

Generalizable Object Re-Identification via Visual In-Context Prompting

Zhizhong Huang Xiaoming Liu
 Michigan State University
 East Lansing, MI, USA
 {huang296, liuxm}@msu.edu

Abstract

Current object re-identification (ReID) methods train domain-specific models (e.g., for persons or vehicles), which lack generalization and demand costly labeled data for new categories. While self-supervised learning reduces annotation needs by learning instance-wise invariance, it struggles to capture identity-sensitive features critical for ReID. This paper proposes Visual In-Context Prompting (VICP), a novel framework where models trained on seen categories can directly generalize to unseen novel categories using only in-context examples as prompts, without requiring parameter adaptation. VICP synergizes LLMs and vision foundation models (VFM): LLMs infer semantic identity rules from few-shot positive/negative pairs through task-specific prompting, which then guides a VFM (e.g., DINO) to extract ID-discriminative features via dynamic visual prompts. By aligning LLM-derived semantic concepts with the VFM’s pre-trained prior, VICP enables generalization to novel categories, eliminating the need for dataset-specific retraining. To support evaluation, we introduce ShopID10K, a dataset of 10K object instances from e-commerce platforms, featuring multi-view images and cross-domain testing. Experiments on ShopID10K and diverse ReID benchmarks demonstrate that VICP outperforms baselines by a clear margin on unseen categories. Code is available at <https://github.com/Hzzone/VICP>.

1. Introduction

Object Re-Identification (ReID) aims to identify and match specific instances of objects across non-overlapping camera views and scenarios, a capability critical for autonomous systems, surveillance [42], and e-commerce. While extensively studied for persons [30] and vehicles [49, 73], ReID applications span diverse object categories like pets [57], products [52], and wildlife [11]. Unlike object classification, ReID demands distinguishing fine-grained intra-class variations—e.g., scratch on vehicles or logos for products—while remaining invariant to viewpoint, lighting, and occlusion.

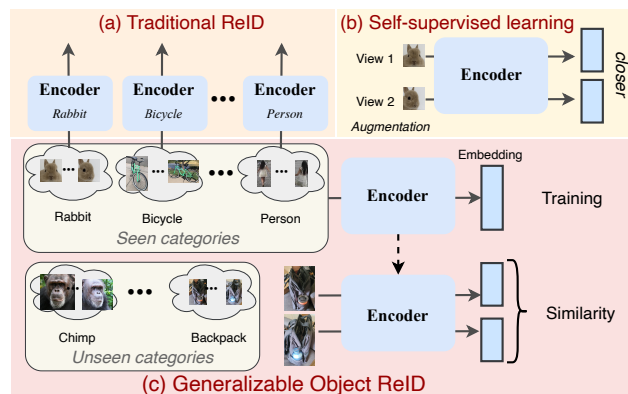


Figure 1. Overview of Object Re-Identification: (a) Traditional ReID methods trained the model on specific categories (e.g., rabbit, bicycle, person); (b) Self-supervised learning encourages pull together different augmentations of the same image; (c) Generalizable Object ReID extends ReID to novel categories (e.g., chimp, backpack), enabling ReID on unseen categories.

Traditional approaches address challenges like visible-infrared ReID [46, 70], text-to-image retrieval [6, 12], un/semi-supervised ReID [47, 76], domain adaptation [64], and clothes-changing scenarios [25, 31]. Yet, these methods remain category-specific: models trained on persons fail on vehicles or products, requiring costly labeled data for each new category, as shown in Fig. 1(a). This specialization hinders deployment in dynamic real-world settings where novel objects (e.g., rare animal species, or emerging retail items) require rapid adaptation.

Self-supervised learning (SSL) methods like DINO [10, 53] and MoCo [16, 18, 28] learn representations by maximizing similarity between augmented views of an image, like contrastive loss [67] (see Fig. 1(b)). While SSL reduces annotation needs and improves generalization to unseen domains, its objective—preserving semantic consistency—aligns poorly with ReID’s core requirement: capturing fine-grained, identity-sensitive features. For instance, person ReID [15, 29, 30, 76] requires distinguishing subtle differences in body shape or accessory, while pet ReID [57] relies on unique fur patterns or facial markings. Conse-

quently, SSL-trained models, usually beneficial for classification/detection/segmentation, often overlook these discriminative local cues, leading to suboptimal ReID performance.

A critical question emerges, as shown in Fig. 1(c): *How can we build a ReID model that generalizes to arbitrary object categories without dataset-specific training?* Vision foundation models (VFM), *e.g.* DINOv2 [53] and CLIP [54], offer strong visual priors, but their general-purpose features lack task-specific adaptation for ReID. In contrast, large language models (LLMs) [22, 78] excel at in-context learning [20]—extracting task rules from minimal examples. We propose that unifying these paradigms can unlock generalization: LLMs can infer identity-discriminative rules from few-shot examples, while VFMs can localize and encode fine-grained visual traits—leading to visual in-context prompting (VICP), a unified ReID framework.

Specifically, in-context learning [20] enables models to solve tasks by conditioning on example input-output prompts without parameter updates. For ReID, this means providing a model with contextual pairs (*e.g.*, positive pairs of the same instance and negative pairs of similar but distinct objects) to infer identity-sensitive attributes. For instance, given images of handbags, an LLM could deduce that “matching stitching patterns and logo placements” define identity, while “color variations under different lighting” are irrelevant. This semantic reasoning can dynamically guide the VFM to focus on task-critical features.

On the other hand, to emphasize ReID-specific traits of pre-trained VFMs, our framework translates LLM-derived semantic rules into dynamic visual prompts—task-specific instructions generated from in-context visual example pairs. For instance, given a few-shot positive pair (*e.g.*, two images of the same handbag under different viewpoints) and a negative pair (*e.g.*, two visually similar but distinct handbags), the LLM analyzes their relationships to infer identity-critical attributes (*e.g.*, “focus on logo placement and stitching patterns while disregarding lighting variations”). These inferred rules are then mapped into visual prompts [34] that tune the VFM’s feature extraction process. Unlike text-based prompting, our visual prompts are derived directly from the input example pairs, enabling the model to prioritize fine-grained local features (*e.g.*, textures, shapes) over globally invariant semantics. This adaptation preserves the VFM’s generalization while aligning it with ReID, achieving strong generalization without any parameter updates.

Despite progress in domain-specific ReID, generalizable object ReID remains underexplored. In this paper, we systematically establish baselines for generalizable object ReID, including self-supervised models, vision foundation models, and their adaptations. Unlike limited categories [65, 75] or well-controlled conditions [39], we further introduce ShopID10K, a dataset with instance labels curated from e-commerce platforms, comprising 10K instances across

34 daily-life categories (bag, shoes, bicycle, *etc.*), featuring multi-view images, occlusions, and high inter-class similarity (*e.g.*, near-identical products differing only in logos). This benchmark enables the rigorous evaluation of cross-category generalization under real-world conditions.

Our contributions are summarized as follows:

- We define a novel task, Generalizable Object ReID, that requires ReID model to adapt to unseen categories using only a few examples.
- We propose a unified ReID framework, VICP, where LLMs infer identity rules from few-shot pairs, and dynamic prompts adapt VFMs for fine-grained ReID.
- We release ShopID10K, a benchmark for evaluating cross-category ReID generalization, fostering research in scalable ReID systems.
- Our method outperforms self-supervised and few-shot baselines by 4% mAP on ShopID10K and standard datasets (MSMT17, VeRi-776), achieving state-of-the-art performance with minimal prompts.

2. Related Work

Self-supervised Learning. Self-supervised learning is a powerful paradigm for learning meaningful representations without manual annotations. Contrastive learning methods, such as MoCo [28] and SimCLR [14], construct positive pairs through data augmentation and maximize agreement between them. Subsequent works improve discriminativity of representations via clustering [9, 32], hard negative [56] or data augmentation [63]. Masked image modeling (MIM) methods [8, 27, 69] like MAE [27] focus on reconstructing masked regions, yet their representations are less discriminative than contrastive methods for fine-grained tasks. Foundation models like DINOv2 [53], pre-trained on massive datasets, demonstrate exceptional generalization in downstream tasks [2, 26, 35, 66, 71]. However, despite their semantic awareness, these representations often fail at instance-level retrieval due to the lack of explicit ID supervision. Our method addresses this gap by leveraging LLM-guided in-context learning to inject ID-sensitive priors into VFMs.

Object Re-Identification. Traditional object ReID methods [38, 43, 44, 49, 51, 59, 73] are category-specific, requiring dedicated training for pedestrians or vehicles. While variants like unsupervised ReID [76] and cross-modal ReID [12] address label scarcity or modality gaps, they remain confined to predefined categories. Deploying these methods to novel categories (*e.g.*, pets [57] or products [52]) necessitates laborious data collection and retraining. This limitation highlights the need for a unified framework capable of generalizing across unseen object categories. Indeed, our method allows a single model to adaptively identify object instances from unseen categories using only a few exemplars.

In-context Learning. In-context learning (ICL) [20], pop-

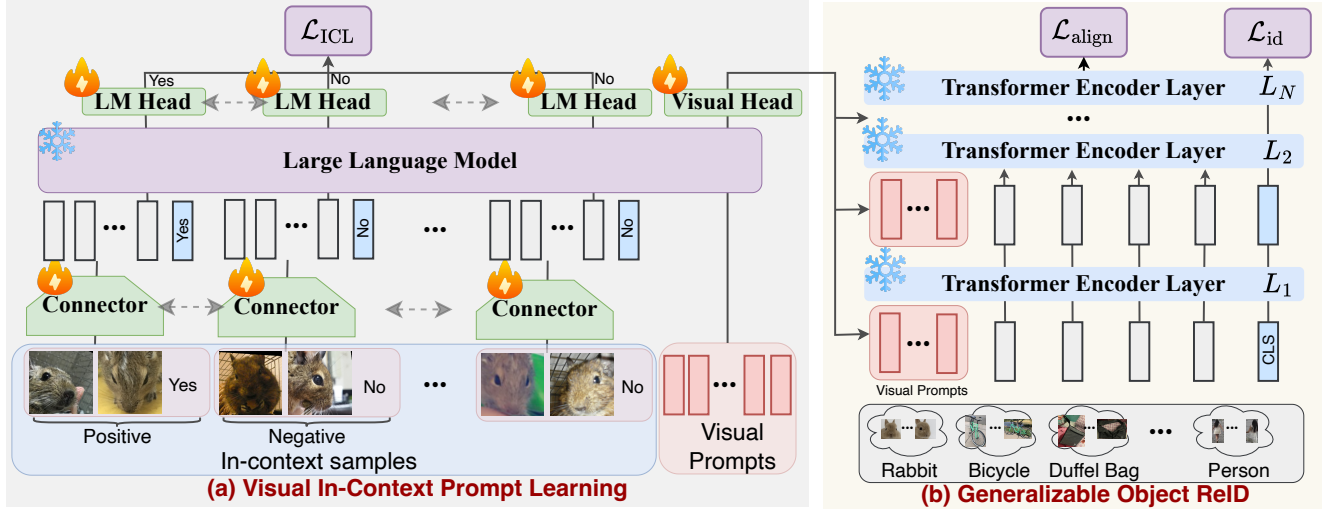


Figure 2. Overview of our proposed framework. a) In-Context Visual Prompt Generation: Given few-shot input pairs, a frozen LLM and a trainable connector/LM head process the context to infer identity-discriminative rules. The visual head generates learnable visual prompts conditioned on the input pairs. b) Generalizable Object ReID: The generated visual prompts are injected into each layer of a vision transformer (e.g., DINOv2), dynamically modulating self-attention and feature aggregation to prioritize ID-sensitive local patterns (e.g., textures, logos) while suppressing irrelevant variations (e.g., viewpoint, illumination).

ularized by LLMs [74], where a model can rapidly adapt to new tasks during inference by conditioning on a small number of examples (*i.e.*, prompts). Recently, researchers have begun to explore this concept in multi-modal scenarios [1, 5, 22, 78]. In these multi-modal in-context learning frameworks, the model is provided with pairs of images and text prompts, guiding it to perform specific tasks, *e.g.*, image classification/OCR [78] and visual question answering [22] by leveraging cross-modal cues. We leverage the ICL to understand image/label pairs and dynamically generate visual prompts for VFMs toward ID-discriminative features.

Generalizable Object ReID Datasets. Generalizable ReID remains underexplored due to the absence of diverse benchmarks. The sole prior work [39] introduces a lab-controlled dataset with 180 instances from 50 categories, which removes background to enhance self-supervised learning models. However, its limited scale (180 instances) and artificial environments hinder practical evaluation. Moreover, precise object segmentation is often infeasible in real-world scenarios. Our work addresses these gaps with a large-scale dataset (10K instances) and a novel visual prompting framework for generalizable object ReID.

3. The Proposed Approach

3.1. Problem Formulation

Since we propose a new task, it is necessary to first formulate our task. Traditional ReID methods require category-specific training with extensive labeled data, limiting generalization to new object categories, while self-supervised models learn generic semantics, lacking fine-grained ID patterns.

In generalizable object ReID, we aim to learn a universal feature extractor $\phi(\cdot)$ that adapts to unseen object categories using only a small support set of examples. During training, the model is exposed to a base dataset $\mathcal{D}_{\text{base}} = \{(x_i, y_i, c_i)\}$, where each image x_i belongs to an instance-level identity y_i within a known object category $c_i \in \mathcal{C}_{\text{base}}$ (*e.g.*, backpacks, shoes). Critically, identities are unique within their categories, and instances across different categories are inherently distinct.

At test time, for a novel category $c' \in \mathcal{C}_{\text{novel}}$ (disjoint from $\mathcal{C}_{\text{base}}$), the model is provided with a support set $\mathcal{S} = \{(x_i, x_j, y_{ij})\}$ containing labeled positive ($y_{ij} = 1$) and negative ($y_{ij} = 0$) pairs, where labels indicate whether x_j and x_i share the same instance ID. The feature extractor $\phi(x|\mathcal{S})$ produces discriminative representations for query/gallery images from c' , enabling accurate similarity computation (*e.g.*, cosine distance) solely conditioned on \mathcal{S} . Our goal is to retrieve another image of the same instance, given a query image of one object instance. With this goal, it's not necessary to compare instances from different categories because cross-category comparison can never retrieve the correct result. The key challenge of object ReID is to distinguish subtle differences across different instances belonging to the same category while being invariant to background, lighting, or pose.

Therefore, a foundational assumption is category-aware inference: during deployment, the object category c' is known a priori (*e.g.*, via a pre-trained detector), ensuring that cross-instance comparisons are restricted to within-category pairs. This aligns with real-world ReID pipelines, where an upstream detection stage first filters candidates to a spe-

cific category (e.g., shoes), drastically reducing the search space and avoiding redundant cross-category matches (e.g., comparing a shoe to a backpack). Consequently, our framework does not need to handle cross-category ambiguity, as identities are only compared within the same category.

Similar to few-shot learning paradigms [58, 62], generalization to novel categories is achieved via dynamic conditioning on the support set \mathcal{S} , which guides $\phi(\cdot)$ to emphasize category-specific discriminative features (e.g., shoe tread patterns, bag stitching details). This parameter-free adaptation mirrors real-world scalability requirements, where deploying ReID systems for new categories can avoid costly retraining.

3.2. In-Context Visual Prompt Generation

In-context learning (ICL) [20] enables models to infer task-specific rules from provided examples without parameter updates. Unlike traditional supervised learning, ICL leverages the inherent reasoning capability of pre-trained LLMs to dynamically adapt to new tasks through sequential prompting. For ReID, this paradigm offers a critical advantage: the ability to encode *identity-discriminative priors* directly from visual context (e.g., “match objects based on logo details”) while avoiding costly fine-tuning of large vision models.

As shown in Fig. 2(a), our method employs a frozen LLM (e.g., LLaMA [60]) to process in-context example pairs and generate semantic guidance for ReID. Given a support set $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{x}_j, y_{ij})\}$ of positive ($y_{ij} = 1$) and negative ($y_{ij} = 0$) image pairs, we first encode each image \mathbf{x}_i into visual tokens using a pre-trained vision encoder (e.g., DINOv2 [53]). To mitigate computational overhead from excessive tokens (e.g., too many input pairs), we introduce a Query-based Connector (Q-Former), inspired by BLIP-2 [41], which compresses each image into a fixed set of N latent tokens. For a pair $(\mathbf{x}_i, \mathbf{x}_j)$, the connector outputs two compressed token sequences $\mathbf{I}_i, \mathbf{I}_j \in \mathbb{R}^{N \times d}$, which are concatenated into a unified sequence:

$$\mathbf{T}_{ij} = [\mathbf{I}_i; \mathbf{I}_j; \mathbf{L}_{ij}] \in \mathbb{R}^{(2N+1) \times d}, \quad (1)$$

where $\mathbf{L}_{ij} \in \mathbb{R}^d$ is a learnable embedding indicating the pair’s label (positive/negative). For K pairs, the full input sequence becomes $\mathbf{T}_{\text{ctx}} = [\mathbf{T}_{ij}^{(1)}; \dots; \mathbf{T}_{ij}^{(K)}]$, forming a contextualized prompt. Instead of adopting LLaVA-style projection [45], the Q-Former can significantly reduce the number of visual tokens, N for Q-Former can be much smaller than the number of feature tokens of the pre-trained model in LLaVA, e.g., 256 for a single image in ViT.

The LLM processes \mathbf{T}_{ctx} to predict the next token in an auto-regressive manner. Crucially, we mask the loss to only supervise the label tokens \mathbf{L}_{ij} , preserving the LLM’s pre-trained semantic knowledge while aligning it for ReID. The

training loss for a sequence of K pairs is:

$$\mathcal{L}_{\text{ICL}} = - \sum_{k=1}^K \log P \left(\mathbf{L}_{ij}^{(k)} \mid \mathbf{T}_{\text{ctx}}^{<k}, \mathbf{I}_i^{(k)}, \mathbf{I}_j^{(k)} \right), \quad (2)$$

where $\mathbf{T}_{\text{ctx}}^{<k}$ denotes all preceding pairs in the context. This forces LLM to reason across multiple pairs, identifying discriminative patterns (e.g., “logo consistency matters more than color”) that generalize beyond individual examples.

To generate adaptive prompts that encode ID-discriminative knowledge, we append a set of M learnable visual prompt tokens $\mathbf{P}_{\text{learn}} \in \mathbb{R}^{M \times d}$ to the end of the input sequence \mathbf{T}_{ctx} . These tokens, initialized randomly, are jointly optimized with the connector and LM head during training. The full input to the LLM becomes:

$$\mathbf{T}_{\text{full}} = [\mathbf{T}_{\text{ctx}}; \mathbf{P}_{\text{learn}}] \in \mathbb{R}^{(2NK+K+M) \times d}. \quad (3)$$

The LLM processes \mathbf{T}_{full} to contextualize the learnable prompts with the provided example pairs. The final hidden states corresponding to $\mathbf{P}_{\text{learn}}$ are then fed into a lightweight Visual Head—a two-layer MLP—to produce the task-specific visual prompts:

$$\mathbf{P}_{\text{task}} = \text{MLP}(\mathbf{P}_{\text{learn}} \cdot \mathbf{W}_{\text{LLM}}^\top) \in \mathbb{R}^{M \times d_{\text{vision}}}, \quad (4)$$

where \mathbf{W}_{LLM} projects the LLM’s hidden dimension d to the vision model’s feature dimension d_{vision} . These prompts \mathbf{P}_{task} implicitly encode *how to compare* instances for ReID, distilling identity-sensitive cues (e.g., “focus on texture consistency”) from the in-context pairs.

3.3. Generalizable Object Re-Identification

Vision foundation models like DINOv2 [53], pre-trained on large-scale datasets, learn rich visual-semantic representations that generalize across domains. However, while these models excel at high-level semantic tasks (e.g., classification or retrieval), their features lack the *fine-grained discriminability* required for ReID. For instance, DINOv2 may group images of “backpacks” by color or shape but fail to distinguish subtle ID-specific traits (e.g., logo placement or stitching patterns). Directly fine-tuning such models on ReID data risks overfitting to specific categories and degrading their generalization ability.

To adapt the pre-trained vision model for ReID, we inject the learned visual prompts $\mathbf{P}_{\text{task}} \in \mathbb{R}^{M \times d_{\text{vision}}}$ into each transformer layer [34], as presented in Fig. 2(b). Let $\mathbf{Z}_l \in \mathbb{R}^{(H \times W + 1) \times d_{\text{vision}}}$ denote the input tokens at layer l , where $H \times W$ are spatial dimensions and $+1$ corresponds to the [CLS] token. The prompts \mathbf{P}_{task} are concatenated with \mathbf{Z}_l to form an augmented token sequence:

$$\mathbf{Z}'_l = [\mathbf{Z}_l; \mathbf{P}_{\text{task}}] \in \mathbb{R}^{(H \times W + 1 + M) \times d_{\text{vision}}}.$$

The self-attention mechanism then computes interactions between all tokens, allowing the prompts to dynamically

Dataset	Categories		Instances	Images	Variation
	Seen	Unseen			
MVImgNet [72]	5	233	205K	800K	P
CUTE [39]	10	40	180	17K	L/P/B
PetFace [57]	5	8	257K	1M	L/P/B
MSMT17 [64]	1 (Person)	-	4K	126K	L/P/B
Market1501 [75]	1 (Person)	-	1.5K	32K	L/P/B
VeRi-776 [49]	1 (Vehicle)	-	776	50K	L/P/B
ShopeID10K (ours)	7	27	10K	45K	L/P/B

Table 1. Comparison of object ReID datasets, with variations in lighting (L), pose (P), and background (B). While MVImgNet has more categories/instances, it only contains pose variations.

reweight spatial features—e.g., amplifying regions with ID-sensitive details (logos, textures) while suppressing irrelevant areas (backgrounds, occlusions). Crucially, the original ViT parameters remain frozen; only the prompts \mathbf{P}_{task} (generated per support set \mathcal{S}) modulate the feature space.

To train the framework, we propose two loss functions that preserve the vision model’s generalization while enhancing ID information.

ReID Loss: We adopt triplet loss to optimize global feature discriminability. For a mini-batch of images within the same category, we sample triplets $(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n)$, where \mathbf{x}_a (anchor) and \mathbf{x}_p (positive) share the same instance ID, and \mathbf{x}_n (negative) has a different ID. The loss enforces a margin α between positive and negative similarities:

$$\mathcal{L}_{\text{ID}} = \sum_{i=1}^B \max \left(0, \alpha - \text{sim}(\phi(\mathbf{x}_a^i), \phi(\mathbf{x}_p^i)) + \text{sim}(\phi(\mathbf{x}_a^i), \phi(\mathbf{x}_n^i)) \right), \quad (5)$$

where B is the number of triple pairs and $\phi(\mathbf{x})$ is [CLS] token embedding. Triplet loss is preferred over classification losses (e.g. ArcFace [19], AdaFace [37]) or contrastive loss [67] as it only penalizes violations of the margin constraint, imposing softer updates that preserve the pre-trained model’s semantic prior, while [19, 37, 67] push the features to align better that may degrade the generalization ability.

Patch Alignment Loss: To further refine local feature discriminability, we compute the optimal transport (OT) distance between patch embeddings of image pairs. For a pair $(\mathbf{x}_i, \mathbf{x}_j)$, let $\mathbf{F}_i, \mathbf{F}_j \in \mathbb{R}^{(H \times W) \times d_{\text{vision}}}$ be their patch-level features (excluding [CLS]). The OT distance reflects the matching cost across local patches. We adopt the inexact proximal point method [68] to compute OT distance as $D_{\text{OT}}(\cdot)$. Based on this, we define the alignment loss:

$$\mathcal{L}_{\text{align}} = \sum_{(\mathbf{x}_i, \mathbf{x}_j)} \left[\mathbb{I}(y_{ij} = 1) \cdot D_{\text{OT}}(\mathbf{F}_i, \mathbf{F}_j) - \mathbb{I}(y_{ij} = 0) \cdot D_{\text{OT}}(\mathbf{F}_i, \mathbf{F}_j) \right], \quad (6)$$

where $\mathbb{I}(\cdot)$ is an indicator function, 1 if condition holds and otherwise 0, and $D_{\text{OT}}(\mathbf{F}_i, \mathbf{F}_j)$ denotes the aforementioned



Figure 3. Examples of 4 categories, 2 instances per category, and 3-4 images per instance from our ShopID10K dataset.

OT distance. This formulation encourages positive pairs, i.e., $\mathbb{I}(y_{ij} = 1)$ to align spatially (e.g., matching similar regions such as logos), while separating negative pairs, i.e., $\mathbb{I}(y_{ij} = 0)$, thereby enhancing local feature consistency.

3.4. Training and Inference

The framework is trained end-to-end with a composite loss that unifies in-context learning, ID discrimination, and local feature alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ID}} + \lambda_{\text{ICL}} \mathcal{L}_{\text{ICL}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}},$$

where λ^* balance the contributions of each objective.

During the testing of a novel category c' , the model processes the support set \mathcal{S} through the LLM-based prompt generator to produce category-specific visual prompts \mathbf{P}_{task} . These prompts are cached and reused for all query-gallery comparisons within c' , incurring only a one-time computational cost for prompt generation. Subsequent feature extraction and similarity computation follow standard ReID pipelines, with the frozen VFM modulated by \mathbf{P}_{task} . Consequently, the cached prompts enable instant deployment to new categories without re-computation, which is ideal for dynamic environments (e.g., retail inventory updates).

4. Experiments

Datasets. We extensively evaluate on seven datasets spanning both general and specific domains, as listed in Tab. 1. MVImgNet [72] provides multi-view object videos captured under controlled lighting, offering pose variations but lacking background/lighting diversity. To adapt it for ReID, we employ GroundingDINO [48] for detection and SAM 2 [55] to remove the background. We then select the four most different frames, focusing on intrinsic ID characteristics. CUTE [39] captures lab-controlled images with varying illuminations/backgrounds per object, yet its limited scale (180 instances) restricts practical use. PetFace [57] aggregates

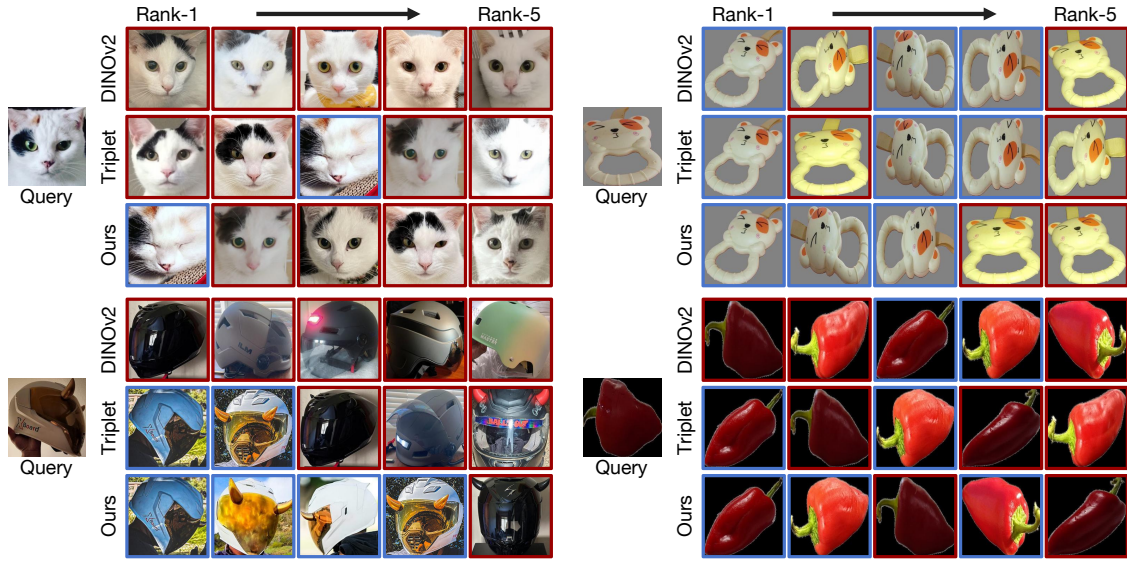


Figure 4. The ranking results of different models on PetFace/MVImageNet/ShopID10K/CUTE datasets. Blue/Red boxes indicate correct/wrong instance ID. The results show that VICP is largely superior to baselines in finding the fine-grained information of local details.

257K pet images from Internet. For domain-specific baselines, MSMT17 [64], Market1501 [75] and VeRi-776 [49] represent person/vehicle ReID benchmarks, although they only contain one category.

ShopeID10K. To establish a comprehensive benchmark for generalizable object ReID, we first define 34 daily object categories, including backpack, bicycle, *etc.* For each category, we collect product images from Amazon customer reviews by searching category keywords. Crucially, we treat all images uploaded by the same reviewer for a specific product as sharing the same instance ID, simulating real-world scenarios where multiple images of an object instance are captured by everyday users. After detection and filtering, we ensure each instance has at least 3 images, leading to 10K instances and 45K images. The key advantage of ShopeID10K is the diversity, as shown in Fig. 3. Each instance exhibits natural variations in lighting, occlusion, pose, and background.

Implementation Details. All experiments are conducted on two H100 GPUs with a fixed learning rate of 10^{-4} , weight decay of 10^{-4} , $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The margin for triplet loss is set to 0.1. We use pre-trained DINOv2 ViT-small [21] as the backbone. We train the model with the batch size 256 for 10 epochs with the image size of 224×224 and random horizontal flip as data augmentation. During training, we randomly sample 64 positive/negative pairs. The number of visual tokens for Q-former is 32.

4.1. Qualitative Results

In Fig. 4, we visualize retrieval results of unsupervised DINOv2, Triplet+, and our methods on ShopID10K. DINOv2 predominantly retrieves images sharing shape similarity or

semantic attributes (*e.g.*, matching object categories) but fails to prioritize identity-specific features, resulting in frequent false positives. In contrast, triplet fine-tuning remarkably mitigates such errors by refining the embedding space to emphasize discriminative ID cues. These qualitative comparisons substantiate the superior discriminative capability of our method: it consistently retrieves identity-consistent instances across extreme variations in viewpoint and illumination while suppressing semantically similar distractors.

4.2. Quantitative Results

We conduct extensive evaluations under the proposed generalizable object ReID paradigm. A subset of categories from each dataset is selected as base categories for training, and the rest are novel categories for testing. We repeat this process 5 times and report the average across different runs.

Baselines. There are three types of baselines for comparison: (1) fully supervised models, which are trained on labeled data encompassing all target categories (including both seen and unseen classes during training). These models establish the empirical performance upper bound for ReID systems. (2) unsupervised or weakly supervised representation learning models [3, 4, 10, 24, 33, 35, 40, 54], such as DINO [35] and CLIP [54], dedicated to extracting generalizable visual features. For example, for downstream tasks like image retrieval, and Unicom [3] proposes clustering-based feature refinement. (3) To build strong baselines, we fine-tune DINOv2 on the base categories with learnable visual prompts [34], augmented with metric loss functions to encourage discriminative representations, *i.e.*, ArcFace [19], AdaFace [37] and supervised contrastive learning (SCL) [36].

	Verification		Identification	
	AUC	ACC	mAP	Top-1
PetFace [57]	92.1	-	-	-
MegaDescriptor [11]	83.7	-	-	-
Supervised	95.5	89.3	57.7	56.3
CLIP [54]	71.5	64.6	7.1	4.4
OpenCLIP [33]	73.4	67.3	7.6	5.8
DINO [10]	82.2	74.7	19.7	16.5
DreamSim [24]	84.2	76.4	17.9	14.9
Unicom [3]	73.3	67.2	7.0	5.4
I-JEPA [4]	74.0	67.1	10.9	8.7
DINOv2 [35]	71.6	65.9	6.5	5.1
Arcface [19]	89.1	81.5	46.6	44.3
Adaface [37]	89.3	82.4	46.9	44.4
SCL [36, 67]	91.1	83.1	46.3	43.4
Triplet [7]	91.7	84.9	48.2	45.9
Triplet+ [7]	92.5	85.6	49.8	47.7
VICP (ours)	93.5	86.0	51.2	49.7

Table 2. Results on PetFace [57]. [Key: **First**, Second]

In particular, Triplet+ fine-tunes the model by incorporating few-shot examples based on the model trained with triplet loss, enabling targeted optimization of the ReID framework.

Results on PetFace (Tab. 2). PetFace [57] collects multiple IDs of the same pet from web sources, comprising 13 categories, 170K unique IDs, and 1 million images. PetFace adopts two evaluation metrics aligned with face recognition paradigms: (1) **Verification** constructs positive/negative pairs per category and computes AUC/Accuracy via 10-fold cross-validation. Our experiments strictly follow PetFace’s verification protocol. (2) **Identification** measures the ability to retrieve the same instance ID gallery images from query samples. Instead of using training data in PetFace, we align with standard person ReID practices. Each image in the test set is used as a query and the remaining forms the gallery, quantified by top-1 accuracy and mAP. Five categories are randomly selected as base categories for training, while the remainder forms novel categories for testing.

Supervised fine-tuning on full data significantly outperforms both original PetFace and MegaDescriptor [11] baselines, demonstrating the efficacy of large-scale pre-trained VFMs. Unsupervised pre-trained models exhibit suboptimal ReID performance. Notably, fine-tuning on ReID data enhances generalization even to novel categories.

Triplet loss surpasses alternative loss functions by selectively optimizing challenging ReID pairs while preserving pre-trained representations’ inherent generalizability. In contrast, other losses will harm the embeddings because they will always minimize loss. Triplet+ further enhances novel category performance via integrating few-shot pairs. Our models, without further parameter updates, achieves superior generalization compared to fine-tuning-based approaches.

Results on MVImageNet and ShopID10K (Tab. 3). Similar to PetFace, five categories are randomly selected as

	MVImageNet		ShopID10K		
	mAP	Rank-1	mAP	Rank-1	Rank-5
Supervised	79.2	88.5	62.6	71.2	89.8
CLIP [54]	39.4	55.6	37.1	48.6	72.1
OpenCLIP [33]	41.8	57.8	40.2	51.3	75.4
DINO [10]	53.0	69.8	41.2	54.4	75.9
DreamSim [24]	56.1	71.7	44.4	56.9	78.7
Unicom [3]	45.5	61.1	43.8	54.6	78.2
I-JEPA [4]	41.2	58.1	32.7	45.6	67.3
DINOv2 [35]	47.3	64.1	34.1	45.7	67.5
Arcface [19]	72.2	58.5	45.3	56.5	78.0
Adaface [37]	73.0	59.0	45.6	57.2	78.7
SCL [36, 67]	68.2	80.4	46.9	56.6	79.5
Triplet [7]	72.9	82.1	50.3	63.1	80.9
Triplet+ [7]	<u>73.2</u>	<u>84.0</u>	<u>54.8</u>	<u>67.4</u>	<u>85.6</u>
VICP (ours)	74.9	85.4	58.5	68.4	87.5

Table 3. Results on MVImageNet [72] and our ShopID10K.

base categories. MVImageNet, constructed from multi-view object videos, primarily captures pose variations with minimal environmental complexity, resulting in relatively lower challengingness. Interestingly, even slight fine-tuning on its constrained category set yields substantial performance gains, indicating its effectiveness as a benchmark for view-invariant representation learning. In contrast, our newly introduced ShopID10K exhibits extreme diversity across backgrounds, lighting conditions, occlusion patterns, and viewpoints. While PetFace focuses on constrained pet facial recognition, our method observes the same trends and achieves greater improvements on ShopID10K.

Results on CUTE (Tab. 4). CUTE dataset provides laboratory-controlled multi-view imagery of objects under varying illumination and pose conditions, designed as a benchmark for intrinsic object similarity metrics. Evaluation is conducted via pairwise comparisons across instances, with performance measured using ReID metrics: mAP and top-1 accuracy. We adopt three distinct evaluation regimes: (1) In-the-wild—images contain simultaneous variations in background, pose, and illumination; (2) Illumination and (3) Pose use controlled illumination or pose variations to assess representation robustness under isolated conditions.

Conventional fine-tuning approaches suffer from severe base-category overfitting, significantly impairing in-the-wild generalization. In contrast, our method achieves consistent performance gains across all regimes, with a 2% improvement in mAP under in-the-wild settings. These results highlight its ability to disentangle intrinsic object features from confounding environmental factors.

Results on Person and Vehicle ReID. We further validate our approach on popular person/vehicle ReID benchmarks. For person ReID, as shown in Tab. 5, our method achieves competitive performance against state-of-the-art person ReID models, where PASS [76] and DINO [10] are pre-trained on human datasets. For vehicle ReID, VICP

	In-the-wild		Illumination		Pose	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
CLIP [54]	70.7	59.6	72.2	71.8	77.8	95.1
OpenCLIP [33]	75.0	68.4	73.6	75.9	79.3	95.7
DINO [10]	71.6	57.1	72.8	68.9	81.1	97.2
DreamSim [24]	73.0	59.6	70.0	65.2	83.4	97.9
Unicom [3]	75.1	69.6	75.3	77.2	82.4	97.3
I-JEPA [4]	61.3	39.6	65.0	57.6	74.6	92.1
DINOv2 [35]	80.5	74.7	81.3	80.4	83.4	96.2
Arcface [19]	78.2	74.3	75.1	75.0	81.5	96.5
Adaface [37]	78.4	74.0	77.6	78.6	82.0	96.3
SCL [36, 67]	79.4	75.9	77.7	77.8	81.6	96.4
Triplet [7]	<u>80.9</u>	<u>76.5</u>	<u>84.2</u>	<u>85.9</u>	<u>84.1</u>	<u>97.5</u>
VICP (ours)	82.5	77.3	89.8	89.2	87.6	98.9

Table 4. Results on CUTE [40].

	MSMT17		Market1501	
	mAP	Rank-1	mAP	Rank-1
DINO [10]	66.1	84.6	<u>91.0</u>	<u>96.0</u>
BOT [50]	50.2	74.1	85.9	94.5
MGN [61]	63.7	85.1	87.5	95.1
SCSN [17]	58.5	83.8	88.5	95.7
ABDNet [13]	60.8	82.3	88.3	95.6
AAformer [77]	63.2	83.6	87.7	95.4
TransReID [29]	63.6	82.5	87.4	94.6
PASS [76]	<u>69.1</u>	<u>86.5</u>	92.2	96.3
VICP (ours)	75.3	89.2	90.3	95.5

Table 5. Results on person ReID datasets.

outperforms TransReID [29] on the VeRi-776 dataset [49], achieving a higher mAP (81.2 vs. 79.6) and slightly better R1 (97.1 vs. 97.0). While existing methods primarily optimize for dataset-specific biases, our approach enables stronger generalization to novel categories, highlighting its adaptability beyond domain-specific constraints.

4.3. Ablation Study

We show three ablation studies in Tab. 6.

Ablation on Different Components. To evaluate our framework, we conduct incremental ablation studies starting from the unsupervised DINOv2 baseline [i] and gradually adding our components. Even minimal fine-tuning on ReID datasets [ii]—despite category mismatches—yields noticeable gains, demonstrating the strong transferability of pre-trained visual priors to identity-sensitive tasks. Few-shot supervision (Triplet+ [iii]) brings moderate improvement, though its scalability is limited in real-world scenarios.

While LLMs exhibit promising in-context learning (ICL [v]) capabilities, performing pairwise query-gallery inference directly is computationally impractical for large-scale ReID. Instead, we utilize LLMs as auxiliary semantic guides, encoding relational structures into ICL-based visual prompts [iv]. This enhances the VFM’s cross-domain reasoning without introducing LLM inference costs at runtime.

	PetFace		ShopID10K	
	mAP	Top-1	mAP	Rank-1
Supervised	57.7	56.3	62.6	71.2
[i] Unsupervised [35]	6.5	5.1	34.1	45.7
[ii]+Triplet	48.2	45.9	50.3	63.1
[iii] Triplet+	49.8	47.7	54.8	67.4
[iv]+ICL visual prompts	50.8	48.6	56.3	67.1
[v] ICL from LLM	42.1	37.5	39.6	48.5
[vi]+Patch Align	51.2	49.7	58.5	68.4
$K=32$	50.4	47.9	55.7	67.5
64	51.2	49.7	58.5	68.4
128	48.8	47.0	56.8	66.9
$N=16$	49.9	48.1	55.9	67.0
32	51.2	49.7	58.5	68.4
64	48.6	47.5	57.1	67.6

Table 6. Results of ablation study.

Finally, the patch alignment loss [vi] enforces spatial consistency across discriminative regions, improving feature localization and leading to better patch-level embedding separation, which ultimately boosts ReID accuracy.

Ablation on Number of Examples. We investigate the impact of the number of in-context examples with $K \in \{32, 64, 128\}$ examples per category. Our method achieves peak performance at 64 examples, with 32 examples yielding marginally lower results and 128 examples causing degradation. With 32 examples, the LLM struggles to capture nuanced ID discrimination patterns, while it is too long for 128 examples to learn the complex patterns.

Ablation on Number of Latent Tokens. We vary the number of latent tokens $N \in \{16, 32, 64\}$ in the Query-based connector. The best performance is achieved at $N = 32$. A smaller N (16) lacks expressiveness, while a larger N (64) leads to longer, noisier inputs that hurt performance and increase cost. This aligns with our findings on in-context example count, indicating both overly simple and overly complex prompts degrade results. In supplementary, we show additional ablation in occlusion robustness, cross-domain generalization, and comparisons with few-shot learning methods.

5. Conclusion

We address the underexplored challenge of Generalizable Object Re-Identification by integrating large language models (LLMs) with vision foundation models. Key contributions include: (1) reformulating ReID as an in-context learning task with LLM-guided feature extraction from few-shot exemplars, and (2) introducing the first large-scale benchmark with 10K real-world instances across 34 categories—surpassing prior lab-controlled datasets in diversity and complexity. Our method outperforms self-supervised baselines on novel categories while requiring less labeled data than conventional ReID. In the future, we aim to tackle the harder zero-shot ReID setting without exemplars.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in neural information processing systems*, 2022.
- [2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.
- [3] Xiang An, Jiankang Deng, Kaicheng Yang, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Unicom: Universal and compact representation learning for image retrieval. In *ICLR*, 2023.
- [4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.
- [5] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- [6] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. *arXiv preprint arXiv:2305.13653*, 2023.
- [7] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.
- [8] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020.
- [10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- [11] Vojtěch Čermák, Lukas Pícek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5953–5963, 2024.
- [12] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards modality-agnostic person re-identification with descriptive query. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15128–15137, 2023.
- [13] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8351–8361, 2019.
- [14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.
- [15] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15050–15061, 2023.
- [16] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [17] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Saliency-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3300–3310, 2020.
- [18] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.
- [19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [20] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.
- [21] 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.
- [22] Sivan Doherty, Shaked Perek, M Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbel, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision & language models. *arXiv preprint arXiv:2403.12736*, 2024.
- [23] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.
- [24] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.
- [25] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1060–1069, 2022.
- [26] Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. Language-guided hierarchical fine-grained image forgery detection and localization. In *International Journal of Computer Vision*, December 2024.
- [27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022.

- [28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019.
- [29] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.
- [30] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17531, 2024.
- [31] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXi-ang Zhang. Clothing status awareness for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11895–11904, 2021.
- [32] Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7509–7524, 2022.
- [33] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.
- [34] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022.
- [35] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. *arXiv preprint arXiv:2412.16334*, 2024.
- [36] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- [37] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.
- [38] Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. Sapiensid: Foundation for human recognition. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Nashville, TN, June 2025.
- [39] Klemen Kotar, Stephen Tian, Hong-Xing Yu, Dan Yamins, and Jiajun Wu. Are these the same apple? comparing images based on object intrinsics. In *Advances in Neural Information Processing Systems*, 2023.
- [40] 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, Nov. 2023.
- [41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [42] Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, Yiyang Su, Pegah Varghaei, Kai Wang, Xingguang Zhang, Stanley Chan, Arun Ross, Humphrey Shi, Zhangyang Wang, Anil Jain, and Xiaoming Liu. Farsight: A physics-driven whole-body biometric system at large distance and altitude. In *In Proceeding of Winter Conference on Applications of Computer Vision*, Waikoloa, Hawaii, January 2024.
- [43] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, and Xiaoming Liu. Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In *In Proceeding of International Conference on Computer Vision*, Paris, France, October 2023.
- [44] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling clip with dual guidance for learning discriminative human body shape representation. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Seattle, WA, June 2024.
- [45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [46] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19366–19375, 2022.
- [47] Kangning Liu, Weicheng Zhu, Yiqiu Shen, Sheng Liu, Narges Razavian, Krzysztof J Geras, and Carlos Fernandez-Granda. Multiple instance learning via iterative self-paced supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3355–3365, 2023.
- [48] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [49] Xinchun Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 869–884. Springer, 2016.
- [50] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
- [51] Kien Nguyen, Clinton Fookes, Sridha Sridharan, Feng Liu, Xiaoming Liu, Arun Ross, Dana Michalski, Huy Nguyen, Debayan Deb, Mahak Kothari, Manisha Saini, Dawei Du, Scott McCloskey, Gabriel Bertocco, Fernanda Andal ´o, Terrence E. Boulton, Anderson Rocha, Haidong Zhu, Zhaocheng

- Zheng, Ram Nevatia, Zaigham Randhawa, Sinan Sabri, and Gianfranco Doretto. Ag-reid 2023: Aerial-ground person re-identification challenge results. In *In Proceeding of International Joint Conference on Biometrics*, Ljubljana, Slovenia, September 2023.
- [52] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.
- [53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [54] 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. In *International conference on machine learning*, pages 8748–8763, 2021.
- [55] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [56] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.
- [57] Risa Shinoda and Kaede Shiohara. Petface: A large-scale dataset and benchmark for animal identification. In *European Conference on Computer Vision*, pages 19–36. Springer, 2024.
- [58] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, 2017.
- [59] Yiyang Su, Yunping Shi, Feng Liu, and Xiaoming Liu. Hamobe: Hierarchical and adaptive mixture of biometric experts for video-based person reid. In *In Proceeding of International Conference on Computer Vision*, Honolulu, Hawaii, October 2025.
- [60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [61] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.
- [62] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019.
- [63] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5549–5560, 2022.
- [64] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.
- [65] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.
- [66] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Biao Gong, Longxiang Tang, Xiang Wang, Haonan Qiu, Hengjia Li, Shuai Tan, Yingya Zhang, et al. Dreamrelation: Relation-centric video customization. *arXiv preprint arXiv:2503.07602*, 2025.
- [67] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.
- [68] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR, 2020.
- [69] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.
- [70] Mouxiang Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022.
- [71] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 200–210, 2024.
- [72] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023.
- [73] Dominik Zapletal and Adam Herout. Vehicle re-identification for automatic video traffic surveillance. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 25–31, 2016.
- [74] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- [75] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.
- [76] 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 *European conference on computer vision*, pages 198–214. Springer, 2022.

- [77] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [78] Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. VL-ICL bench: The devil in the details of multimodal in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025.