FDSNorm) introduces a region-aware frequency-domain normalization mechanism that suppresses environment induced style variations while preserving structure-sensitive semantics. (2) Cross-species Neighborhood Modeling (CNM) captures relational regularities within and across species by constructing dynamic feature neighborhoods, enabling the model to align transferable semantics while maintaining intra-species discriminative structure.

3.2 Foreground–Background Decoupled Spectral Normalization

In multi-species generalized ReID, heterogeneous textures, morphologies, and environmental conditions introduce substantial species-dependent biases, leading to unstable feature distributions. Recent studies show that such variations are closely related to the spectral characteristics of visual representations [24, 30]. Specifically, the amplitude spectrum mainly captures style-related factors such as illumination and background statistics, whereas the phase spectrum preserves semantic structure. Consequently, normalization strategies that suppress feature statistics uniformly may inadvertently distort phase-dependent semantics. To address this limitation, we introduce Foreground–Background Decoupled Spectral Normalization (FDSNorm), a frequency-domain normalization mechanism tailored for multi-species ReID. Unlike existing normalization methods that suppress style variations globally in spatial or spectral spaces, FDSNorm explicitly accounts for semantic differences across regions. By preserving the original phase information and adaptively modulating amplitude spectra in foreground and background regions, the proposed mechanism achieves structure-preserving style control under cross-species domain shifts.

Frequency-Domain Feature Normalization. Given an input image $\mathbf{I} \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, the Vision Transformer [14] divides it into N non-overlapping patches of size $P \times P$, each projected into a C -dimensional embedding space. After positional encoding and class-token concatenation, the resulting sequence is processed through L Transformer layers. Let $\mathbf{F}^{(l)} \in \mathbb{R}^{B \times C \times H \times W}$ denote the reshaped token feature map at the l -th layer, where B is the batch size and (H, W) represent the spatial grid reconstructed from tokens.

To adaptively suppress style-induced domain bias while preserving semantic consistency, we employ a learnable frequency-domain normalization strategy. First, we obtain a style-normalized version of the feature map via spatial nor-

malization:

$$\tilde{\mathbf{F}}^{(l)} = \text{Norm}(\mathbf{F}^{(l)}), \quad (1)$$

where $\text{Norm}(\cdot)$ denotes instance normalization applied across spatial dimensions. We then perform Discrete Fourier Transform (DFT) on both the original and normalized features:

$$\begin{aligned} \mathcal{F}_{\text{org}}^{(l)}(u, v) &= \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \mathbf{F}^{(l)}(x, y) e^{-j2\pi(\frac{ux}{H} + \frac{vy}{W})}, \\ \mathcal{F}_{\text{norm}}^{(l)}(u, v) &= \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \tilde{\mathbf{F}}^{(l)}(x, y) e^{-j2\pi(\frac{ux}{H} + \frac{vy}{W})}, \end{aligned} \quad (2)$$

where (x, y) and (u, v) denote the spatial and frequency coordinates, respectively, and $j = \sqrt{-1}$. Each spectral representation is decomposed into amplitude and phase components:

$$\begin{aligned} \mathcal{F}_{\text{org}}^{(l)}(u, v) &= A_{\text{org}}^{(l)}(u, v) e^{j\Phi_{\text{org}}^{(l)}(u, v)}, \\ \mathcal{F}_{\text{norm}}^{(l)}(u, v) &= A_{\text{norm}}^{(l)}(u, v) e^{j\Phi_{\text{norm}}^{(l)}(u, v)}, \end{aligned} \quad (3)$$

where $A^{(l)}(u, v) = |\mathcal{F}^{(l)}(u, v)|$ captures the magnitude spectrum and $\Phi^{(l)}(u, v) = \angle \mathcal{F}^{(l)}(u, v)$ encodes structural information. Following [24], we treat the amplitude spectrum as a style carrier reflecting environmental variations (*e.g.*, illumination and background), while the phase spectrum represents semantic structure that should remain invariant across domains.

Given the CLS token $\mathbf{c}^{(l)}$ and patch tokens $\{\mathbf{t}_i^{(l)}\}_{i=1}^N$ at layer l , we build a soft foreground mask from CLS-to-patch cosine similarity:

$$s_i = \left\langle \frac{\mathbf{t}_i^{(l)}}{\|\mathbf{t}_i^{(l)}\|}, \frac{\mathbf{c}^{(l)}}{\|\mathbf{c}^{(l)}\|} \right\rangle, \quad \tilde{s}_i = \frac{s_i - \min(\mathbf{s})}{\max(\mathbf{s}) - \min(\mathbf{s}) + \epsilon}, \quad (4)$$

$$\tau = \text{Quantile}_{1-r}(\tilde{\mathbf{s}}), \quad m_i = \sigma\left(\frac{\tilde{s}_i - \tau}{T_m}\right), \quad (5)$$

where r is the foreground ratio and T_m is the mask temperature. Reshaping $\{m_i\}$ gives $\mathbf{M}^{(l)} \in [0, 1]^{1 \times H \times W}$. We keep the CLS token unchanged to preserve the global semantic representation, and apply spectral normalization only to patch tokens.

We then apply spatial split before frequency mixing:

$$\mathbf{F}_{\text{fg,org}}^{(l)} = \mathbf{M}^{(l)} \odot \mathbf{F}^{(l)}, \quad \mathbf{F}_{\text{fg,norm}}^{(l)} = \mathbf{M}^{(l)} \odot \tilde{\mathbf{F}}^{(l)}, \quad (6)$$

$$\mathbf{F}_{\text{bg,org}}^{(l)} = (1 - \mathbf{M}^{(l)}) \odot \mathbf{F}^{(l)}, \quad \mathbf{F}_{\text{bg,norm}}^{(l)} = (1 - \mathbf{M}^{(l)}) \odot \tilde{\mathbf{F}}^{(l)}. \quad (7)$$

After spatial decoupling, we independently perform DFT on the foreground and background branches to obtain branch-wise amplitude and phase spectra:

$$\mathcal{F}_{\text{fg},*}^{(l)} = A_{\text{fg},*}^{(l)} e^{j\Phi_{\text{fg},*}^{(l)}}, \quad \mathcal{F}_{\text{bg},*}^{(l)} = A_{\text{bg},*}^{(l)} e^{j\Phi_{\text{bg},*}^{(l)}}, \quad (8)$$

where $*$ $\in$ {org, norm} denotes original and normalized branches. Foreground and background use independent mixing strengths:

$$\alpha_{\text{fg}} = [\text{softmax}(\boldsymbol{\lambda}_{\text{fg}}/T_s)]_0, \quad \alpha_{\text{bg}} = [\text{softmax}(\boldsymbol{\lambda}_{\text{bg}}/T_s)]_0, \quad (9)$$

where $\boldsymbol{\lambda}_{\text{fg}}, \boldsymbol{\lambda}_{\text{bg}} \in \mathbb{R}^2$ are learnable two-dimensional parameter vectors initialized to zeros, T_s is the temperature parameter in softmax, and $[\cdot]_0$ selects the first element as the normalized mixing weight.

$$\hat{A}_{\text{fg}}^{(l)} = \alpha_{\text{fg}} A_{\text{fg}, \text{norm}}^{(l)} + (1 - \alpha_{\text{fg}}) A_{\text{fg}, \text{org}}^{(l)}, \quad \hat{A}_{\text{bg}}^{(l)} = \alpha_{\text{bg}} A_{\text{bg}, \text{norm}}^{(l)} + (1 - \alpha_{\text{bg}}) A_{\text{bg}, \text{org}}^{(l)}. \quad (10)$$

In this work, we use the unconstrained variant, i.e., no explicit ordering constraint is imposed between α_{fg} and α_{bg} , allowing the network to adaptively discover when background regions require stronger style suppression. Reconstruction preserves branch-wise original phase:

$$\hat{\mathbf{F}}_{\text{fg}}^{(l)} = \mathcal{F}^{-1} \left(\hat{A}_{\text{fg}}^{(l)} e^{j\Phi_{\text{fg}, \text{org}}^{(l)}} \right), \quad \hat{\mathbf{F}}_{\text{bg}}^{(l)} = \mathcal{F}^{-1} \left(\hat{A}_{\text{bg}}^{(l)} e^{j\Phi_{\text{bg}, \text{org}}^{(l)}} \right), \quad (11)$$

$$\hat{\mathbf{F}}^{(l)} = \hat{\mathbf{F}}_{\text{fg}}^{(l)} + \hat{\mathbf{F}}_{\text{bg}}^{(l)}. \quad (12)$$

Through empirical analysis, the FDSNorm module is inserted at multiple Transformer depths ($l \in \{0, 4, 8\}$), corresponding to the patch-embedding output and intermediate blocks, thereby forming a progressive de-stylization pipeline. Specifically, shallow layers mainly clean pixel-level domain shifts (e.g., illumination and color temperature); middle layers align structure-level shifts (e.g., parts and shape); and deeper layers compensate for residual shifts leaked by residual connections. In parallel, as the CLS token passes through more attention layers, its semantic awareness becomes stronger. This allows the foreground mask to be refined from coarse to fine, yielding a coordinated progression between de-stylization strength and mask quality. In addition, the foreground ratio r and branch-wise mixing strengths (α_{fg} and α_{bg}) are dynamically learnable parameters, while the mask temperature T_m remains fixed.

Temporal Semantic Distillation. While normalization suppresses style-related instability within the backbone, temporal inconsistency may still arise from noisy or domain-biased updates during optimization. To stabilize the evolution of semantic representations, we employ a teacher-student framework in which the teacher network maintains an exponential moving average (EMA) of the student parameters:

$$\boldsymbol{\theta}_t \leftarrow \mu \boldsymbol{\theta}_t + (1 - \mu) \boldsymbol{\theta}_s, \quad (13)$$

where $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_s$ denote the teacher and student parameters, respectively, and $\mu \in (0, 1)$ is a momentum coefficient. The teacher network provides temporally smoothed features that serve as stable semantic anchors for subsequent neighborhood consistency learning, effectively distilling long-term structural knowledge into the student without introducing additional supervision.

3.3 Cross-species Neighborhood Modeling

In multi-species ReID, each species forms a visually coherent cluster within the feature space, yet these clusters remain topologically isolated due to the absence of shared semantic anchors. This leads to semantic fragmentation: features are discriminative within species but unaligned across them, hindering transfer to unseen species. Conventional ReID metric learning objectives such as triplet loss [16] rely on instance-level correspondences and fail to exploit the latent relational regularities shared across species. To bridge this gap, we propose Cross-species Neighborhood Modeling, which constructs dynamic relational structures using a teacher-maintained memory to jointly enforce intra-species consistency and cross-species semantic connectivity.

Dynamic Memory Construction. Let $\mathbf{f}_t$ and $\mathbf{z}_s$ denote the global features from the teacher and student networks, respectively. We maintain a feature memory queue $\mathcal{M} = \{(\mathbf{f}_i, sp_i, y_i)\}_{i=1}^{|\mathcal{M}|}$, where each entry consists of a normalized teacher feature $\mathbf{f}_i$, its species label sp_i , and identity label y_i . After each iteration, newly computed teacher features are enqueued, while the oldest entries are dequeued to maintain a fixed capacity, ensuring $\mathcal{M}$ captures long-term semantic structure across species. For each student feature $\mathbf{z}_s$ with species label sp , the cosine similarity to all memory entries is computed as:

$$\text{sim}(\mathbf{z}_s, \mathbf{f}_i) = \frac{\mathbf{z}_s^\top \mathbf{f}_i}{\|\mathbf{z}_s\| \|\mathbf{f}_i\|}. \quad (14)$$

Mutual Neighborhood Search. Let $\mathcal{P} = \{\mathbf{f}_t\} \cup \mathcal{M}$ denote the union of the current-batch teacher features and the memory queue. Each anchor $\mathbf{z}_s$ retrieves two types of neighborhoods from $\mathcal{P}$: an *intra-species neighborhood* $\mathcal{N}_{intra}(\mathbf{z}_s)$ consisting of top- K_1 nearest neighbors whose species label matches the anchor, i.e., $sp_i = sp$, and a *cross-species neighborhood* $\mathcal{N}_{cross}(\mathbf{z}_s)$ of top- K_1 nearest neighbors where $sp_i \neq sp$. To suppress incidental correlations in the cross-species neighborhood, we adopt a reciprocal nearest-neighbor rule: for an anchor A and a candidate neighbor B retrieved from the top- K_1 cross-species list of A , B is retained only if A also appears in the top- K_2 cross-species nearest list of B within $\mathcal{P}$. Here, K_1 determines the size of the candidate neighborhood for retrieval, whereas K_2 ($K_2 \leq K_1$) controls the stringency of the reciprocal verification: a larger K_1 broadens the candidate pool, while a smaller K_2 retains only strongly mutual cross-species pairs. This reciprocal filtering is applied solely to the cross-species neighborhood. This bidirectional filtering yields more stable semantic neighborhoods that reflect intrinsic relational similarity across species. The neighborhood centers are defined as:

$$\mathbf{c}_{intra} = \frac{1}{|\mathcal{N}_{intra}|} \sum_{\mathbf{f}_i \in \mathcal{N}_{intra}} \mathbf{f}_i, \quad \mathbf{c}_{cross} = \frac{1}{|\mathcal{N}_{cross}|} \sum_{\mathbf{f}_i \in \mathcal{N}_{cross}} \mathbf{f}_i. \quad (15)$$

Loss Formulation. Cross-species neighborhood modeling jointly optimizes two complementary objectives: intra-species compactness and cross-species relational consistency. 1) *Intra-species Compactness.* For each sample, we align

the student feature with its intra-species neighborhood center to enforce intra-species compactness, where the center is computed over same-species neighbors regardless of identity:

$$\mathcal{L}_{\text{intra}} = 1 - \frac{1}{B} \sum_{i=1}^B \text{sim}(\mathbf{z}_i, \mathbf{c}_{\text{intra}}^{(i)}), \quad (16)$$

where B denotes the batch size and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. 2) *Cross-species Relational Constraint*. To bridge gaps across species, we introduce a margin-based relational constraint:

$$\mathcal{L}_{\text{cross}} = \frac{1}{B} \sum_{i=1}^B \max\left(0, m - \text{sim}(\mathbf{z}_i, \mathbf{c}_{\text{cross}}^{(i)})\right), \quad (17)$$

where m is a margin. Although cross-species centers provide transferable semantic cues, overly strong attraction toward them may pull features of different identities too close, impairing identity discrimination. The hinge thus acts as a bounded attraction: each feature is pulled toward its cross-species relational center only until the similarity reaches the target margin m , after which no gradient is applied. This injects a controlled level of cross-species connectivity (i.e., $\text{sim} \geq m$), while the small margin and the saturation of the hinge prevent features from being driven into full alignment, thereby preserving identity-level discrimination. This formulation softly aligns inter-species manifolds without collapsing structural diversity. The overall CNM objective is:

$$\mathcal{L}_{\text{CNM}} = \lambda_{\text{intra}} \mathcal{L}_{\text{intra}} + \lambda_{\text{cross}} \mathcal{L}_{\text{cross}}. \quad (18)$$

In summary, our learning objective is the total loss $\mathcal{L}$, formulated as a weighted sum of the identification loss, the triplet loss, and the CNM loss.

$$\mathcal{L} = \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{tri}} \mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{CNM}}. \quad (19)$$

Memory Update. After each iteration, normalized teacher features with their species and identity labels are added to the memory:

$$\mathcal{M} \leftarrow \text{enqueue}(\text{Norm}(\mathbf{f}_t), sp_t, y_t), \quad \text{dequeue oldest}. \quad (20)$$

This online update maintains a temporally smoothed and semantically consistent teacher space, providing robust relational guidance for the student network.

4 Evaluation Protocol

4.1 Datasets and Evaluation Protocols

Datasets and Splits. To comprehensively evaluate cross-species generalization, we conduct experiments on 11 publicly available animal ReID datasets

Table 1: Dataset Statistics

Dataset	#Image	#ID	Species
PetFace [42]	115,708	55,686	13
Wildlife71 [20]	108,096	1,924	67
iPanda-50 [45]	6,874	50	1
ELPephants [22]	2,078	276	1
SealID [34]	2,080	61	1
GZGC [41]	4,948	1,762	2
ATRW [26]	2,950	135	1
HyenaID2022 [5]	3,129	256	1
LeopardID2022 [6]	6,806	430	1
SeaTurtleID2022 [1]	8,729	438	1
WhaleSharkID [17]	7,693	543	1

Table 2: Evaluation Protocols

Protocol	Training Data	Testing Data
1	Wildlife71	PetFace, iPanda-50, ELPephants, SealID, SeaTurtleID2022, ATRW, HyenaID2022, LeopardID2022, GZGC, WhaleSharkID,
2	iPanda-50+ELPephants +LeopardID2022+GZGC +ATRW+HyenaID2022 +SealID+SeaTurtleID2022 +WhaleSharkID	PetFace, Wildlife71

covering diverse habitats and species morphologies. This diverse dataset collection provides a challenging evaluation setting with substantial variations in visual appearance and environmental conditions. (1) *Wildlife71 Dataset* [20]. This large-scale benchmark comprises 71 species. Following its official split protocol, we use the predefined 67 seen species for training. (2) *PetFace Dataset* [42]. This dataset contains facial images from 13 domestic animal species, and we test the model separately on each species. (3) *Nine public datasets*. We further evaluate on iPanda-50 [45], ELPephants [22], SealID [34], GZGC (zebra and giraffe domains) [41], WhaleSharkID [17], ATRW [26], HyenaID2022 [5], LeopardID2022 [6], and SeaTurtleID2022 [1]. As shown in Tab. 2, we design two complementary evaluation protocols to simulate different cross-species generalization scenarios. Both protocols enforce a fully open-set setting in which identities and species in the test set are disjoint from those used during training, ensuring that the evaluation reflects genuine cross-species generalization rather than dataset-specific overlap. Since existing animal ReID datasets adopt inconsistent split strategies or lack official training/testing partitions, we use the full datasets under both protocols to maintain consistent evaluation conditions. The only exception is PetFace, for which we follow the official split. This unified data usage ensures fair and reproducible comparisons across different methods.

Evaluation Metrics. We adopt Cumulative Matching Characteristics (CMC) at Rank-1 and mean Average Precision (mAP) as standard metrics. Since most animal ReID datasets lack explicit camera annotations, we include all valid gallery matches in evaluation rather than only cross-camera ones.

5 Experiments

5.1 Implementation Details

All experiments were conducted on four NVIDIA 4090 GPUs using PyTorch. We employ a ViT [14] pre-trained on ImageNet-1K as the backbone; unless a method has specific architectural constraints, all baselines share the same backbone for fair comparison. Input images are resized to 256×256 with patch size 16×16 . Training augmentations include random horizontal flipping (50%) and 10-pixel padding. For CNM hyperparameters, we set $K_1=8$, $K_2=3$, margin = 0.2, and memory size = 4096. The model is trained for 60 epochs with SGD (initial lr

Table 3: Protocol-1 results on 10 unseen domains. The symbol † denotes results obtained from re-implemented versions of the corresponding methods.

Method	Venue	ELPephants [22]		SealID [34]		GZGC (zebra) [41]		ATRW [26]		GZGC (giraffe) [41]	
		Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
Base [15]	ICCV 2021	31.9	7.9	79.4	24.8	12.6	<u>7.1</u>	96.2	58.9	21.0	25.3
TransReID [15]	ICCV 2021	32.3	8.1	78.6	23.3	12.2	6.9	<u>96.5</u>	59.1	20.0	25.6
META [46]	ECCV 2022	26.4	5.8	79.0	20.2	8.0	3.2	95.7	51.6	13.5	9.3
CLIP [27]	AAAI 2023	28.9	6.8	77.6	20.8	11.9	6.9	95.6	58.1	22.9	25.7
PartAware [36]	ICCV 2023	32.0	7.9	79.7	24.9	12.4	<u>7.1</u>	96.2	58.9	20.6	25.4
UniReID† [20]	NeurIPS 2023	25.2	6.1	79.4	23.6	12.0	6.8	96.1	55.9	24.7	26.0
BAU [11]	NeurIPS 2024	13.2	3.7	<u>80.4</u>	30.8	10.1	5.1	90.6	49.9	20.6	22.2
AdaFreq [25]	ECCV 2024	<u>33.3</u>	<u>8.4</u>	78.6	22.8	12.1	6.8	96.4	58.0	21.6	26.0
ReNorm [38]	ECCV 2024	20.7	5.0	78.6	23.6	10.6	5.5	96.3	56.1	<u>23.5</u>	24.1
Megadescriptor† [7]	WACV 2024	32.0	8.2	76.5	22.1	<u>12.9</u>	7.0	95.9	57.4	<u>23.5</u>	<u>26.8</u>
MiewID† [39]	CoRR 2024	16.2	4.3	72.9	19.9	9.1	4.7	94.6	51.2	17.9	20.8
CLIP-FGDI [53]	TIFS 2025	16.2	4.3	72.4	22.2	10.6	5.6	88.5	45.8	21.3	24.5
ARBase† [18]	ICCV 2025	25.4	5.5	77.9	21.4	10.0	3.9	96.1	53.8	12.7	9.1
Ours	-	34.4	9.0	81.5	<u>25.6</u>	13.1	7.8	98.0	60.3	22.9	27.3
Method	Venue	iPanda-50 [45]		HyenaID2022 [5]		LeopardID2022 [6]		SeaTurtleID2022 [1]		WhaleSharkID [17]	
		Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
Base [15]	ICCV 2021	<u>91.8</u>	<u>13.1</u>	60.4	22.6	76.0	17.5	46.7	7.3	37.4	6.9
TransReID [15]	ICCV 2021	<u>91.8</u>	<u>13.1</u>	<u>62.9</u>	<u>23.7</u>	<u>78.0</u>	<u>18.4</u>	51.2	8.1	42.2	7.8
META [46]	ECCV 2022	87.9	11.4	50.8	15.5	64.0	12.0	39.3	5.0	35.1	5.4
CLIP [27]	AAAI 2023	90.2	12.9	59.3	21.3	74.7	16.5	45.9	6.9	38.8	7.0
PartAware [36]	ICCV 2023	91.7	<u>13.1</u>	61.8	23.0	77.6	18.3	51.1	8.0	40.2	7.3
UniReID† [20]	NeurIPS 2023	88.8	12.9	59.2	21.2	74.6	16.4	45.8	6.9	30.9	5.2
BAU [11]	NeurIPS 2024	81.9	10.2	49.8	15.0	65.6	11.9	51.2	7.3	22.4	3.6
AdaFreq [25]	ECCV 2024	90.6	12.8	59.9	23.0	76.5	17.4	53.1	8.4	<u>42.5</u>	<u>8.0</u>
ReNorm [38]	ECCV 2024	82.1	9.9	59.9	21.4	72.5	13.6	60.4	<u>9.3</u>	32.9	5.4
Megadescriptor† [7]	WACV 2024	89.6	13.0	61.3	23.0	77.1	18.0	45.3	6.9	39.1	7.1
MiewID† [39]	CoRR 2024	74.2	8.9	53.0	17.1	66.3	12.0	42.0	5.5	27.2	4.5
CLIP-FGDI [53]	TIFS 2025	59.9	6.9	47.8	15.2	65.8	13.1	20.2	3.0	25.9	4.9
ARBase† [18]	ICCV 2025	88.2	11.7	51.1	15.6	64.3	12.1	39.6	5.1	34.4	5.8
Ours	-	92.5	13.8	64.0	24.8	78.9	19.4	<u>58.8</u>	10.0	45.4	8.9

0.004, cosine decay) and a total batch size of 128 (8 identities $\times$ 4 images per GPU $\times$ 4 GPUs). Both CNM and FDSNorm use a 3-epoch warm-up. At test time, only original features are used for distance computation.

5.2 Comparison with SOTA Methods

Table 3 and Fig. 3 report Protocol-1 results; Table 4 reports Protocol-2 results on Wildlife71. We compare four categories of methods: (1) *General ReID* (e.g., TransReID, CLIP-ReID). These methods employ strong architectures and often achieve competitive results, yet they remain species-specific and do not capture fine-grained cues that transfer across species. (2) *Animal ReID* (e.g., UniReID, AdaFreq, ARBase). These methods enhance within-species discrimination but lack cross-species structural alignment, leading to negative transfer on unseen domains. Foundation models such as MegaDescriptor [7] and MiewID [39] benefit from large-scale data but still optimize instance-level metrics without explicit cross-species alignment. (3) *DG person ReID* (e.g., ReNorm, META). These methods improve statistical invariance and handle style variations, but do not address the semantic shifts induced by species changes and therefore underperform on unknown species. (4) *DG animal ReID* (e.g., UniReID). UniReID relies on dataset-specific textual descriptions; when applied to a face-centric dataset such as PetFace, its generic whole-body description becomes mismatched, leading to a substantial performance drop. Overall, existing methods may excel on

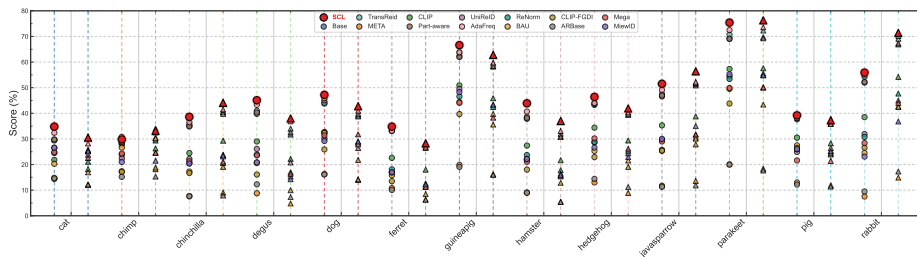


Fig. 3: Protocol-1 results on PetFace (13 species). Circles: mAP; triangles: Rank-1.

Table 4: Protocol-2 results on Wildlife71.

Method	mAP	mINP	Rank1
Base [15]	90.1	75.1	96.6
TransReID [15]	<u>91.6</u>	72.6	96.5
META [46]	83.7	51.0	96.5
CLIP [27]	86.4	54.1	96.7
PartAware [36]	90.1	70.3	96.8
UniReID [†] [20]	84.4	53.6	96.6
BAU [11]	86.7	56.4	96.2
AdaFreq [25]	91.2	79.2	97.3
ReNorm [38]	71.9	25.3	95.6
Megadescriptor [†] [7]	87.3	60.6	97.3
MiewID [†] [39]	82.5	43.3	96.1
CLIP-FGDI [53]	73.6	27.7	<u>97.4</u>
ARBase [†] [18]	86.4	59.4	96.8
Ours	93.8	<u>78.5</u>	97.6

Table 5: Ablation on CNM memory.

Setting	Same-ID (%)	mAP (%)	Rank1 (%)
k-NN	65.12	18.36	53.53
CNM w/o Mem.	78.24	19.33	56.01
Ours (CNM)	89.36	20.69	58.75

Table 6: Layer replacement and EMA ablation results.

Setting	mAP (%)	Rank1 (%)
$l \in \{0, 4, 8\}$	20.69	58.75
$l \in \{2, 4, 6\}$	19.77	56.58
$l \in \{0, \dots, 11\}$	19.37	54.53
$l \in \{0, 4, 8\}$ w/o EMA	20.14	57.53

selected species yet degrade significantly when the species changes. In contrast, our approach learns a species-agnostic model without any target-domain adaptation and achieves leading performance across diverse unseen species.

5.3 Ablation Experiments

Effectiveness of Each Component. Ablation results in Table 7 confirm the complementary contributions of FDSNorm and CNM to cross-species generalization. FDSNorm stabilizes style statistics to yield domain-robust features, while CNM promotes intra-species cohesion and cross-species semantic connectivity. Table 5 further shows that CNM with memory significantly improves nearest-neighbor identity purity over vanilla k-NN; crucially, removing the memory queue leads to a noticeable drop in retrieval purity, proving it essential for providing stable, long-term relational anchors. Table 6 reveals that sparse layer placement ($l \in \{0, 4, 8\}$) is optimal for FDSNorm—applying it too densely or across all layers over-suppresses structural semantics. Removing EMA further destabilizes feature evolution, confirming the necessity of temporal smoothing.

Foreground-Only Analysis. We use MVANet [49] to segment foreground regions and construct foreground-only inputs. Table 8 further evaluates robustness when background cues are largely removed. All methods degrade under foreground-only input, indicating that context still contributes to matching. Nevertheless, our method achieves the strongest absolute foreground performance

Table 7: Ablation study of different components on ten wildlife datasets under Protocol-1. Intra: Intra-species consistency; Cross: Cross-species consistency.

ID	FDSNorm	Intra	Cross	ELPephants [22]		SealID [34]		Wildlife71 [20]		ATRW [26]		GZGC (giraffe) [41]	
				Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
(a)	–	–	–	31.9	7.9	79.4	24.8	96.6	90.1	96.2	58.9	21.0	25.3
(b)	✓	–	–	32.8	8.4	80.9	25.3	97.0	91.4	97.1	59.3	22.4	26.8
(c)	–	✓	–	33.1	8.5	80.8	24.8	97.2	92.3	97.7	59.6	22.2	26.0
(d)	–	✓	✓	33.3	8.6	81.2	25.0	97.4	93.0	97.6	60.0	22.5	27.0
(e)	✓	✓	✓	34.2	9.0	81.3	25.6	97.6	93.8	97.8	60.3	22.7	27.3

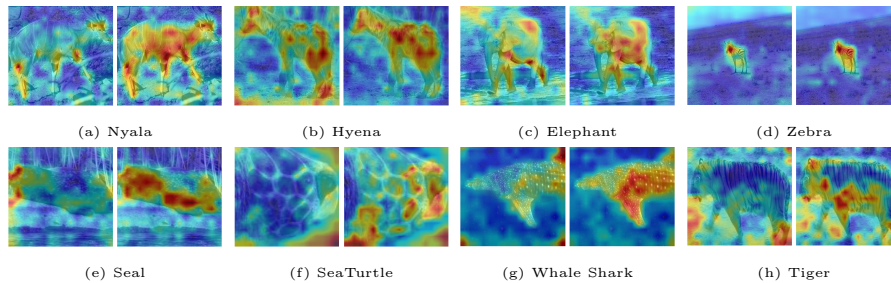
ID	FDSNorm	Intra	Cross	PetFace [42]		HyenaID2022 [5]		LeopardID2022 [6]		SeaTurtleID2022 [1]		WhaleSharkID [17]	
				Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
(a)	–	–	–	41.9	39.8	60.4	22.6	76.0	17.5	46.7	7.3	37.4	6.9
(b)	✓	–	–	43.6	42.8	62.6	24.1	77.8	18.7	55.5	9.2	44.0	8.4
(c)	–	✓	–	44.5	44.8	63.2	23.9	78.2	18.6	56.7	8.9	44.3	8.5
(d)	–	✓	✓	45.3	45.9	63.5	24.2	78.5	18.9	57.4	9.3	44.7	8.7
(e)	✓	✓	✓	46.2	46.9	63.8	24.8	78.7	19.4	58.6	10.0	45.2	8.9

Table 8: Original vs. foreground-only input under Protocol-2.

Model	Wildlife71 [20]			Wildlife71 (Foreground) [20]			Drop		
	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1
Base [15]	90.1	75.1	96.6	71.2	28.6	95.8	18.9	46.5	0.8
Megadescriptor [7]	87.3	60.6	97.3	69.8	28.7	95.2	17.5	31.9	2.1
SCL (Ours)	93.8	78.5	97.6	79.0	45.8	96.1	14.8	32.7	1.5

and the smallest mAP drop, showing that SCL relies more on transferable identity structure than on scene-specific shortcuts.

Visualization Analysis. Figure 4 shows eight subfigures: Compared with Base, SCL consistently shifts attention toward semantically meaningful animal regions (*e.g.*, torso contours, texture-rich parts, and limbs) and reduces diffuse activation on irrelevant background areas. This trend is stable across elephant, zebra, seal, sea turtle, whale shark, hyena, nyala and tiger, indicating stronger structure-focused consistency under large appearance and habitat variations.

**Fig. 4:** Last layer activation-map comparison between Base (left) and SCL (right).

Feature Distribution Analysis. Figure 5 shows what CNM learns in the embedding space. For this experiment, Base denotes the plain ViT baseline. Both the baseline and our model are trained on Wildlife71, and we randomly sample

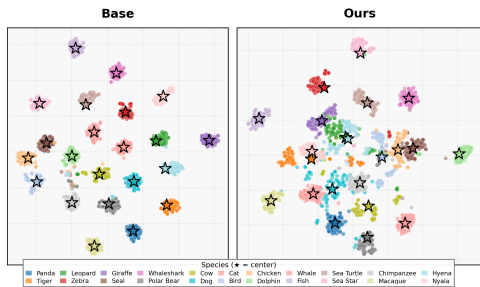


Fig. 5: Embedding space visualization of samples from diverse species.

Table 9: Ratio of inter-ID distance to intra-ID distances on Protocol-1. Higher is better.

Dataset	SCL	Mega	MiewID
iPanda-50 [45]	1.265	1.145	1.062
ELPephants [22]	1.220	1.122	1.049
SealID [34]	1.379	1.263	1.134
GZGC-Zebra [41]	1.242	1.106	1.048
ATRW [26]	2.816	2.106	1.269
GZGC-Giraffe [41]	1.838	1.726	1.326
HyenaID2022 [5]	1.741	1.412	1.126
LeopardID2022 [6]	1.456	1.272	1.091
SeaTurtleID2022 [1]	1.311	1.234	1.108
WhaleSharkID [17]	1.196	1.064	1.039

100 instances for each of the 22 species for visualization. The Base representation exhibits clear species-wise separation, where samples from different species form isolated clusters. In contrast, CNM produces a more mixed embedding distribution across species while achieving better ReID performance, suggesting that it reduces species-specific clustering tendencies and learns more effective cross-species representations.

To further examine representation quality under unseen domains, we report the ratio of inter-ID distance to intra-ID distance in Table 9. A higher ratio indicates a better clustering structure, namely tighter intra-identity compactness together with clearer inter-identity separation. Compared with strong animal foundation models such as Megadescriptor and MiewID, SCL consistently achieves higher ratios on all ten unseen datasets. This result supports the same conclusion as our qualitative visualizations: our method improves cluster compactness while preserving clearer boundaries between identities, thereby mitigating species-isolated feature fragmentation.

6 Conclusion

This paper introduced the Semantic Consistency Learning framework for cross-species generalization in animal ReID. Through Foreground–Background Decoupled Spectral Normalization, SCL suppresses feature instability caused by environmental variations while aggregating long-range semantic cues shared across species. Cross-species Neighborhood Modeling further shifts the objective from instance-level discrimination to relational semantic understanding, enabling the model to capture structural regularities that generalize across species and yield a unified, transferable embedding space. Promising future directions include extending SCL to fully open-set ReID, integrating large-scale ecological foundation models, and enabling adaptive deployment across diverse platforms such as UAVs and camera traps. We hope this work encourages the community to move beyond species-specific ReID and toward general, cross-species visual understanding that can support large-scale biodiversity monitoring in real-world ecosystems.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China under Grant T2541022.

References

1. Adam, L., Čermák, V., Papafitsoros, K., Pícek, L.: Seaturtleid2022: A long-span dataset for reliable sea turtle re-identification. In: WACV. pp. 7146–7156 (2024)
2. Adam, L., Papafitsoros, K., Kovář, R., Čermák, V., Pícek, L.: Overview of AnimalCLEF 2025: Recognizing individual animals in images. In: Working Notes of CLEF (2025)
3. Andrew, W., Hannuna, S., Campbell, N., Burghardt, T.: Friesian: A novel dataset and a two-stage deep learning framework for cattle re-identification. In: ICIP. pp. 3103–3107 (2021)
4. Bai, Y., Jiao, J., Ce, W., Liu, J., Lou, Y., Feng, X., Duan, L.Y.: Person30k: A dual-meta generalization network for person re-identification. In: CVPR (2021)
5. Botswana Predator Conservation Trust: Panthera pardus csv custom export (2022), <https://lila.science/datasets/hyena-id-2022>, retrieved from African Carnivore Wildbook. Dataset export dated 2022-04-28. Accessed: June 29, 2026
6. Botswana Predator Conservation Trust: Panthera pardus csv custom export (2022), <https://lila.science/datasets/leopard-id-2022>, retrieved from African Carnivore Wildbook. Dataset export dated 2022-04-28. Accessed: June 29, 2026
7. Čermák, V., Pícek, L., Adam, L., Papafitsoros, K.: Wildlifedatasets: An open-source toolkit for animal re-identification. In: WACV. pp. 5953–5963 (2024)
8. Chen, S., Wu, Y., Ye, M.: Object-generalized re-identification: A step towards universal instance perception. In: CVPR. pp. 18481–18491 (2026)
9. Chen, S., Ye, M., Du, B.: Rotation invariant transformer for recognizing object in uavs. In: ACM MM. pp. 2565–2574 (2022)
10. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE TPAMI **40**(2), 392–408 (2017)
11. Cho, Y., Kim, J., Kim, W.J., Jung, J., eui Yoon, S.: Generalizable person re-identification via balancing alignment and uniformity. In: NeurIPS (2024)
12. Choi, S., Kim, T., Jeong, M., Park, H., Kim, C.: Meta batch-instance normalization for generalizable person re-identification. In: CVPR. pp. 3425–3435 (2021)
13. Dai, Y., Li, X., Liu, J., Tong, Z., Duan, L.Y.: Generalizable person re-identification with relevance-aware mixture of experts. In: CVPR. pp. 16145–16154 (2021)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2020)
15. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: ICCV. pp. 15013–15022 (2021)
16. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
17. Holmberg, J., Norman, B., Arzoumanian, Z.: Estimating population size, structure, and residency time for whale sharks rhincodon typus through collaborative photo-identification. Endangered Species Research **7**(1), 39–53 (2009)
18. Hou, S., Huang, P., Wang, Z., Liu, Y., Li, Z., Zhang, M., Huang, Y.: Openanimals: Revisiting person re-identification for animals towards better generalization. ICCV (2024)

19. Jiang, Y., Cheng, X., Yu, H., Liu, X., Chen, H., Zhao, G.: Domain shifting: A generalized solution for heterogeneous cross-modality person re-identification. In: ECCV. pp. 289–306 (2024)
20. Jiao, B., Liu, L., Gao, L., Wu, R., Lin, G., Wang, P., Zhang, Y.: Toward re-identifying any animal. *NeurIPS* **36**, 40042–40053 (2023)
21. Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification. In: CVPR. pp. 3143–3152 (2020)
22. Korschens, M., Denzler, J.: Elpephants: A fine-grained dataset for elephant re-identification. In: ICCVW. pp. 0–0 (2019)
23. Lee, H., Park, J., Oh, J., Eom, C.: Domain generalization for person re-identification: A survey towards domain-agnostic person matching. *Neurocomputing* p. 130763 (2025)
24. Lee, S., Bae, J., Kim, H.Y.: Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In: CVPR. pp. 11776–11785 (2023)
25. Li, C., Chen, S., Ye, M.: Adaptive high-frequency transformer for diverse wildlife re-identification. In: ECCV. pp. 296–313. Springer (2024)
26. Li, S., Li, J.W., Wu, C., Zheng, W.S.: ATRW: A benchmark for amur tiger re-identification in the wild. In: ACM MM. pp. 1297–1305 (2021)
27. Li, S., Sun, L., Li, Q.: Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In: AAAI. vol. 37, pp. 1405–1413 (2023)
28. Liao, S., Shao, L.: Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In: ECCV. pp. 456–474. Springer (2020)
29. Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., Cai, J.: Domain-invariant disentangled network for generalizable object detection. In: ICCV. pp. 8771–8780 (2021)
30. Lin, S., Zhang, Z., Huang, Z., Lu, Y., Lan, C., Chu, P., You, Q., Wang, J., Liu, Z., Parulkar, A., et al.: Deep frequency filtering for domain generalization. In: CVPR. pp. 11797–11807 (2023)
31. Lou, Y., Bai, Y., Liu, J., Wang, S., Duan, L.: Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In: CVPR. pp. 3235–3243 (2019)
32. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPRW. pp. 122–130 (2019)
33. Nepovninnykh, E., Chelak, I., Eerola, T., Immonen, V., Kälviäinen, H., Kholiavchenko, M., Stewart, C.V.: Species-agnostic patterned animal re-identification by aggregating deep local features. *IJCV* **132**(9), 4003–4018 (2024)
34. Nepovninnykh, E., Eerola, T., Biard, V., Mutka, P., Niemi, M., Kunnasranta, M., Kälviäinen, H.: Sealid: Saimaa ringed seal re-identification dataset. *Sensors* **22**(19), 7602 (2022)
35. Nguyen, V.D., Mirza, S., Zakeri, A., Gupta, A., Khaldi, K., Aloui, R., Mantini, P., Shah, S.K., Merchant, F.: Tackling domain shifts in person re-identification: A survey and analysis. In: CVPRW. pp. 4149–4159 (2024)
36. Ni, H., Li, Y., Gao, L., Shen, H.T., Song, J.: Part-aware transformer for generalizable person re-identification. In: ICCV. pp. 11280–11289 (2023)
37. Ni, H., Song, J., Luo, X., Zheng, F., Li, W., Shen, H.T.: Meta distribution alignment for generalizable person re-identification. In: CVPR. pp. 2487–2496 (2022)
38. Nie, R., Ding, J., Zhou, X., Li, X.: Rethinking normalization layers for domain generalizable person re-identification. In: ECCV. pp. 267–284. Springer (2024)
39. Otarashvili, L., Subramanian, T., Holmberg, J., Levenson, J.J., Stewart, C.V.: Multispecies animal re-id using a large community-curated dataset. *CoRR* (2024)

40. Papafitsoros, K., Adam, L., Čermák, V., Pícek, L.: Seaturtleid: A novel long-span dataset highlighting the importance of timestamps in wildlife re-identification. arXiv preprint arXiv:2211.10307 (2022)
41. Parham, J., Crall, J., Stewart, C., Berger-Wolf, T., Rubenstein, D.I.: Animal population censusing at scale with citizen science and photographic identification. In: AAAI (2017)
42. Shinoda, R., Shiohara, K.: Petface: A large-scale dataset and benchmark for animal identification. In: ECCV. pp. 19–36. Springer (2025)
43. Song, J., Yang, Y., Li, Y.Z., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: CVPR. pp. 718–727 (2019)
44. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM MM. pp. 274–282 (2018)
45. Wang, L., Ding, R., Zhai, Y., Zhang, Q., Tang, W., Zheng, N., Hua, G.: Giant panda identification. IEEE TIP **30**, 2837–2849 (2021)
46. Xu, B., Liang, J., He, L., Sun, Z.: Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In: ECCV. pp. 372–388. Springer (2022)
47. Yang, Z., Wu, D., Wu, C., Lin, Z., Gu, J., Wang, W.: A pedestrian is worth one prompt: Towards language guidance person re-identification. In: CVPR. pp. 17343–17353 (2024)
48. Ye, M., Chen, S., Li, C., Zheng, W.S., Crandall, D., Du, B.: Transformer for object re-identification: A survey. arXiv preprint arXiv:2401.06960 (2024)
49. Yu, Q., Zhao, X., Pang, Y., Zhang, L., Lu, H.: Multi-view aggregation network for dichotomous image segmentation. In: CVPR. pp. 3921–3930 (2024)
50. Yuan, C., Zhang, G., Ma, C., Zhang, T., Niu, G.: From poses to identity: Training-free person re-identification via feature centralization. In: CVPR (2025)
51. Zhang, Q., Wang, L., Patel, V.M., Xie, X., Lai, J.: View-decoupled transformer for person re-identification under aerial-ground camera network. In: CVPR. pp. 22000–22009 (2024)
52. Zhao, H., Qi, L., Geng, X.: Clip-dfgs: A hard sample mining method for clip in generalizable person re-identification. ACM T MULTIM COMP **21**(1), 1–20 (2024)
53. Zhao, H., Qi, L., Geng, X.: Cilp-fgdi: Exploiting vision-language model for generalizable person re-identification. IEEE TIFS (2025)
54. Zhao, Y., Zhang, J., et al.: Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In: CVPR (2021)
55. Zheng, Z., Zheng, Z., Zheng, W.S., Tao, D.: Deep smal-based 3d reconstruction for animal re-identification. In: ACM MM. pp. 4680–4688 (2021)
56. Zhu, H., Budhwant, P., Zheng, Z., Nevatia, R.: Seas: Shape-aligned supervision for person re-identification. In: CVPR. pp. 164–174 (2024)
57. Zou, Y., Yang, X., Yu, Z., Kumar, B.V., Kautz, J.: Joint disentangling and adaptation for cross-domain person re-identification. In: ECCV. pp. 87–104. Springer (2020)