This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Object-Generalized Re-Identification: A Step Towards Universal Instance Perception

Shuoyi Chen , Yurui Wu, Mang Ye*
School of Computer Science, Wuhan University, Wuhan, China
{chenshuoyi, wuyurui, yemang}@whu.edu.cn
https://github.com/whucsy/OG-ReID

Abstract

The object re-identification (ReID) task aims to recognize the same individual object across diverse viewpoints and sensing conditions. Although person and vehicle ReID have achieved remarkable success, most existing methods are built on the assumption that training and testing data come from the same object category. This constraint requires separate models for each category, which limits scalability and generalization. To address this limitation, we introduce Object-Generalized Re-Identification (OG-ReID), a new paradigm that learns unified identity representations transferable across different object categories. Unlike conventional domain generalization that focuses on appearance variations within a single category, OG-ReID deals with category shifts caused by intrinsic structural differences in identity cues. To achieve this goal, we introduce the Meta-Generalized Object Re-Identification (MGOR) framework, which treats meta-learning as semantic distributional regularization, exposing the model to controlled category shifts so that invariance emerges as an equilibrium between semantic diversity and identity discrimination. Extensive evaluations on more than 100 unseen object categories from multiple domains show that MGOR outperforms existing ReID approaches without any target-domain adaptation, advancing toward universal identity perception beyond domain and category boundaries.

1. Introduction

The object re-identification (ReID) task aims to recognize the same individual object across diverse viewpoints and sensing devices [50, 51]. Over the past decade, significant progress has been made in person and vehicle ReID, where deep learning [15, 52] frameworks and metric learning methods have achieved high instance-level accuracy within specific semantic domains [13, 54]. However, most

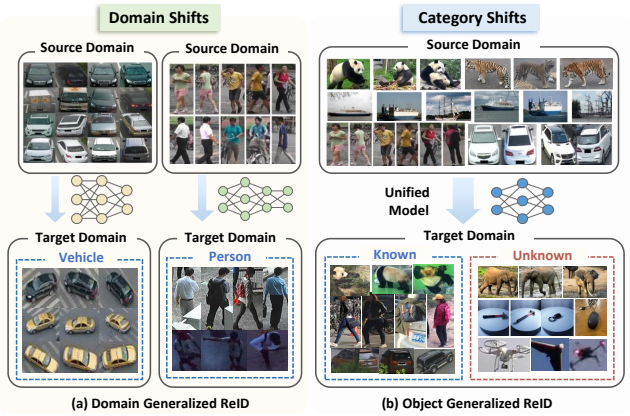


Figure 1. Comparison between DG-ReID and OG-ReID. (a) DG-ReID assumes a single category and addresses domain-level variations. (b) OG-ReID learns category-agnostic representations from mixed-category data, enabling generalization across object types.

ReID methods are built upon a category-consistent assumption that training and testing data originate from the same object category [7, 22, 31]. This paradigm requires training and maintaining separate models for each object type, which is annotation-intensive, computationally costly, and difficult to scale. In real-world applications such as smart cities, intelligent transportation, and ecological monitoring, visual perception must simultaneously handle a wide spectrum of categories, including persons, vehicles, vessels, buildings, animals, and various objects. To bridge this gap, future ReID systems should learn to generalize across heterogeneous objects with a unified model, an essential step toward universal visual intelligence.

In this paper, we introduce a new paradigm termed *Object-Generalized Re-Identification (OG-ReID)*, which studies instance-level matching in the absence of category isolation or category-specific priors. Unlike conventional ReID, which operates within a single semantic category, OG-ReID learns category-agnostic identity representations from mixed-category data and generalizes across heterogeneous object types. The core challenge underlying this

*Corresponding Author.

problem is the presence of category shifts, which fundamentally differ from domain shifts [41] such as variations in viewpoint or illumination. Category shifts arise from intrinsic differences in visual structures and identity-defining cues across objects: persons are recognized by body parts [25], vehicles by geometric contours [38], and animals by texture patterns [32]. Such heterogeneity disrupts the shared feature assumptions underlying conventional ReID models, making it difficult to preserve identity consistency across categories.

Domain generalization has long been explored in person ReID to enhance robustness against unseen domains when target-domain data are unavailable [29, 34, 39]. These approaches share an objective of learning domain-invariant representations [56]. To this end, a variety of strategies have been explored, such as style normalization [26, 35] and data augmentation [21] to mitigate domain-specific appearance bias; feature alignment and regularization [5] to promote inter-domain consistency; and disentanglement-based modeling [3] to separate domain-dependent and identity-related factors. Despite their effectiveness against domain variations, these approaches generalize poorly across category shifts, where the defining cues of identity differ fundamentally among object types, undermining the consistency of learned representations.

Several studies have explored meta-learning frameworks [9, 33, 59] that simulate episodic domain shifts to improve generalization. However, these methods inherently assume that repeated adaptation to synthetic domain splits suffices for generalization, thus confining meta-learning to appearance-level variations. We argue that, when reinterpreted beyond task adaptation, meta-learning can serve as a principled framework for semantic-invariant representation learning, enabling models to preserve identity consistency under evolving semantic distributions. In this paper, we propose Meta-Generalized Object Re-Identification (MGOR), a meta-learning framework that generalizes identity representations across object categories. To achieve semantic-level generalization, MGOR constructs dynamic meta-episodes that expose the model to continuously shifting category distributions, enabling it to internalize semantic variability as part of the learning process rather than noise to be suppressed.

Specifically, we design an Episodic Distribution Perturbation Sampling (EDPS) strategy that builds multi-category balanced mini-batches and introduces semantic perturbation by randomly dividing them into meta-train and meta-test subsets, effectively turning each episode into a controlled distributional shift. Building on this, we propose Meta-Object Invariance Learning (MOIL), which treats each iteration as a distribution consistency test. The model learns discriminative identity cues in the meta train phase and is validated on perturbed meta test subsets with altered category composition, promoting semantically invariant identity learning. To preserve a stable decision structure under EDPS perturbations, we further introduce conditional pairwise alignment regularization, which enforces conditional stability for comparable semantic units. On the other hand, a fundamental tension in object-level generalization lies in reconciling semantic discrimination within categories and invariance across categories. We address this tension through a Hierarchical Semantic Equilibrium (HSE) mechanism that implicitly aligns domain-sensitive and domain-invariant objectives via dual-level supervision. By leveraging category-aware cues within episodic training while consolidating invariant identity structure across episodes, MGOR internalizes semantic relations, yielding representations that are both semantically agnostic and identity discriminative across heterogeneous objects. In summary, our main contributions are as follows:

- We introduce a new paradigm, *Object-Generalized Re-Identification*, bridging the gap between category-specific ReID and universal visual understanding, developing MGOR as a unified framework for this setting.
- We propose Meta Object Invariance Learning (MOIL), which treats semantic perturbations as distributional shifts, promoting identity consistency under varying category distributions with conditional pairwise alignment regularization.
- We design Hierarchical Semantic Equilibrium, which implicitly aligns domain-sensitive and invariant objectives through dual-level supervision, promoting stable and semantically disentangled identity representations.
- Extensive experiments over 100 heterogeneous categories show that MGOR achieves state-of-the-art cross-category generalization and strong scalability to unseen objects.

2. Related Work

Object ReID aims to match the same object instance across different camera views despite appearance changes. Current approaches to improving ReID fall into two categories: designing advanced architectures (hybrid CNN and Transformer, hierarchical, or graph-based) [13, 38, 53] and enhancing feature representations by extracting stable, expressive object cues (multi-spatial, rotation-invariant, and part-level) [6, 24, 63]. Some methods use frequency-domain cues [20, 54], such as Haar wavelet or Fourier transforms, to enhance high-frequency texture details. Furthermore, multi-modal approaches have emerged, utilizing text information from Vision-Language Models to aid ReID by generating label text [23], constructing textual image representations [49], or employing object-level descriptive prompts [17]. These methods are largely developed for single-category ReID, they often exploit category-dependent appearance priors and dataset-specific biases, which weakens their ability to generalize across heterogeneous object types

and hampers scalable universal deployment.

Domain Generalized ReID focuses on training a model on one or more labeled source domains that can directly generalize to an unseen target domain without further adaptation. To enhance generalization, some methods incorporate meta-learning into training [33, 59], often simulating train-test splits and using high-order derivatives for model updates. Furthermore, Instance Normalization alleviates domain-specific information, but often compromises object discriminative details. Solutions combine Instance Normalization and Batch Normalization [9, 16, 35] to achieve domain generalization while preserving identity cues. Another prominent approach introduces the Mixture of Experts (MoE) mechanism to learn domain-discriminative features using multiple experts [10]. To mitigate parameter increase, specialized methods propose retaining expert-specific Batch Normalization while sharing other parameters [47] or designing parameter-efficient fine-tuning for MoE [45]. Recent studies leverage multi-modal models such as CLIP [37] to enhance domain-invariant feature learning, including domain-invariant prompts [58] and DFS sampling for improved fine-grained representations [57]. Collectively, these methods address domain shifts within fixed categories, leaving DG-ReID vulnerable to category shifts. Specifically, DG-ReID assumes a single category, while OG-ReID learns from mixed-category data and generalizes across heterogeneous categories.

Generalizable Object ReID defines the advanced goal of creating a single model capable of cross-category instance recognition, moving beyond the fixed-category focus of traditional ReID. This task is practically important because maintaining separate category-specific models is often unrealistic in open-world applications with diverse or unknown object types, and theoretically valuable because it studies category-agnostic identity representation under semantically heterogeneous conditions. The development of this field is underpinned by benchmarks like CUTE [19], which features 50 object categories and 180 identities with variations in lighting, pose, and background. However, the scene-constrained nature of this dataset may limit its ability to fully simulate complex real-world environments. Recent work [62] leverages Large Language Models [1, 2, 42], using in-context visual examples to construct category-specific prompts that guide object feature extraction and demonstrate better generalization to unseen domains. However, this formulation still assumes category-isolated retrieval, where category-specific visual cues are available to guide representation learning. As a result, it relies on category-conditioned modeling and does not directly apply to scenarios involving mixed or unseen categories without category priors. In contrast, our proposed OG-ReID focuses on category-agnostic instance matching under semantically heterogeneous settings, posing a fundamentally more challenging problem.

3. Method

3.1. Overview

The objective of Object-Generalized ReID is to learn instance-discriminative representations that generalize to unseen object categories with diverse semantics and appearances. However, training on heterogeneous object types inevitably induces semantic shifts, as each category exhibits distinct structural patterns and identity-defining cues. Such intrinsic heterogeneity disrupts feature consistency and limits the transferability of existing domain generalization ReID methods, which primarily address appearance-level variations while assuming shared semantic structures across domains. To overcome these limitations, we propose the Meta-Generalized Object Re-Identification (MGOR) framework, which reinterprets meta-learning as a process of semantic distributional regularization. Instead of suppressing inter-category variation, it internalizes semantic perturbations into the learning process, allowing the model to acquire invariance through exposure to dynamically shifting category distributions.

Concretely, MGOR constructs episodic training scenarios through *(1) Episodic Distribution Perturbation Sampling (EDPS)*, where multi-category balanced mini-batches are perturbed across meta-train and meta-test splits. This controlled semantic disturbance exposes the model to evolving category distributions and regularizes it to maintain identity stability under distributional shifts. Building on this foundation, *(2) Meta-Object Invariance Learning (MOIL)* models each episode as a semantic consistency process, driving the network to internalize identity invariance across object categories rather than relying on explicit alignment or adversarial objectives. Furthermore, *(3) Hierarchical Semantic Equilibrium (HSE)* introduces dual-level supervision that implicitly aligns domain-sensitive and invariant objectives, forming a semantic equilibrium that stabilizes the learned representation space. The overall framework is shown in Fig. 2.

3.2. Episodic Distribution Perturbation Sampling

Object-Generalized ReID requires learning identity representations that remain instance-discriminative while being insensitive to shifts in the category mixture (e.g., persons, vehicles, animals). Conventional $P \times K$ sampling guaranties identity diversity within a mini-batch but leaves the semantic composition largely unconstrained, so the training process rarely encounters the distributional changes to which it must generalize. In this part, we introduce Episodic Distribution Perturbation Sampling (EDPS). EDPS remedies this by constructing categories-balanced, identity-paired mini-batches and imposing an episodic

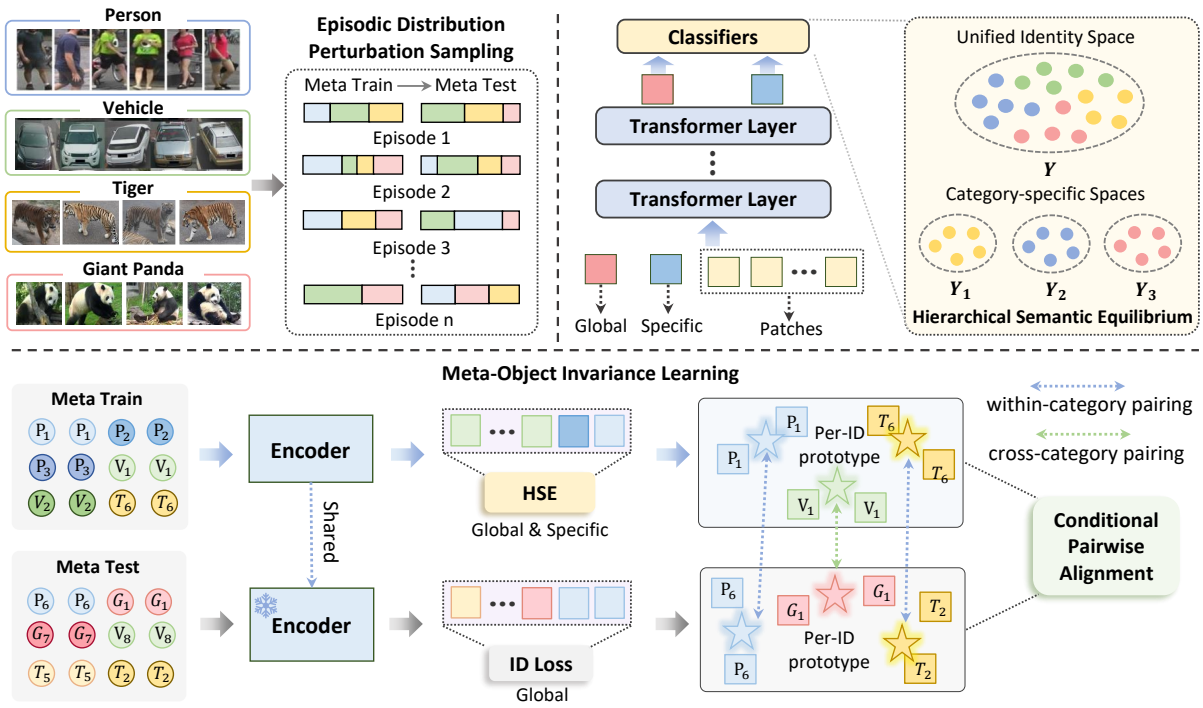


Figure 2. The MGOR framework with three components: EDPS for distribution-perturbed episodes, MOIL for cross-category identity invariance, and HSE for dual-level supervision that equilibrates domain-sensitive and invariant goals.

meta-train/meta-test partition. Each mini-batch thereby becomes a controlled perturbation of the category distribution, repeatedly testing and regularizing identity stability under evolving semantics.

Let $\mathcal{S}$ be the set of object categories; for each $s \in \mathcal{S}$, let $\mathcal{Y}_s$ denote the set of identity labels within s. We jointly train on the union of multiple datasets

$$\mathcal{D} = \bigcup_{s\in\mathcal{S}} \bigcup_{y\in\mathcal{Y}_s} \{(x, y, s)\}, \tag{1}$$

where x is an image, y its identity, and s its category. An encoder f_θ maps x to an identity representation $z = f_\theta(x)$. Given a target mini-batch size B and per-identity multiplicity $K \geq 2$, EDPS samples a subset $\mathcal{S}_b \subseteq \mathcal{S}$, and for each $s \in \mathcal{S}_b$ selects at least two identities,

$$\forall s \in \mathcal{S}_b: \quad \mathcal{Y}_b(s) = \{y_1^{(s)}, \ldots, y_{P_s}^{(s)}\} \subseteq \mathcal{Y}_s, \qquad P_s \geq 2, \tag{2}$$

then draws K images for each selected identity,

$$\mathcal{X}(y) = \{x_1^{(y)}, \ldots, x_K^{(y)}\}. \tag{3}$$

Within each categories, identities are paired (y_i, y_j) and packed to form the mini-batch

$$\mathcal{B} = \bigcup_{s\in\mathcal{S}_b} \bigcup_{(y_i,y_j)\in Pairs(\mathcal{Y}_b(s))} (\mathcal{X}(y_i) \cup \mathcal{X}(y_j)), \quad |\mathcal{X}(y_i) \cup \mathcal{X}(y_j)| = 2K, \tag{4}$$

until $|\mathcal{B}| \leq B$. This preserves a $P\times K$-style structure while explicitly balancing semantics at the batch level.

EDPS produces an episode by uniformly partitioning $\mathcal{B}$ into two equal, disjoint halves ,

$$\mathcal{B}^{\text{train}},\ \mathcal{B}^{\text{test}} \subset \mathcal{B}, \mathcal{B}^{\text{train}} \cap \mathcal{B}^{\text{test}} = \varnothing, \quad |\mathcal{B}^{\text{train}}| = |\mathcal{B}^{\text{test}}| = \tfrac{1}{2}|\mathcal{B}|. \tag{5}$$

3.3. Meta-Object Invariance Learning

Building on EDPS, which provides for each episode t a meta-train set $\mathcal{B}_t^{\text{train}}$ and a meta-test set $\mathcal{B}_t^{\text{test}}$ with perturbed category mixtures, MOIL reinterprets each training iteration as a distributional consistency test. The encoder f_θ must *(i) learn fine-grained identity discrimination on the meta-train set* and *(ii) retain this discrimination under a perturbed semantic composition on the meta-test set.* Departing from adaptation-oriented meta-learning [12], MOIL formulates a distributional-invariance objective: a single θ is optimized to preserve identity discriminability across shifts in the semantic mixture.

Let $(x_i, y_i, s_i) \in \mathcal{D}$ (Eq. 1) and $z_i = f_\theta(x_i) \in \mathbb{R}^d$. To ensure discrimination on $\mathcal{B}_t^{\text{train}}$, we optimize identity classification $\mathcal{L}_{\text{id}}^{\text{train}}$ and a commonly used metric criterion $\mathcal{L}_{\text{met}}^{\text{train}}$ (e.g., triplet/batch-hard). Keeping θ fixed, we evaluate the global identity discrimination capability of $\mathcal{B}_t^{\text{test}}$ by computing $\mathcal{L}_{\text{id}}^{\text{test}}$ to test its robustness to the introduced semantic perturbations. This denies readaptation to the new mixture, making semantic-level invariance an explicit stability

requirement for the same parameters θ.

Conditional Pairwise Alignment. Although the identity losses applied in the meta-train and meta-test phases encourage discrimination, each operates within its own distribution. To establish cross-perturbation discriminative stability, we propose Conditional Pairwise Alignment: under EDPS-induced mixture shifts, it first constructs cross-split semantic pairings between comparable units across $\mathcal{B}_t^{\text{train}}$ and $\mathcal{B}_t^{\text{test}}$ identity prototypes by default, or nearest-neighbor mini-clusters, and then applies lightweight stability only within those matched units. This preserves the global semantic diversity while reinforcing the local discriminative geometry that supports identity recognition.

Residual view and paired geometry. Let $z = f_\theta(x)$ and decompose each representation into an identity prototype and a residual,

$$z_i = \mu_{y_i} + r_i, \qquad \mu_y = \tfrac{1}{|\mathcal{X}(y)|} \sum_{x_j \in \mathcal{X}(y)} f_\theta(x_j), \tag{6}$$

where r_i captures fine-grained, identity-preserving variation. In each split, we form positive/negative distance sets

$$\begin{aligned} D^+_{\text{train}} &= \{\|z_i - z_j\| : y_i = y_j\}, \\ D^-_{\text{train}} &= \{\|z_i - z_j\| : y_i \neq y_j\}, \end{aligned} \tag{7}$$

and analogously $D^+_{\text{test}}, D^-_{\text{test}}$ on $\mathcal{B}_t^{\text{test}}$. This promotes conditional stability by aligning the shapes of these distributions via the 1D sliced Wasserstein distance:

$$\mathcal{L}_{\text{pair}} = \mathrm{S}\big(D^+_{\text{train}}, D^+_{\text{test}}\big) + \mathrm{S}\big(D^-_{\text{train}}, D^-_{\text{test}}\big). \tag{8}$$

3.4. Hierarchical Semantic Equilibrium

A central design in HSE is to factorize the label space into two complementary levels that mirror the objectives of OG-ReID. Let $\mathcal{S}$ denote the set of categories. For each $s \in \mathcal{S}$, let $\mathcal{Y}_s$ be the set of instance identities within s, and define the global identity space as the disjoint union

$$\mathcal{Y} = \bigsqcup_{s \in \mathcal{S}} \mathcal{Y}_s, \qquad \pi : \mathcal{Y} \to \mathcal{S}, \;\; \pi(y) = s \text{ if } y \in \mathcal{Y}_s. \tag{9}$$

Every identity $y \in \mathcal{Y}$ has a unique category tag $\pi(y) = s$, and a category-local index $\tilde{y} \in \{1, \ldots, |\mathcal{Y}_s|\}$ inside its own category. This hierarchical labeling induces two supervision views over the same representation $z = f_\theta(x)$: *(i) a global view that treats all identities in $\mathcal{Y}$ as mutually exclusive classes, abstracting away the category; (ii) a category-local view that focuses only on $\mathcal{Y}_s$ when $\pi(y) = s$, sharpening fine-grained separation within s.* HSE leverages these two views to reconcile cross-category invariance with intra-category discrimination through dual-level supervision applied asymmetrically across the EDPS episode.

To operationalize the factorization, we adopt a Vision Transformer with two class tokens: a global token $\mathbf{c}^{\mathrm{G}}$ for category-agnostic identity evidence and a local token $\mathbf{c}^{\mathrm{L}}$ for categories-specific detail. Both tokens attend to the same patches through the shared backbone; after the final layer they yield readouts $z^{\mathrm{G}}, z^{\mathrm{L}} \in \mathbb{R}^d$ that feed a global classifier $g : \mathbb{R}^d \to \Delta^{|\mathcal{Y}|}$ and categories-conditional classifiers $\{g^{(s)} : \mathbb{R}^d \to \Delta^{|\mathcal{Y}_s|}\}_{s \in \mathcal{S}}$. Supervision-aligned readouts encourage the backbone to allocate complementary subspaces to category-invariant and category-sensitive factors with minimal architectural overhead. In inference, only the global path z^{G} is used.

For (x, y, s) with $y \in \mathcal{Y}_s$ and categories-local index $\tilde{y}$, we adopt the standard global identity loss and its categories-local counterpart:

$$\begin{aligned} \ell_{\text{glob}}(x, y) &= -\log g(z^{\mathrm{G}})_y, \\ \ell_{\text{spec}}(x, \tilde{y}, s) &= -\log g^{(s)}(z^{\mathrm{L}})_{\tilde{y}}. \end{aligned} \tag{10}$$

The global term ℓ_{glob} is the same identity objective used in MOIL on both episode splits:

$$\begin{aligned} \mathcal{L}_{\text{id}}^{\text{train}} &= \mathbb{E}_{(x,y,\cdot) \sim \mathcal{B}_t^{\text{train}}}\big[\ell_{\text{glob}}(x, y)\big], \\ \mathcal{L}_{\text{id}}^{\text{test}} &= \mathbb{E}_{(x,y,\cdot) \sim \mathcal{B}_t^{\text{test}}}\big[\ell_{\text{glob}}(x, y)\big]. \end{aligned} \tag{11}$$

Meanwhile, the categories-local term is only active in meta-train:

$$\mathcal{L}_{\text{spec}}^{\text{train}} = \mathbb{E}_{(x,y,s) \sim \mathcal{B}_t^{\text{train}}}\big[\ell_{\text{spec}}(x, \tilde{y}, s)\big], \tag{12}$$

The global term pulls z^{G} toward a category-agnostic identity topology, while the categories term sharpens local margins in z^{L}. With a shared backbone, their complementary gradients separate invariant from category-sensitive cues without heavy architectural branching. The episode-level loss combines meta-train discrimination, meta-test stability, and a lightweight alignment penalty:

$$\mathcal{L}_{\mathrm{M}} = \mathcal{L}_{\text{id}}^{\text{train}} + \mathcal{L}_{\text{met}}^{\text{train}} + \mathcal{L}_{\text{spec}}^{\text{train}} + \mathcal{L}_{\text{id}}^{\text{test}} + \lambda \mathcal{L}_{\text{pair}}. \tag{13}$$

HSE complements MOIL's distributional consistency by injecting structure from the label space: meta-train exploits both global and categories-local signals to sculpt the representation, while meta-test validates only the global pathway under semantic perturbations. The episode-level consistency (MOIL) and the label-level factorization (HSE) yield embeddings that are category-agnostic yet instance-discriminative, with clean test-time fidelity and reduced negative transfer across heterogeneous object types.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. Our experiments assess generalization using datasets that span multiple object domains and highly diverse scenarios (Fig. 3), with dataset statistics summarized

Table 1. Training and test set statistics.

Split	Subset	# IDs	# Images	# Categories
Train	iPanda-50 [44]	35	5,209	1
	Market-1501 [60]	751	12,936	1
	VeRi [30]	576	37,715	1
	ATRW [22]	107	1,887	1
	VesselReID [55]	623	14,570	1
	Total (Train)	2,092	72,317	5
Test	MSMT17 [46]	3,060	93,820	1
	UAV-VeID [43]	2,208	19,058	1
	LeopardID2022 [14]	130	2,673	1
	ELPephants [18]	83	591	1
	University-1652 [61]	701	52,056	1
	CUTE [19]	180	18,000	50
	GZGC [36]	529	1,409	2
	PetFace [40]	55,686	115,708	13
	Wildlife71 [17]	1,924	108,096	67
	Total (Test)	64,501	411,411	137

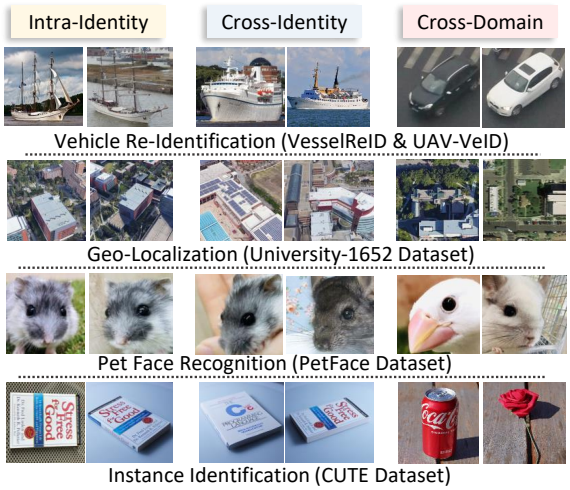


Figure 3. Evaluation on diverse domain generalization scenarios.

in Table 1. We constructed the training data by selecting representative benchmark from each of the most widely used object-level ReID datasets, resulting in five category ReID datasets: Market-1501 [60], VeRi [30], VesselReID [55], iPanda-50 [44], and ATRW [22]. Specifically, Market-1501[60], VeRi [30], and VesselReID[55] represent the person, vehicle, and maritime object domains, respectively, while iPanda-50 [44] and ATRW [22] represent two distinct animal domains. For testing, we validate our approach on nine diverse datasets to ensure robustness against both unseen categories and domain shifts. We select highly varied test data to evaluate generalization across complex scenarios. For known categories, we include MSMT17 [46], a large-scale, scene-complex person ReID benchmark, and UAV-VeID [43] for vehicle ReID, which presents the extreme domain challenge of drone perspectives. To assess generalization to unseen categories, our evaluation covers over 69 types of wild animals, including the specific animal datasets LeopardID2022 [14] and ELPephants [18], alongside Wildlife71 [17], which features numerous species in varying wild environments. We also evaluate our model on multi-category object datasets. CUTE [19] introduces 50 diverse object categories in a lab-controlled setting. Furthermore, we test our model's adaptability beyond the ReID task. We evaluate its performance on a Geo-localization task using the University-1652 [61] dataset, which focuses on building recognition from satellite view to drone view. We also include PetFace [40] for the face recognition task on fine-grained pet instances.

Evaluation Metrics. We evaluate the performance across nine distinct datasets, employing the standard Re-ID metrics, mAP (mean Average Precision) and CMC (Cumulative Matching Characteristics). For datasets without camera annotations, we further use mINP (mean Inverse Negative Penalty) [50], which evaluates accuracy on the hardest negative samples. The dataset-specific evaluation details are provided in the Appendix.

4.2. Implementation Details.

The backbone network used in our experiments is ViT-B/16 [11] pretrained on ImageNet-1K. Input images are resized to 256×256 pixels, padded by 10 pixels, and randomly flipped horizontally with a probability of 0.5. Each training iteration constructs an episode consisting of a category-balanced meta-training batch with $P \times K$ samples ($K = 4$) and a meta-test batch with perturbed class proportions. The total batch size is 64. Optimization is performed using SGD with an initial learning rate of 10^{-3} and weight decay of 10^{-4}. Backpropagation is applied once per episode, and training proceeds for 60 epochs.

4.3. Comparison with SOTA DG-ReID Methods

In Table 2, we group the compared approaches into two categories: conventional domain-specific ReID methods and domain-generalizable ReID methods. Evaluation is conducted under two regimes: Single-Category and Multi-Category. In the Single-Category regime, we evaluate the model on both seen and unseen categories. MSMT17 [46] and UAV-VeID [43] are used for seen-category evaluation, with the same target categories as in training but markedly different data settings. Several animal datasets are used for unseen-category evaluation. In the Multi-Category regime, each dataset exhibits substantially higher diversity and complexity. The suite includes multi pet face recognition that captures variation across different pet types [40], instance-level identification of regular objects [19], and large-scale wildlife ReID that spans terrestrial, aerial, and marine environments [17]. All comparison results were reproduced under the same settings as our method, except for VICP [62], which additionally requires one batch of target-domain data

Table 2. Comparison with state-of-the-art methods on nine datasets. The upper table shows single-category ReID results, and the lower table shows multi-category ReID results. † denotes methods adapted with a few target-domain samples.

Method	Time	MSMT17 [46]			UAV-VeID [43]			LeopardID2022 [14]			ELPephants[18]			University-1652 [61]		
		mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1
State-of-the-art methods for domain specific ReID																
TransReID[13]	ICCV 2021	12.2	0.7	29.6	35.0	35.0	27.4	24.6	9.3	**63.0**	11.1	2.7	33.0	11.5	2.1	28.2
PASS-ReID[63]	ECCV 2022	3.5	0.2	10.3	26.0	26.0	19.1	23.2	9.1	62.7	9.5	2.2	27.2	6.3	0.6	23.4
PHA[54]	CVPR 2023	8.2	0.5	21.0	33.9	33.9	26.4	23.9	9.2	61.8	11.4	2.6	33.8	8.8	1.2	23.4
CLIP-ReID[23]	AAAI 2023	**19.4**	**1.4**	**43.4**	35.3	35.3	27.6	24.7	9.3	62.4	8.7	2.0	24.4	15.4	1.9	31.7
State-of-the-art methods for domain generalizable ReID																
TransMatcher [27]	NeurIPS 2021	15.1	0.4	42.4	27.8	27.8	23.6	10.9	2.9	42.6	10.2	2.2	26.1	10.6	0.4	50.8
QAConv [28]	CVPR 2022	14.8	0.3	43.0	28.3	28.3	24.3	10.3	2.9	43.8	10.1	2.2	26.4	10.4	0.3	56.8
PAT [34]	ICCV 2023	14.1	0.7	35.3	34.9	34.9	27.8	24.9	9.0	62.2	11.4	2.6	33.8	14.0	1.7	32.4
BAU [8]	NeurIPS 2024	2.7	0.1	9.1	9.0	9.0	6.1	12.3	8.3	42.0	4.3	1.8	9.0	0.4	0.1	2.9
ReNorm [35]	ECCV 2024	6.8	0.2	24.1	25.3	25.3	21.5	16.7	8.5	55.7	6.8	2.2	16.2	3.0	0.1	24.4
DACS [48]	AAAI 2024	12.5	0.7	31.8	33.5	33.5	26.9	**25.4**	**9.7**	62.6	10.9	2.4	31.8	14.9	1.9	32.2
CLIP-FGDI [58]	TIFS 2025	15.8	0.7	38.9	23.9	23.9	18.0	17.5	8.3	57.1	6.4	1.9	15.4	4.6	0.3	20.8
ADSR [4]	ACMMM 2025	1.2	0.1	4.5	10.8	10.8	6.7	23.3	9.0	61.6	8.6	2.1	23.5	5.7	0.2	30.7
VICP †[62]	ICCV 2025	16.0	0.8	40.4	19.5	19.5	13.6	23.6	9.2	62.6	11.1	2.3	32.3	13.7	2.6	28.4
Ours	-	18.1	1.1	41.3	**38.6**	**38.6**	**32.0**	24.7	9.2	61.9	**14.0**	**2.8**	**38.6**	**19.5**	**3.5**	**38.7**

Method	Time	CUTE [19]			GZGC [36]			PetFace [40]			Wildlife71 [17]			Avg.		
		mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1
State-of-the-art methods for domain specific ReID																
TransReID [13]	ICCV 2021	19.6	0.6	83.9	14.1	**7.3**	25.0	29.1	27.6	26.6	83.3	49.1	96.1	26.7	14.9	45.9
PASS-ReID [63]	ECCV 2022	13.1	0.6	77.3	12.9	6.5	22.6	22.4	21.1	20.7	81.1	44.9	96.3	22.0	12.4	40.0
PHA [54]	CVPR 2023	19.2	0.6	84.0	13.9	7.2	25.0	26.1	24.6	23.9	80.9	44.1	96.1	25.1	13.8	43.9
CLIP-ReID [23]	AAAI 2023	23.7	0.9	84.8	13.6	7.0	22.7	22.9	21.6	20.5	75.5	32.3	95.9	26.6	12.4	45.9
State-of-the-art methods for domain generalizable ReID																
TransMatcher [27]	NeurIPS 2021	17.9	0.6	74.4	11.3	6.0	20.1	34.1	32.7	31.2	72.0	29.8	95.0	23.3	11.4	45.1
QAConv [28]	CVPR 2022	14.5	0.6	73.8	12.5	6.5	22.7	33.4	32.2	31.0	71.8	29.2	94.7	22.9	11.4	46.3
PAT [34]	ICCV 2023	24.1	0.7	83.7	13.5	6.8	24.3	35.8	34.2	32.8	83.3	46.8	95.9	28.4	15.3	47.6
BAU [8]	NeurIPS 2024	11.1	0.6	69.3	5.0	2.7	7.6	3.1	3.0	2.7	16.6	0.9	68.8	7.2	2.9	24.2
ReNorm [35]	ECCV 2024	17.1	0.6	79.0	8.2	3.9	15.6	17.6	16.7	16.3	51.4	9.7	92.4	17.0	7.5	38.4
DACS [48]	AAAI 2024	23.5	0.7	84.9	13.2	6.7	23.0	32.4	30.8	29.5	82.7	46.7	95.8	27.7	14.8	46.5
CLIP-FGDI [58]	TIFS 2025	24.9	0.9	82.6	10.5	5.4	19.1	11.5	10.8	10.0	42.5	6.4	90.4	17.5	6.5	39.1
ADSR [4]	ACMMM 2025	14.9	0.6	79.1	9.2	4.2	16.4	20.8	19.8	18.3	62.1	12.7	94.5	17.4	6.6	37.3
VICP †[62]	ICCV 2025	30.1	0.8	**89.7**	7.4	3.6	13.0	30.9	29.4	27.9	38.7	1.2	83.3	21.2	6.8	43.5
Ours	-	**32.9**	**1.3**	87.9	**14.2**	7.1	**25.2**	**41.4**	**39.7**	**38.2**	**86.0**	**54.5**	**95.8**	**32.2**	**17.5**	**51.1**

(128 images) for adaptation. Results show that under OG-ReID, many existing DG methods fail to generalize and even perform worse than several domain-specific baselines.

Single-Category Evaluation. For ReID tasks involving a single object category, several domain generalization methods such as PAT [34] and CLIP-FGDI [58] still maintain competitive performance, due to their algorithmic designs with broad applicability. For instance, PAT emphasizes local discriminative cues, while CLIP-FGDI benefits from the strong cross-category knowledge encoded in CLIP. However, other DG-oriented methods focusing primarily on style normalization, such as ReNorm [35], BAU [8], and ADSR [4], show significant performance degradation, often falling behind conventional supervised ReID methods. This observation suggests that most existing DG-ReID techniques, when trained jointly over multiple object categories, struggle to learn robust discriminative representations due to semantic entanglement across domains, leading to reduced generalization. In contrast, conventional supervised approaches such as TransReID [13], which adopt transformer-based backbone architectures, demonstrate greater structural stability. Building on this robust backbone, our method explicitly learns object-invariant features that generalize across diverse categories, achieving state-of-the-art performance. While CLIP-ReID achieves slightly higher accuracy than ours on MSMT17, our method demonstrates more consistent and overall superior performance across diverse domains, highlighting its stronger generalization capability beyond specific categories.

Multi-Category Evaluation. In multi-category ReID scenarios, our approach exhibits even more pronounced advantages. Although VICP [62], a general object ReID

Table 3. Experimental results of the ablation study for the proposed Meta-Generalized Object Re-Identification framework.

Method	CUTE [19]		MSMT17 [46]		PetFace [40]		University-1652 [61]		Wildlife71 [17]		ELPephants [18]	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
Baseline	23.3	85.6	14.2	34.1	31.5	28.8	13.5	32.0	83.3	95.0	11.8	34.7
+ EDPS & MOIL	31.6	88.0	16.2	39.4	40.5	37.4	18.2	37.9	85.3	95.7	13.6	36.5
+ EDPS & MOIL & HSE	32.9	87.9	18.1	41.3	41.4	39.7	19.5	38.7	86.0	95.8	14.0	38.6

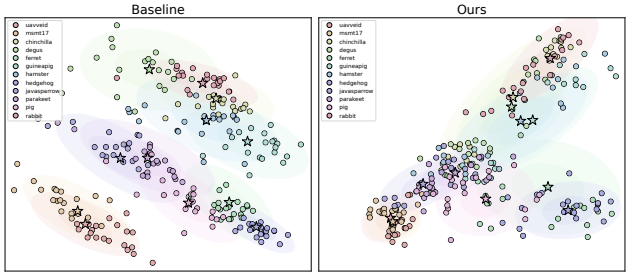


Figure 4. Domain distribution visualization of samples: Baseline (left) vs. Ours (right). Samples from different domains (colors) are more mixed within each cluster, indicating smaller domain gaps.

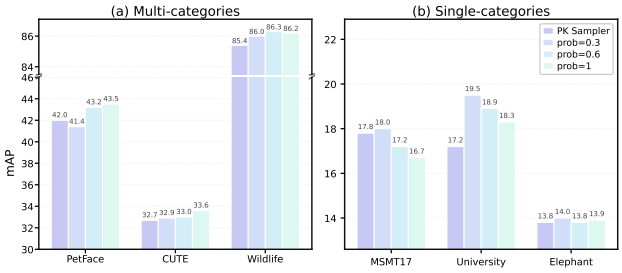


Figure 5. Parameter analysis of EDPS under different category-disjoint probabilities between the meta-train and meta-test sets.

method with goals similar to ours, performs well in single category scenarios, it suffers a pronounced drop in mixed category settings. This decline may result from prompt-based adaptation with limited target-domain data, which is less effective in multi-category settings. In contrast, our method needs no target-domain adaptation and still achieves comparable or better accuracy, showing strong practical potential for open-world ReID.

4.4. Ablation Study

Effectiveness of Each Component. To evaluate the contribution of each component in the proposed MGOR framework, we conduct an ablation study across six benchmark datasets. Our baseline is the plain ViT-based ReID model from TransReID[13], without any additional modules. As shown in Table 3, incorporating Episodic Distribution Perturbation Sampling (EDPS) and Meta-Object Invariance Learning (MOIL) consistently improves the baseline across all datasets. This demonstrates that episodic perturbation and meta-level invariance learning effectively regularize the feature space against distributional and semantic shifts. Further integrating Hierarchical Semantic Equilibrium (HSE) yields the most significant overall improvement, achieving state-of-the-art performance across both evaluation regimes. The dual-level supervision enhances fine-grained identity discrimination while stabilizing cross-category representation learning.

Visualization Analysis. We randomly select domains spanning multiple categories. Features extracted by baseline and ours are embedded into a shared 2D space for visual comparison. As shown in Fig. 4, each color corresponds to a category. Compared with the baseline, our method produces stronger cross-domain color mixing and markedly weaker domain boundaries. This indicates that the proposed framework learns representations that are more invariant across domains, enabling improved generalization under previously unseen semantic conditions.

Parameter Analysis. In Fig. 5, we study the effect of the category-disjoint probability between the meta-train and meta-test sets in EDPS. We first replace EDPS with the widely used PK sampling strategy in ReID. Unlike PK sampling, which introduces random perturbations between the meta-train and meta-test sets, EDPS provides more controllable category allocation. Increasing the disjointness between meta-train and meta-test degrades performance on the seen-category dataset MSMT17, but improves performance on multi-category evaluation.

5. Conclusion

We introduced the Object-Generalized Re-Identification paradigm and the MGOR framework, which promotes category-agnostic identity generalization through semantic distributional regularization. This setting is important because it moves ReID beyond category-bounded modeling toward a more unified and scalable form of identity perception across heterogeneous objects. Despite its promise, our framework still depends on existing annotated datasets, whose limited scale and category coverage restrict the diversity of semantic shifts available for training. Future work may further advance this direction through larger open benchmarks, reduced reliance on annotation via self-supervised learning or continual meta-learning, and stronger semantic transfer from vision-language foundation models. We hope this study can serve as a step toward more universal and practical ReID systems.

Acknowledgements. This work is partially supported by National Natural Science Foundation of China under Grant (T2541022, 62361166629), Major Project of Science and Technology Innovation of Hubei Province (2025BEA002).

References

[1] AI@Meta. Llama 3 model card. *arXiv preprint arXiv:2407.21783*, 2024. 3

[2] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. pages 2397–2430. PMLR, 2023. 3

[3] Patrick PK Chan, Xiaoman Hu, Haorui Song, Peng Peng, and Keke Chen. Learning disentangled features for person re-identification under clothes changing. In *ACM MM*, pages 1–21. ACM New York, NY, 2023. 2

[4] Long Chen, De Cheng, Shizhou Zhang, Yinghui Xing, Di Xu, and Yanning Zhang. Amplitude-aware domain style replay for lifelong person re-identification. In *ACM MM*, page 2026–2035, 2025. 7

[5] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Mingliang Xu, Qi Tian, and Rongrong Ji. Dual distribution alignment network for generalizable person re-identification. In *AAAI*, pages 1054–1062, 2021. 2

[6] Shuoyi Chen, Mang Ye, and Bo Du. Rotation invariant transformer for recognizing object in uavs. In *ACM MM*, pages 2565–2574, 2022. 2

[7] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE TPAMI*, 40(2):392–408, 2017. 1

[8] Yoonki Cho, Jaeyoon Kim, Woo Jae Kim, Junsik Jung, and Sung eui Yoon. Generalizable person re-identification via balancing alignment and uniformity. In *NeurIPS*, 2024. 7

[9] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *CVPR*, pages 3425–3435, 2021. 2, 3

[10] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, 2021. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. pages 1126–1135. PMLR, 2017. 4

[13] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021. 1, 2, 7, 8

[14] Saihui Hou, Panjian Huang, Zengbin Wang, Yuan Liu, Zeyu Li, Man Zhang, and Yongzhen Huang. Openanimals: Revisiting person re-identification for animals. In *ICCV*, 2025. 6, 7

[15] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Self-knowledge distillation with dimensional history knowledge. *Science China Information Sciences*, 68(9):192101, 2025. 1

[16] Bingliang Jiao, Lingqiao Liu, Liying Gao, Guosheng Lin, Lu Yang, Shizhou Zhang, Peng Wang, and Yanning Zhang. Dynamically transformed instance normalization network for generalizable person re-identification. In *ECCV*, 2022. 3

[17] Bingliang Jiao, Lingqiao Liu, Liying Gao, Ruiqi Wu, Guosheng Lin, Peng Wang, and Yanning Zhang. Toward re-identifying any animal. In *NeurIPS*, 2023. 2, 6, 7, 8

[18] Matthias Korschens and Joachim Denzler. Elpephants: A fine-grained dataset for elephant re-identification. pages 0–0, 2019. 6, 7, 8

[19] Klemen Kotar, Stephen Tian, Hong-Xing Yu, Daniel L.K. Yamins, and Jiajun Wu. Are these the same apple? comparing images based on object intrinsics. In *NeurIPS*, 2023. 3, 6, 7, 8

[20] Chenyue Li, Shuoyi Chen, and Mang Ye. Adaptive high-frequency transformer for diverse wildlife re-identification. In *ECCV*, pages 296–313, 2024. 2

[21] He Li, Mang Ye, Cong Wang, and Bo Du. Pyramidal transformer with conv-patchify for person re-identification. In *ACM MM*, pages 7317–7326, 2022. 2

[22] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: a benchmark for amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019. 1, 6

[23] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, pages 1405–1413, 2023. 2, 7

[24] Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. Dc-former: Diverse and compact transformer for person re-identification. In *AAAI*, pages 1415–1423, 2023. 2

[25] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021. 2

[26] Yuke Li, Jingkuan Song, Hao Ni, and Heng Tao Shen. Style-controllable generalized person re-identification. In *ACM MM*, pages 7912–7921, 2023. 2

[27] Shengcai Liao and Ling Shao. Transmatcher: Deep image matching through transformers for generalizable person re-identification. *NeurIPS*, 34:1992–2003, 2021. 7

[28] Shengcai Liao and Ling Shao. Graph sampling based deep metric learning for generalizable person re-identification. In *CVPR*, pages 7359–7368, 2022. 7

[29] Fangyi Liu, Mang Ye, and Bo Du. Domain generalized federated learning for person re-identification. *Computer Vision and Image Understanding*, 241:103969, 2024. 2

[30] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE TMM*, 20(3):645–658, 2018. 6

[31] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *CVPR*, pages 3235–3243, 2019. 1

[32] Ekaterina Nepovinnykh, Tuomas Eerola, Vincent Biard, Piia Mutka, Marja Niemi, Mervi Kunnasranta, and Heikki Kälviäinen. Sealid: Saimaa ringed seal re-identification dataset. *Sensors*, 22(19):7602, 2022. 2

[33] Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, and Heng Tao Shen. Meta distribution alignment for generalizable person re-identification. In *CVPR*, pages 2487–2496, 2022. 2, 3

[34] Hao Ni, Yuke Li, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Part-aware transformer for generalizable person re-identification. In *ICCV*, pages 11280–11289, 2023. 2, 7

[35] Ren Nie, Jin Ding, Xue Zhou, and Xi Li. Rethinking normalization layers for domain generalizable person re-identification. In *ECCV*, pages 267–284. Springer, 2024. 2, 3, 7

[36] Jason Parham, Jonathan Crall, Charles Stewart, Tanya Berger-Wolf, and Daniel Ian Rubenstein. Animal population censusing at scale with citizen science and photographic identification. In *AAAI*, 2017. 6, 7

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763. PMLR, 2021. 3

[38] Fei Shen, Yi Xie, Jianqing Zhu, Xiaobin Zhu, and Huanqiang Zeng. Git: Graph interactive transformer for vehicle re-identification. *IEEE TIP*, 32:1039–1051, 2023. 2

[39] Wuxuan Shi and Mang Ye. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In *ICCV*, pages 1772–1781, 2023. 2

[40] Risa Shinoda and Kaede Shiohara. Petface: A large-scale dataset and benchmark for animal identification. In *ECCV*, pages 19–36. Springer, 2025. 6, 7, 8

[41] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, pages 719–728, 2019. 2

[42] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[43] Shangzhi Teng, Shiliang Zhang, Qingming Huang, and Nicu Sebe. Viewpoint and scale consistency reinforcement for uav vehicle re-identification. *IJCV*, 129(3):719–735, 2021. 6, 7

[44] Le Wang, Rizhi Ding, Yuanhao Zhai, Qilin Zhang, Wei Tang, Nanning Zheng, and Gang Hua. Giant panda identification. *IEEE TIP*, 30:2837–2849, 2021. 6

[45] Yang Wang, Yixing Zhang, Xudie Ren, and Yuxin Deng. Moda: Mixture of domain adapters for parameter-efficient generalizable person re-identification. In *ACM Trans. Multimedia Comput.*, 2025. 3

[46] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 6, 7, 8

[47] Boqiang Xu, Jian Liang, Lingxiao He, and Zhenan Sun. Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *European Conference on Computer Vision*, pages 372–388. Springer, 2022. 3

[48] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yifan He, Shaozi Li, and Nicu Sebe. Diversity-authenticity co-constrained stylization for federated domain generalization in person re-identification. In *AAAI*, pages 6477–6485, 2024. 7

[49] Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *CVPR*, pages 17343–17353, 2024. 2

[50] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 44(6): 2872–2893, 2022. 1, 6

[51] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *arXiv preprint arXiv:2401.06960*, 2024. 1

[52] Mang Ye, Wenke Huang, Zekun Shi, He Li, and Bo Du. Revisiting federated learning with label skew: an over-confidence perspective. *Science China Information Sciences*, 68(9):192102, 2025. 1

[53] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, pages 516–525, 2021. 2

[54] Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *CVPR*, pages 14133–14142, 2023. 1, 2, 7

[55] Qian Zhang, Mingxin Zhang, Jinghe Liu, Xuanyu He, Ran Song, and Wei Zhang. Unsupervised maritime vessel re-identification with multi-level contrastive learning. *IEEE Transactions on Intelligent Transportation Systems*, 24(5): 5406–5418, 2023. 6

[56] Yi-Fan Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. Learning domain invariant representations for generalizable person re-identification. *IEEE TIP*, 32: 509–523, 2022. 2

[57] Huazhong Zhao, Lei Qi, and Xin Geng. Clip-dfgs: A hard sample mining method for clip in generalizable person re-identification. In *ACM Trans. Multimedia Comput.*, 2024. 3

[58] Huazhong Zhao, Lei Qi, and Xin Geng. Cilp-fgdi: Exploiting vision-language model for generalizable person re-identification. *IEEE TIFS*, 2025. 3, 7

[59] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, pages 6277–6286, 2021. 2, 3

[60] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 6

[61] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *ACM MM*, pages 1395–1403, 2020. 6, 7, 8

[62] Huang Zhizhong and Xiaoming Liu. Vicp: Generalizable object re-identification via visual in-context prompting. In *ICCV*, 2025. 3, 6, 7

[63] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *ECCV*, 2022. 2, 7