DePro: Domain Ensemble using Decoupled Prompts for Universal Cross-Domain Retrieval

Kaixiang Chen
School of Computer Science and Engineering, Southeast University
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
Nanjing, China
kxchen@seu.edu.cn

Pengfei Fang*
School of Computer Science and Engineering, Southeast University
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
Nanjing, China
fangpengfei@seu.edu.cn

Hui Xue*
School of Computer Science and Engineering, Southeast University
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
Nanjing, China
hxue@seu.edu.cn

Abstract

This paper investigates the potential of vision-language models (VLMs) in addressing the challenges of universal cross-domain retrieval (UCDR), where queries originate from unseen domains or classes. For adapting VLMs like CLIP to downstream tasks prompt tuning is frequently adopted as a lightweight alternative to full fine-tuning. However, this approach often struggles with the domain and semantic shifts inherent in UCDR. To overcome these limitations, we propose a novel prompt decoupling strategy that separates prompts into universal domain prompts (UDPs) and class prompts (CPs). Specifically, UDPs are designed to unify features from both seen and unseen domains, while CPs are tailored to capture class-specific visual characteristics, enabling robust retrieval across both known and unknown classes. To ensure effective decoupling, we introduce a dedicated decoupling loss that enforces the domain-agnostic nature of CPs. Additionally, we employ a regulation loss to align features from the frozen CLIP domain with those of the universal domain by selectively integrating or excluding UDPs. This mechanism fosters a synergistic domain ensemble effect, enhancing retrieval generalization across diverse domains. Finally, we propose the domain-aware triplet-hard (DaTri) loss to mitigate overfitting by reducing the risk of class collapse. The proposed framework, referred to as **D**omain **E**nsemble using Decoupled **Pro**mpts (DePro), demonstrates state-of-the-art performance and effectively enhances the model's generalization capacity across unseen domains and classes, as validated through extensive experiments. Code is **here**.

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**.

*Co-corresponding authors

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3729946

Keywords

Vision-Language Models, Universal Cross-Domain Retrieval, Prompt Decoupling, Domain Ensemble

ACM Reference Format:

Kaixiang Chen, Pengfei Fang, and Hui Xue. 2025. DePro: Domain Ensemble using Decoupled Prompts for Universal Cross-Domain Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3729946

1 Introduction

Cross-domain retrieval (CDR) [12, 21] is crucial for the advancement of information retrieval (IR) systems [3, 13, 35], which aims to retrieve relevant instances from one domain in response to queries originating from another domain. While CDR has demonstrated promising results, it typically operates under the assumption that the testing phase involves known domains and classes. This assumption significantly restricts the applicability of CDR methods in real-world scenarios, where it is natural to encounter unseen domains and classes. To address this limitation, we shift our focus to more practical yet challenging CDR scenarios that utilize data from unseen domains, unseen classes, or a combination of both, as queries to retrieve semantically similar examples from the *Real* domain. These scenarios are termed as universal cross-domain retrieval (UCDR) [8, 26, 32], U^{C}CDR and U^{D}CDR (see **Fig.1**).

Recently, pretrained vision-language models (VLMs), such as CLIP [28], have exhibited exceptional performance across a diverse range of downstream tasks [22, 36, 41]. Their remarkable "zero-shot" generalization capabilities position them as promising tools to address the UCDR problem. However, pretrained VLMs, even when enhanced with prompt tuning [18, 20, 37, 41], may struggle to effectively tackle the challenges of domain shift and semantic shift in UCDR. Currently, there are few prompt tuning-based studies focused on the UCDR problem [8], highlighting a significant demand and challenge in developing effective prompt tuning methods to overcome these shifts. To address this, we introduce a novel prompt decoupling strategy that divides prompts into two types: universal domain prompts (UDPs) and class prompts (CPs), designed to separately address domain-specific and class-specific challenges.

In general prompt tuning methods such as CoOp [41] and VPT [18], prompts play a crucial role in tailoring model outputs toward

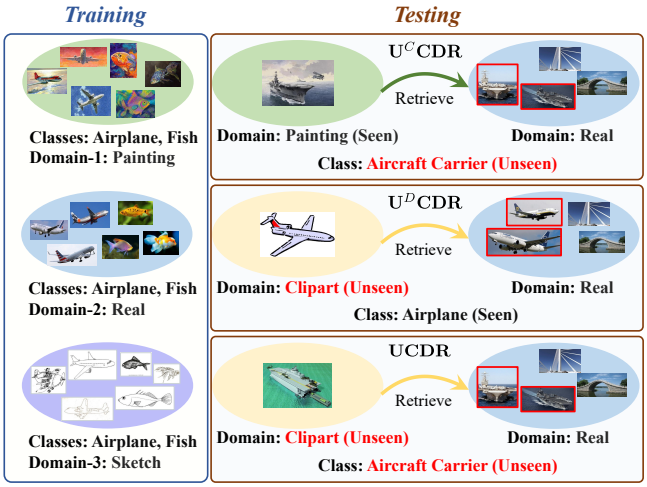


Figure 1: Illustration of three practical CDR scenarios. After training, retrievals are performed using data from unseen classes (U^{C}CDR), unseen domains (U^{D}CDR), or both (UCDR), targeting data of the same class within the *Real* domain.

target domains. When applied to UCDR, which involves multiple training domains, an intuitive approach is to utilize domain-specific prompts (see **Fig. 2-a**) for training and utilize all these prompts collectively for handling unseen domains at inference. This strategy has been also adopted in ProS [8], the only UCDR method based on VLMs with prompt tuning. Given the inherent focus on cross-domain retrieval, the primary objective of UCDR is to synthesize domain-invariant features. We argue that domain-specific prompts, while tailored to their respective domains, may fall short in promoting the synthesis of domain-invariant features during inference, as they risk overfitting to the source domains and thus limit generalization to unseen domains. Therefore, we introduce universal domain prompts (UDPs, see **Fig. 2-b**), which provide a consistent representation across diverse domains, ensuring a unified embedding space capable of accommodating unseen domains during inference. Unlike domain-specific prompts, UDPs are jointly optimized across all source domains, enabling them to capture the shared characteristics of the source domains and effectively map instances into a more robust, universal domain. This is supported by the empirical study shown in **Fig. 5**.

While UDPs effectively address the domain shift, tackling unseen classes requires a complementary approach. To this end, we introduce class prompts (CPs), designed to extract class-specific insights based on individual images. Specifically, we input each image into a trainable adaptive semantic-prompter (ASP) to generate adaptive CPs, which enables the generation of CPs for previously unseen classes. It's worth noting that the generated CPs may inevitably incorporate domain-specific information, leading to overlapping roles or even conflicts with the UDPs. To facilitate prompt decoupling, we introduce a decoupling loss defined as a cross-entropy loss applied to the outputs of the text and image encoders after removing UDPs. This loss ensures that the visual outputs, derived from prompts containing only CPs, are aligned with the frozen CLIP visual outputs, thereby maintaining the domain-agnostic nature of the learned CPs. Note that visual outputs with UDPs represent the learned universal domain, while outputs without UDPs correspond to the frozen CLIP domain. During training, we introduce a regulation loss between these outputs, facilitating a domain ensemble effect, improving the model's generalization ability.

This paper also systematically investigates the triplet-hard loss [16], a loss function widely used in metric learning tasks [5, 23, 24, 30]. Typically, it employs a *PK* sampler, selecting P classes and K images per class, forming a mini-batch of size $B = P \times K$. In the context of UCDR, we consider two variants: the *PK* domain sampler and the *DPK* domain sampler (**Fig. 2-c**). The *PK* domain sampler generates PK samples independently for each domain, while the *DPK* domain sampler ensures that each domain includes the same classes. Our results indicate that the triplet-hard loss based on the *DPK* domain sampler performs significantly better. This improvement likely arises because when the *DPK* domain sampler is adopted, the least similar positive samples to the anchor ***always*** come from other domains (89.88% in our experiments), which reduces the distance between domains during optimization and mitigates overfitting by preventing class collapse [4] when positive samples are from the same domain. Building on this, we introduce the domain-aware triplet-hard (DaTri) loss. In DaTri, the proportion of least similar positive samples from other domains is maintained at 100% using an in-domain mask, further enhancing the benefits of the *DPK* domain sampler.

The whole presented framework, referred to as **D**omain **E**nsemble using Decoupled **Pro**mpts (DePro). Experiments on three large-scale multi-domain datasets demonstrate the effectiveness of our DePro. In summary, our contributions are featured as follows: ❶ We introduce the DePro framework, which utilizes two types of prompts: UDPs for unifying image features across domains, ensuring consistent representation, and CPs for maintaining class-specific discrimination within the universal domain. ❷ We introduce a decoupling loss that ensures the generated CPs remain domain-agnostic. ❸ We propose a regulation loss to to play a domain ensemble effect, thereby enabling more generalizable retrieval. ❹ We systematically investigate the use of triplet-hard loss and propose a DaTri loss to amplify the benefits brought by the cooperation of triplet-hard loss and *DPK* domain sampler.

2 Related Work

Universal Cross-Domain Retrieval. While Cross-Domain Retrieval (CDR) [10, 17] has achieved promising progress, it typically assumes that the domains and classes involved in both the training and testing phases are known and have labeled data available. This assumption limits the generalizability of CDR methods to real-world scenarios, where new, unseen domains and classes are often encountered. To address this limitation, researchers have turned their attention to Universal Cross-Domain Retrieval (UCDR), which aims to facilitate retrieval across various unseen domains and unseen classes. SnMpNet [26] represents a pioneering effort in tackling UCDR by advocating for the representation of unseen classes based on their relational positions with seen classes, leveraging the mix-up technique to acquire a domain-invariant embedding. SASA [32] introduces a novel approach utilizing ViT [6] to preserve the global structural information of images. Moreover, ProS [8] marks the

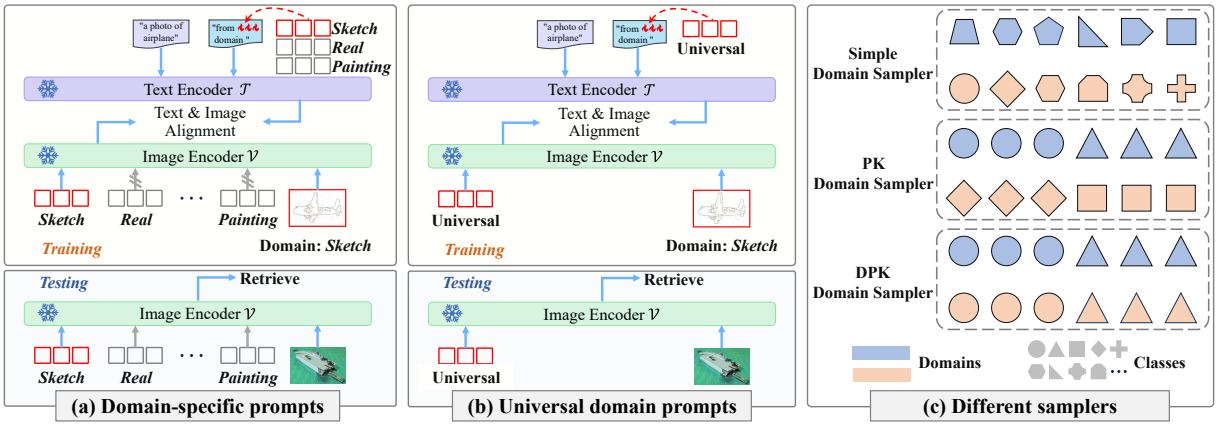


Figure 2: (a) During training, only the domain-related prompts are passed to the network, while during testing, all visual prompts are utilized. (b) Universal domain prompts are shared across all training domains and are also employed during testing. (c) Different samplers, from top to bottom: the simple domain sampler is used in ProS [8]; the *PK* domain sampler generates PK samples for each domain; the *DPK* sampler ensures that each domain contains the same classes.

initial exploration into fine-tuning VLMs by employing adaptive domain-specific prompts and tailored semantic prompts.

Downstream Adaptation of CLIP. As a milestone in vision-language learning, CLIP [28] has significantly enhanced various downstream tasks through prompt tuning techniques. Notable examples include CoOp [41], which adds prompts to the text encoder, and VPT [18], which appends prompts to the vision encoder. Additionally, MaPLe [19] builds upon CoOp and VPT by introducing multi-modal prompting to better align vision and language representations, achieving great success. Regarding adapters, CLIP-Adapter [11] pioneers a lightweight adapter module for generating adapted multi-modal features. Tip-Adapter [40] and CaFo [39] significantly reduce training costs by employing adapters via a key-value cache mechanism. CALIP [14] innovates with a parameter-free attention mechanism for cross-modal interactions. While these approaches have demonstrated impressive performance, directly applying them to the UCDR problem may not be sufficient to address the inherent domain and semantic shifts. To fill this gap, we introduce the DePro framework, which leverages decoupled prompts to effectively tackle these challenges, significantly enhancing CLIP's generalizable retrieval capability.

3 Method

In this section, we first introduce the concept of Universal Cross-Domain Retrieval (UCDR) and revisit CLIP as the preliminary. Then we present the details of our DePro framework.

3.1 Problem Definition

UCDR. In UCDR, the training setup includes at least two source domains ($N_S \geq 2$), collectively represented as $\mathbb{D}_S=\{\mathcal{D}_S^j\}_{j=1}^{N_S}$. Each source domain is defined as $\mathcal{D}_S^j = \{(x_i^j, y_i^j)\}_{i=1}^{P_j}$, where x_i^j denotes the i-th image out of a total of P_j images in the j-th source domain, and y_i^j represents its class label. All source domains share a common label space $\mathcal{Y}_S$. The *Real* domain $\mathcal{D}_R$, consisting of natural images, plays a dual role. It is partitioned into two subsets based on classes: $\mathcal{D}_R^+$, which contributes to training as part of $\mathbb{D}_S$ (i.e., $\mathcal{D}_R^+ \in \mathbb{D}_S$), and $\mathcal{D}_R^-$, which serves as the gallery set during testing. For evaluation, a query set $\mathcal{D}_Q = \{(x_i^q, y_i^q)\}_{i=1}^{P_q}$ is provided to enable image retrieval from $\mathcal{D}_R^-$, where P_q samples are drawn from an unseen domain and correspond to unseen classes. By default, the classes in $\mathcal{D}_R^-$ are identical to those in $\mathcal{D}_Q$, and we denote this scenario as `UnseenGallery`. However, in a realistic scenario, $\mathcal{D}_R^-$ may include additional classes, such as those from the training phase, which is refereed to as `MixedGallery`.

U^{D}CDR and U^{C}CDR are two degraded variations of UCDR. In U^{D}CDR, the query class is known from the training phase, but the domain of the query is unseen. Conversely, in U^{C}CDR, the query domain is known from the training phase, while the class of the query is unseen.

3.2 CLIP Preliminary

CLIP. Our approach is built upon CLIP, a pre-trained vision-language model (VLM) that leverages 400 million image-text pairs for contrastive pre-training. CLIP aligns texts and images using a image encoder $\mathcal{V}(\cdot)$ (ViT [6] by default) and a text encoder $\mathcal{T}(\cdot)$. For zero-shot inference with C classes, CLIP inserts all class names into a pre-defined textual template, e.g., '`a photo of a <category>`', generating C inputs $\{t_i\}_{i=1}^{C}$ for the text encoder $\mathcal{T}(\cdot)$. With $\mathcal{V}(I)$ denotes the visual representation for the image I, the text features of the text templates with class labels $\{1, 2, ..., C\}$ are matched using the formula $p(y|I) = \frac{\exp(sim(\mathcal{T}(t_y),\mathcal{V}(I)))}{\sum_{i=1}^{C}\exp(sim(\mathcal{T}(t_i),\mathcal{V}(I)))}$, where $y \in \{1, 2, ..., C\}$, $\mathcal{T}(t_y) \in \mathbb{R}^d$, $\mathcal{V}(I) \in \mathbb{R}^d$, d is the output dimension shared by both encoders, and $sim(\cdot, \cdot)$ refers to cosine similarity. For simplicity, we will use $\mathcal{T}(y)$ in place of $\mathcal{T}(t_y)$ in the following content.

Textual Prompt Tuning. To efficiently adapt the pre-trained VLM for downstream tasks, CoOp [41] avoids fine-tuning the entire network by only fine-tuning the textual template. Specifically, CoOp replaces the phrase '`a photo of a`' with M_t trainable vectors, i.e., '$[V]_1[V]_2...[V]_{M_t}$ `<category>`'. Each $[V]_{i\in\{1,2,...,M_t\}} \in \mathbb{R}^{d_t}$ is a

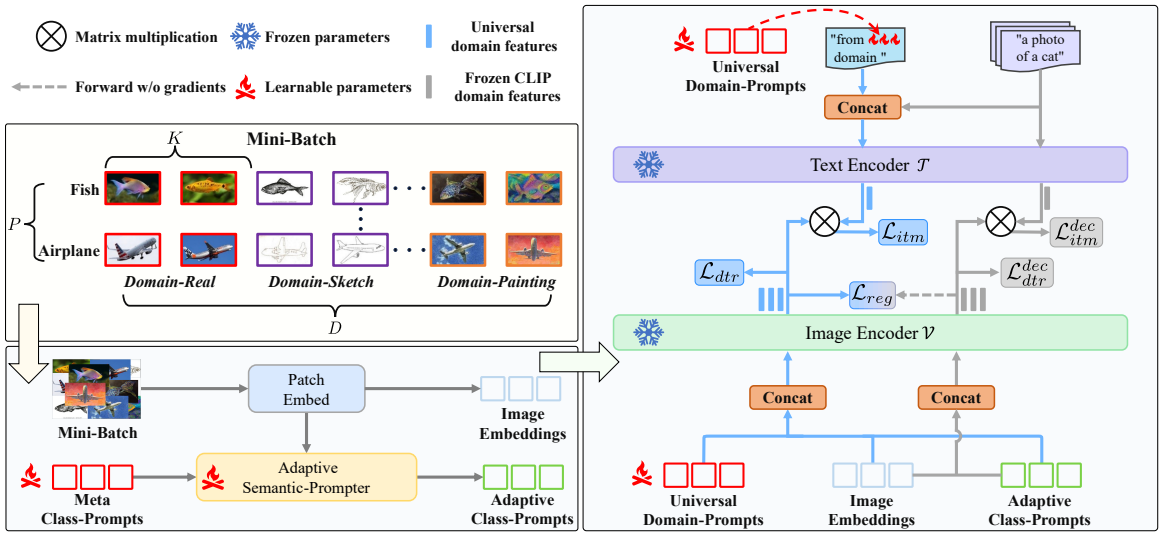


Figure 3: Overview of DePro: DePro decouples prompts into universal domain prompts and class prompts to address the domain and semantic shifts inherent in UCDR. To facilitate this decoupling, a decoupling loss, $\mathcal{L}_{itm}^{dec}$, is introduced, along with a regulation loss, $\mathcal{L}_{reg}$, to enable a domain ensemble effect. Additionally, based on the *DPK* domain sampler, the DaTri losses, $\mathcal{L}_{dtr}$ and $\mathcal{L}_{dtr}^{dec}$, are introduced to mitigate the risk of class collapse.

learnable vector with the same dimension d_t as the word embeddings. The text encoder $\mathcal{T}(\cdot)$, consisting of L Transformer layers [34] $\{\mathcal{T}_i\}_{i=1}^{L}$, performs a fine-tuning forward as:

$$(P_j, W_j|_y) = \mathcal{T}_j(P_{j-1}, W_{j-1}|_y), \quad j = 1, ..., L, \tag{1}$$

where $W_0|_y = [w_0^1|_y, w_0^2|_y, ..., w_0^{N_t}|_y]^T \in \mathbb{R}^{N_t \times d_t}$ represents the text embeddings conditioned on class y, with N_t indicates embedding length, and $P_0 = [V]_1[V]_2...[V]_{M_t}$. The output of the text encoder is then calculated as:

$$\mathcal{T}(y, P_0) = \texttt{TextProj}(w_L^{N_t}|_y), \tag{2}$$

where TextProj is a fully-connected layer ($d_t \rightarrow d$).

Visual Prompt Tuning. Given that the image encoder $\mathcal{V}(\cdot)$ also consists of L Transformer layers $\{\mathcal{V}_i\}_{i=1}^{L}$, VPT [18] introduces the application of prompt tuning to the visual backbone. Specifically, VPT introduces M_v learnable vectors $\hat{P}_0 = \{\hat{P}_0^i \in \mathbb{R}^{d_v}\}_{i=1}^{M_v}$, with d_v denotes the visual embedding dimension. Similarly, the fine-tuning forward is formulated as:

$$[x_j, \hat{P}_j, E_j|_I] = \mathcal{V}_j(x_{j-1}, \hat{P}_{j-1}, E_{j-1}|_I), \ j = 1, ..., L, \tag{3}$$

where $E_0|_I \in \mathbb{R}^{N_v \times d_v}$ is the patch embeddings of image I, and $x_0 \in \mathbb{R}^{d_v}$ is the initial CLS Token. The CLS Token x_L from the last Transformer layer is utilized to compute visual representation:

$$\mathcal{V}(I, \hat{P}_0) = \texttt{ImageProj}(x_L), \tag{4}$$

where ImageProj is also a fully-connected layer ($d_v \rightarrow d$).

3.3 Prompts Decoupling

Motivation. UCDR requires distinguishing images when classes or domains are unseen. This objective can be broken down into two key sub-goals: first, mapping image features from different domains, including unseen ones, into an optimal universal domain; meanwhile, discriminating image features within this domain. Accordingly, we propose the **D**omain **E**nsemble using Decoupled **Pro**mpts (DePro) framework (**Fig. 3**), which utilizes two types of prompts to achieve these goals: universal domain prompts and class prompts.

Universal Domain Prompts. In UCDR, domain-specific prompts, tailored to their respective domains, risk overfitting to the source domains, limiting their ability to generalize to unseen domains, even when jointly used during inference. To overcome this, we propose universal domain prompts (UDPs), learned from multiple source domains to capture shared characteristics and map instances into a more robust, universal space. We reuse $\hat{P}_0 \in \mathbb{R}^{M_v \times d_v}$ and $P_0 \in \mathbb{R}^{M_t \times d_t}$ as the UDPs for both encoders. Following ProS [8], the textual template is defined as "a photo of [CLASS] from P_0 domain."

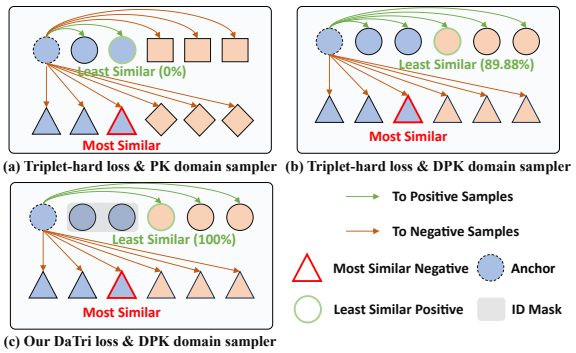
Class Prompts. Class prompts (CPs) play the role in guiding the extraction of class information. To handle unseen classes, we introduce an adaptive semantic-prompter (ASP), $\mathcal{F} = \{\mathcal{F}_j\}_{j=1}^{\ell}$, composed of ℓ Transformer layers, to dynamically generate adaptive CPs that are tailored to the unique characteristics of each input, thereby enhancing the model's ability to capture fine-grained class-specific details. Specifically, we define M_c meta class prompts $\bar{P}_0 = \{\bar{P}_0^i \in \mathbb{R}^{d_v}\}_{i=1}^{M_c}$ as the input for the ASP:

$$[\bar{P}_j, E_j|_I] = \mathcal{F}_j(\bar{P}_{j-1}, E_{j-1}|_I), \ j = 1, ..., \ell, \tag{5}$$

where $\bar{P}_\ell$ denotes the adaptive CPs. The image-text matching loss $\mathcal{L}_{itm}$ is then introduced to optimize both the UDPs and CPs:

$$g_{P_0}(y) = \mathcal{T}(y, P_0), \quad f_{\hat{P}_0}(I) = \mathcal{V}(I, \hat{P}_0, \bar{P}_\ell),$$
$$\mathcal{L}_{itm}(I) = -\sum_{y=1}^{C} p_y \log\left(\frac{\exp(sim(g_{P_0}(y), f_{\hat{P}_0}(I)))}{\sum_{\hat{y}=1}^{C} \exp(sim(g_{P_0}(\hat{y}), f_{\hat{P}_0}(I)))}\right), \tag{6}$$

where p_y denotes the ground truth for the image I.



(a) Triplet-hard loss & PK domain sampler (b) Triplet-hard loss & DPK domain sampler

(c) Our DaTri loss & DPK domain sampler

Figure 4: (a) For the triplet-hard loss with the *PK* domain sampler, the probability of the least similar positive sample originating from a different domain is 0%. (b) With the *DPK* domain sampler, this probability increases to 88.98%. (c) For our proposed DaTri loss, leveraging the *DPK* sampler with an in-domain (ID) mask, this probability is maximized to 100%.

Decoupling Loss. While we can simply adopt $\mathcal{L}_{itm}$ for optimization, our ablation experiments reveal that retrieval results with CPs are nearly identical to those without them. This is because the generated CPs inevitably incorporate domain-specific information, leading to overlapping roles or even conflicts with the UDPs. Therefore, we propose the decoupling loss, to ensure that CPs remain domain-agnostic. Specifically, we unplug the UDPs for both encoders and compute the decoupling loss $\mathcal{L}_{itm}^{dec}$ as follows:

$$g(y) = \mathcal{T}(y), \quad f(I) = \mathcal{V}(I, \tilde{P}_\ell),$$
$$\mathcal{L}_{itm}^{dec}(I) = -\sum_{y=1}^{C} p_y \log\left(\frac{\exp(sim(g(y), f(I)))}{\sum_{\hat{y}=1}^{C} \exp(sim(g(\hat{y}), f(I)))}\right). \tag{7}$$

Intuitively, $\mathcal{L}_{itm}^{dec}$ serves to align the inter-class distribution of CPs-guided visual outputs with that of the frozen CLIP textual outputs. Given that the inter-class distributions of the frozen CLIP visual and textual components are well-aligned during CLIP's pre-training, $\mathcal{L}_{itm}^{dec}$ not only ensures consistency in the inter-class distribution between the CPs-guided visual outputs and the frozen CLIP visual outputs—thereby upholding the domain-agnostic property of CPs—but also facilitates a reduction in intra-class distances.

Domain Ensemble. For the image I, $f_{\hat{P}_0}(I)$ represents the features mapped to the learned universal domain, while $f(I)$ corresponds to the frozen CLIP domain. In practical test scenarios, it is challenging to assert that the query domain closely aligns with either the universal domain or the frozen CLIP domain. A simple solution would be to conduct the concatenation between $f_{\hat{P}_0}(I)$ and $f(I)$, however, this operation requires to perform two forward passes during testing, which doubles the computational load. Therefore, we propose a surrogate loss named regulation loss $\mathcal{L}_{reg}$ to play a similar domain ensemble effect by regulating $f_{\hat{P}_0}(I)$ toward $f(I)$:

$$\mathcal{L}_{reg}(I) = \|f_{\hat{P}_0}(I) - f(I)\|_2. \tag{8}$$

Note that $\mathcal{L}_{reg}$ specifically targets $f_{\hat{P}_0}(I)$, while no gradients are propagated through the $f(I)$ branch. This design prevents any interference with the domain-agnostic properties of the CPs.

3.4 Domain-Aware Triplet-Hard Loss

In this section, we systematically examine the application of triplet-hard loss [16] within the context of UCDR, a loss function widely used in metric learning tasks [5, 23, 24, 30] to effectively minimize intra-class distances while maximizing inter-class distances. The standard approach involves employing a *PK* sampler to select P classes and sample K images per class, forming a mini-batch of size $B = P \times K$. The triplet-hard loss is then defined as follows:

$$\mathcal{L}_{tri\_hard} = \frac{1}{B}\sum_{i=1}^{P}\sum_{a=1}^{K}[\rho - \min_{p=1\ldots K} sim(I_i^a, I_i^p) + \max_{\substack{j=1\ldots P \\ j\neq i \\ n=1\ldots K}} sim(I_i^a, I_j^n)]_+, \tag{9}$$

where ρ is the margin hyper-parameter, $[\cdot]_+$ denotes the $max(0,\cdot)$ function, $sim(\cdot,\cdot)$ represents cosine similarity, and I_i^a identifies the anchor image, specifically the a-th image from the i-th class within the batch. Additionally, I_i^p refers to the positive sample, while I_j^n corresponds to the negative sample.

In UCDR, which involves multiple source domains, we explore two sampling variants: the *PK* domain sampler and the *DPK* domain sampler (**Fig. 2-c**). The *PK* domain sampler independently selects PK samples from each domain and computes the triplet-hard loss separately. In comparison, the *DPK* domain sampler ensures that all domains share the same set of classes, effectively functioning as a single $P\hat{K}$ sampler, where $\hat{K} = D \times K$, with $D = N_S$ denotes the number of source domains within the batch. *Our results indicate that the triplet-hard loss based on the DPK domain sampler performs significantly better.* This improvement can likely be attributed to the fact that when the *DPK* domain sampler is employed, the least similar positive samples to the anchor predominantly come from other domains (89.88% in our experiments). This provides two significant benefits: ❶ it reduces inter-domain discrepancies during optimization, promoting better generalization across domains; and ❷ it mitigates overfitting by preventing class collapse (i.e., samples from the same class are compressed into a single point in the embedding space, losing their internal similarity structure [4]), a risk particularly heightened when the least similar positive samples are confined to the same domain.

Building on this insight, we further introduce the domain-aware triplet-hard (DaTri) loss $\mathcal{L}_{dtr}$ (**Fig. 4**) as follows:

$$\mathcal{L}_{dtr} = \frac{1}{B}\sum_{d=1}^{D}\sum_{i=1}^{P}\sum_{a=1}^{K}[\rho - \min_{\substack{b=1\ldots D \\ b\neq d \\ p=1\ldots K}} sim(I_{d,i}^a, I_{b,i}^p) + \max_{\substack{b=1\ldots D \\ j=1\ldots P \\ j\neq i \\ n=1\ldots K}} sim(I_{d,i}^a, I_{b,j}^n)]_+, \tag{10}$$

where $I_{d,i}^a$ identifies the anchor image, specifically the a-th image from the i-th class and d-th domain within the batch. In DaTri, the proportion of least similar positive samples from other domains is maintained at 100% using an in-domain mask ($b \neq d$ in **Eq.** (10)), further enhancing the benefits of the *DPK* domain sampler.

3.5 Loss Function

Based on the *DPK* domain sampler, given a mini-batch of size $[B, d]$ as the outputs from the image encoder, where $B = D \times P \times K$, the

overall loss function for our framework is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{itm} + \mathcal{L}_{itm}^{dec} + \mathcal{L}_{dtr} + \mathcal{L}_{dtr}^{dec} + \mathcal{L}_{reg}, \tag{11}$$

where we also apply DaTri loss to the branch without UDPs ($\mathcal{L}_{dtr}^{dec}$).

4 Experiments

4.1 Datasets, Baselines, and Evaluation Metrics

Datasets. To fully evaluate the effectiveness of proposed DePro, we experiment on three benchmarks, including DomainNet [27], Sketchy [31], and TU-Berlin [7].

DomainNet dataset is utilized for UCDR and U^{D}CDR evaluations. It comprises 596,006 images across six domains: *Real*, *Sketch*, *Quickdraw*, *Infograph*, *Clipart*, and *Painting*, covering a total of 345 categories. The dataset is divided into three subsets: 245 categories for training, 55 categories for validation, and 45 categories for testing. In line with the leave-one-out setting employed in ProS [8], five domains are treated as source domains, while the remaining domain serves as the unseen query domain. For the UCDR evaluation, the `UnseenGallery` consists of images from the *Real* domain with unseen classes, and the `MixedGallery` is created by combining the `UnseenGallery` with 8% of samples from each seen class in the *Real* domain. For the U^{D}CDR evaluation, we select 45 training classes and utilize 25% of samples from each class for both the query domain (10% for *Quickdraw*) and the *Real* domain.

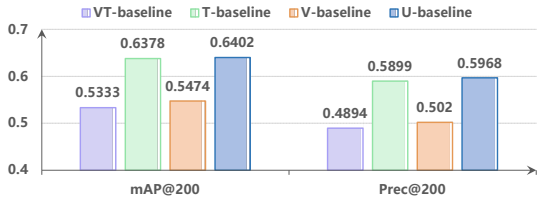
Sketchy and **TU-Berlin** datasets are utilized for the U^{C}CDR evaluation. The TU-Berlin dataset comprises 250 categories, each containing 80 free-hand sketches, along with 204,489 extended natural images as provided by [38]. Similarly, the Sketchy dataset includes 125 categories, totaling 75,471 hand-drawn sketches and 12,500 corresponding images. We use its extended version [25] having an additional 60,502 natural images sourced from from ImageNet[29]. We partition the categories in the TU-Berlin and Sketchy datasets into training (200 / 93 categories), validation (20 / 11 categories), and testing (30 / 21 categories) sets.

Baselines. We have established four baselines to evaluate the effect of domain-specific prompts: ❶ *VT-baseline*, where domain-specific prompts are applied to both encoders; ❷ *V-baseline* / ❸ *T-baseline*, where domain-specific prompts are applied only to the visual or textual encoder, while the other encoder uses universal domain prompts; and ❹ *U-baseline*, where both encoders utilize universal domain prompts. During training, we employ the sampler from ProS (see **Fig. 2-c**) and utilize only the image-text matching loss.

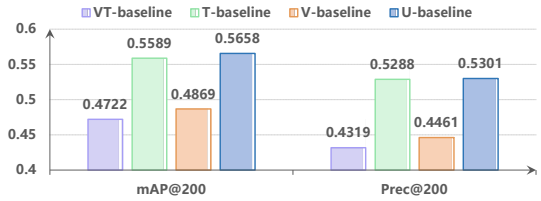
Evaluation Metrics. Following prior works [8, 32], we adopt consistent evaluation metrics. For Sketchy and DomainNet, we calculate precision (Prec@200) and mean Average Precision (mAP@200) using the top-200 retrieved candidates. For TU-Berlin, the metrics used are Prec@100 and mAP@all.

4.2 Implementation Details

We perform prompt tuning on a pre-trained CLIP model with a ViT-B/32 backbone [6], where the model dimensions are set as $d = d_t = 512$ for language branch and $d_v = 768$ for vision branch. The textual UDPs have a length of 1 ($M_t = 1$), while CPs and visual UDPs and are set to a length of 4 ($M_c = M_v = 4$). The adaptive semantic-prompter (ASP) includes 2 Transformer layers ($\ell = 2$). The margin hyper-parameter ρ is set to 0.5 in **Eq.** (10). Training is



(a) Unseen Gallery



(b) Mixed Gallery

Figure 5: Baselines comparisons on DomainNet under UCDR.

conducted over 10 epochs with early stopping after 2 epochs. For the batch size, we use $P = 3$ and $K = 4$ for DomainNet (5 source domains), resulting in a batch size of 60 samples. For Sketchy and TU-Berlin datasets, which involve 2 source domains, we set $P = 6$ and $K = 4$, leading to a batch size of 48 samples. The Adam optimizer is employed with a learning rate of 1e-3, accompanied by a cosine learning rate decay strategy. During training, we fine-tune all the layer normalization parameters as a trick, referred to as the LN-trick. All experiments are performed on an NVIDIA RTX-4090 GPU (24GB), with the random seed fixed at 0 for reproducibility.

4.3 Comparison with the State-of-the-Arts

We compare our model with state-of-the-art (SOTA) methods, which fall into two categories: ***traditional methods***, including SnMpNet [26], SCNNet[1] and SASA [32], and ***CLIP-based methods***, such as VPT [18], ProS [8],and CLIP-F [28]. Notably, CLIP-F refers to the fine-tuning of the entire CLIP model.

Comparison under UCDR and U^{D}CDR. We compare the UCDR and U^{D}CDR performance of our DePro against existing methods on DomainNet, as shown in **Tab. 1**, yielding several key insights: ❶ *CLIP-based methods outperform tradition methods.* Unlike traditional approaches relying on single-modality backbones (e.g., ResNet [15] or ViT [6]), CLIP leverages multi-modality pre-training, enabling even frozen backbone to significantly outperform fine-tuned single-modality methods. ❷ *Our method consistently outperforms existing SOTA methods.* For instance, under the UCDR *Sketch*-query scenario with the `UnseenGallery`, DePro achieves a mAP@200 improvement of 4.79% over ProS. This highlights DePro's effectiveness in addressing the dual challenges of unseen domains and classes in UCDR. Moreover, under U^{D}CDR, where only domain shift is present, DePro achieves a mAP@200 gain of 4.15% over ProS under the *Sketch*-query scenario.

Comparison under U^{C}CDR. As shown in **Tab. 2**, DePro also outperforms all other methods. On the Sketchy dataset, DePro outperforms ProS by 4.79% and 5.90% on mAP@200 and Prec@200, respectively. On the TU-Berlin dataset, DePro outperforms ProS by

Table 1: Comparison with state-of-the-art (SOTA) methods under UCDR and U^{D}CDR. The symbol ⋆ denotes the traditional methods, and † denotes the CLIP-based methods. The best performance is marked as bold and the second best performance is marked as underline, while scores from our DePro are highlighted with a light purple background.

Training Domains	Query Domain	Method	Venue	UCDR				U^{D}CDR	
				Unseen Gallery		Mixed Gallery			
				mAP@200	Prec@200	mAP@200	Prec@200	mAP@200	Prec@200
Real, *Quickdraw*, *Infograph*, *Painting*, *Clipart*	*Sketch*	SnMpNet⋆	ICCV'21	0.3007	0.2432	0.2624	0.2134	0.3529	0.1657
		SASA⋆	SIGIR'22	0.5262	0.4468	0.4732	0.4025	0.5733	0.5290
		SCNNet⋆	ACCV'23	0.4075	0.4120	0.3422	0.2534	-	-
		CLIP†	ICML'21	0.4220	0.3528	0.3662	0.2979	0.4760	0.2871
		CLIP-F†	ICML'21	0.5367	0.4666	0.4788	0.4136	0.6128	0.3806
		VPT†	ECCV'22	0.6216	0.5676	0.5609	0.5130	0.6769	0.4405
		ProS†	CVPR'24	0.6457	0.6001	0.5843	0.5463	0.7385	0.4911
		DePro†	**Ours**	**0.6936**	**0.6521**	**0.6209**	**0.5830**	**0.7881**	**0.5240**
Real, *Sketch*, *Infograph*, *Painting*, *Clipart*	*Quickdraw*	SnMpNet⋆	ICCV'21	0.1736	0.1284	0.1512	0.1111	0.1077	0.0509
		SASA⋆	SIGIR'22	0.2564	0.1970	0.2116	0.1651	0.1805	0.1549
		SCNNet⋆	ACCV'23	0.1998	0.1580	0.1698	0.1411	-	-
		CLIP†	ICML'21	0.0744	0.0561	0.0600	0.0317	0.0867	0.0450
		CLIP-F†	ICML'21	0.2011	0.1522	0.1622	0.1196	0.1820	0.0723
		VPT†	ECCV'22	0.2467	0.2092	0.1953	0.1688	0.2367	0.0982
		ProS†	CVPR'24	0.2842	0.2544	0.2318	0.2127	0.2889	0.1186
		DePro†	**Ours**	**0.3165**	**0.2949**	**0.2432**	**0.2331**	**0.3393**	**0.1333**
Real, *Sketch*, *Infograph*, *Quickdraw*, *Clipart*	*Painting*	SnMpNet⋆	ICCV'21	0.4031	0.3332	0.3635	0.3019	0.4808	0.4424
		SASA⋆	SIGIR'22	0.5898	0.5188	0.5463	0.4804	0.5596	0.5178
		SCNNet⋆	ACCV'23	0.4242	0.4409	0.3731	0.3964	-	-
		CLIP†	ICML'21	0.6168	0.5507	0.5653	0.5014	0.5569	0.3170
		CLIP-F†	ICML'21	0.6558	0.5926	0.6083	0.5478	0.6189	0.3688
		VPT†	ECCV'22	0.7138	0.6503	0.6752	0.6153	0.6618	0.4105
		ProS†	CVPR'24	0.7516	0.6955	0.7120	0.6612	0.7227	0.4615
		DePro†	**Ours**	**0.7699**	**0.7188**	**0.7178**	**0.6728**	**0.7691**	**0.4867**
Real, *Sketch*, *Painting*, *Quickdraw*, *Clipart*	*Infograph*	SnMpNet⋆	ICCV'21	0.2079	0.1717	0.1800	0.1496	0.1957	0.1764
		SASA⋆	SIGIR'22	0.2823	0.2425	0.2491	0.2113	0.2340	0.2093
		SCNNet⋆	ACCV'23	0.2737	0.2476	0.2369	0.1983	-	-
		CLIP†	ICML'21	0.5008	0.4474	0.4375	0.3891	0.4756	0.2936
		CLIP-F†	ICML'21	0.5332	0.4893	0.4718	0.4309	0.5311	0.3330
		VPT†	ECCV'22	0.5434	0.4957	0.4870	0.4468	0.5690	0.3566
		ProS†	CVPR'24	0.5798	0.5442	0.5219	0.4956	0.6056	0.3962
		DePro†	**Ours**	**0.6213**	**0.5946**	**0.5445**	**0.5228**	**0.6711**	**0.4361**
Real, *Sketch*, *Painting*, *Quickdraw*, *Infograph*	*Clipart*	SnMpNet⋆	ICCV'21	0.4198	0.3323	0.3765	0.2959	0.5520	0.5074
		SASA⋆	SIGIR'22	0.4397	0.3670	0.3940	0.3295	0.6840	0.6361
		SCNNet⋆	ACCV'23	0.3579	0.3449	0.3108	0.2981	-	-
		CLIP†	ICML'21	0.6037	0.5130	0.5608	0.4691	0.5581	0.3110
		CLIP-F†	ICML'21	0.6880	0.6200	0.6423	0.5755	0.6922	0.4174
		VPT†	ECCV'22	0.7344	0.6785	0.6942	0.6409	0.7536	0.4770
		ProS†	CVPR'24	0.7648	0.7186	0.7228	0.6815	0.8105	0.5298
		DePro†	**Ours**	**0.7920**	**0.7510**	**0.7389**	**0.7026**	**0.8432**	**0.5504**

3.39% and 1.28% on mAP@200 and Prec@200, respectively, indicating that the semantic shift problem is well alleviated.

Comparison between baselines. We compare the performance of four baselines on DomainNet under UCDR, with *Sketch* as the query

Table 2: Comparison with SOTA methods under U^{C}CDR.

Method	Sketchy		TU-Berlin	
	mAP@200	Prec@200	mAP@all	Prec@100
SnMpNet*	0.5781	0.5155	0.3568	0.5226
SASA*	0.6910	0.6090	0.4715	0.6682
CLIP†	0.3582	0.3308	0.3145	0.4612
CLIP-F†	0.6553	0.6145	0.6076	0.7158
VPT†	0.6588	0.6105	0.5574	0.6815
ProS†	0.6991	0.6545	0.6675	0.7442
DePro†	**0.7470**	**0.7135**	**0.7014**	**0.7570**

Table 3: Ablation studies on prompt decoupling and LN-trick.

Method	Unseen Gallery		Mixed Gallery	
	mAP@200	Prec@200	mAP@200	Prec@200
ProS	0.6457	0.6001	0.5843	0.5463
U-baseline	0.6402	0.5968	0.5658	0.5301
+ASP	0.6408	0.5983	0.5684	0.5328
+$\mathcal{L}_{itm}^{dec}$	0.6582	0.6182	0.5814	0.5503
+$\mathcal{L}_{reg}$	0.6634	0.6250	0.5865	0.5543
+ LN-trick	0.6740	0.6351	0.6002	0.5655

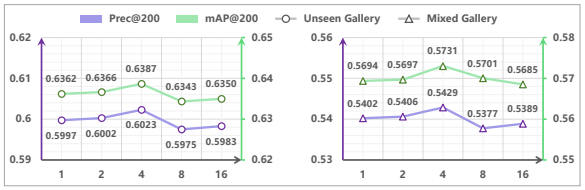
Table 4: Ablation studies on samplers and DaTri loss. The '-' denotes that we do not utilize either the triplet-hard loss or the DaTri loss. The sampler used, as illustrated in Fig. 2-c, is enclosed in square brackets (i.e., []).

Method	Unseen Gallery		Mixed Gallery	
	mAP@200	Prec@200	mAP@200	Prec@200
- [simple]	0.6740	0.6351	0.6002	0.5655
- [PK]	0.6733	0.6341	0.5958	0.5645
- [DPK]	0.6699	0.6309	0.5972	0.5604
Tri [PK]	0.6816	0.6400	0.6021	0.5695
Tri [DPK]	0.6874	0.6465	0.6137	0.5775
DaTri [DPK]	0.6936	0.6521	0.6209	0.5830

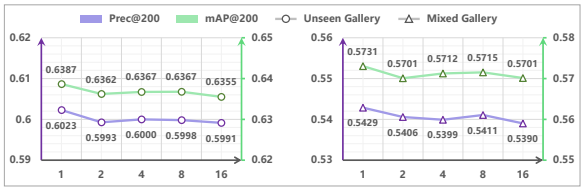
domain, as shown in **Fig. 5**. Key conclusions are: ❶*U-baseline delivers the best performance.* Domain-specific prompts, whether applied to the text or image encoder, limit the model's ability to generalize to unseen domains and classes. ❷*V-baseline underperforms compared to T-baseline.* Visual Domain-specific prompts risk overfitting to source domains, limiting their generalization to unseen domains, even when used jointly during inference. This issue is further exacerbated when combined with domain-specific text prompts, which can create domain-specific prototypes (VT-baseline).

4.4 Ablation Study

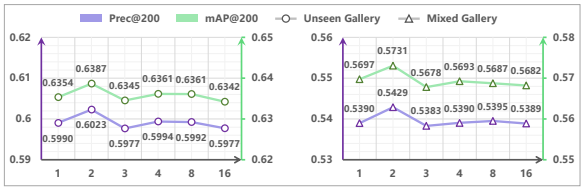
In this section, we assess the performance of each component within our proposed framework. By default, the experiments are conducted under UCDR, with *Sketch* as the query domain.



(a) Class prompt length



(b) Textual universal domain prompt length



(c) ASP depth

Figure 6: Evaluation results of class prompt length (a), universal domain prompt length (b) and ASP depth (c) on DomainNet under UCDR. We conduct the comparison by averaging the performance across five query domains.

Effectiveness of Prompt Decoupling and LN-trick. We evaluate the effectiveness of our proposed prompt decoupling strategy and LN-trick using U-baseline as the benchmark. The results, presented in **Tab. 3**, reveal several key insights: ❶ *The benefit of ASP is limited.* The ASP-generated CPs retain excessive domain-specific information, leading to functional overlap with UDPs and thereby restricting their overall effectiveness. ❷ *The decoupling loss ($\mathcal{L}_{itm}^{dec}$) yields substantial benefits.* For instance, with the `UnseenGallery`, mAP@200 improves by 1.74%, underscoring the critical role of effective prompt decoupling. ❸ *Domain ensemble ($\mathcal{L}_{reg}$) offers additional gains.* With the `UnseenGallery`, it provides a 0.52% improvement in mAP@200 , demonstrating its utility in handling unseen domains. ❹ *The LN-trick is essential.* Consistent with prior work [2, 9, 30], fine-tuning normalization parameters proves highly effective for adapting to unseen domains and classes.

Effectiveness of DaTri Loss. Based on the LN-trick results in **Tab. 3**, we independently incorporate the triplet-hard loss and DaTri loss using various samplers (illustrated in **Fig. 2-c**), including the simple domain sampler (simple), the PK domain sampler (PK), and the DPK domain sampler (DPK). The results, summarized in **Tab. 4**, lead to the following observations: ❶ When neither triplet-hard loss nor DaTri loss is applied, the simple domain sampler achieves the best performance due to its high degree of randomness. ❷ The DPK domain sampler outperforms the PK domain sampler when paired with the triplet-hard loss, as it ensures that the least similar positive sample for an anchor is *always* selected from a

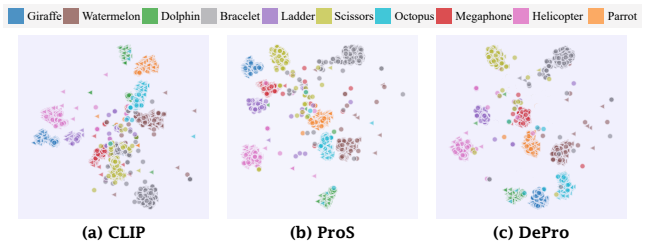


Figure 7: The t-SNE visualization for 10 randomly selected unseen classes of *Clipart* (query) domains and *Real* (gallery) domain. Different colors represent different categories, while △ and ○ represent samples from *Real* and *Clipart* domains, respectively.

different domain. ❸ Combining our DaTri loss with the DPK domain sampler further boosts performance, enlarging the effective synergy between the triplet-hard loss and the DPK domain sampler.

The Impact of Prompt Length. We also conduct experiments to evaluate the impact of prompt lengths in DePro, specifically the lengths of the class prompts (M_c) and the textual universal domain prompts (M_t). As shown in **Fig. 6-a**, increasing the class prompt length initially improves performance, peaking at a length of 4, after which it gradually declines. Furthermore, **Fig. 6-b** demonstrates that the optimal length for the textual universal domain prompt is 1, with overall performance decreasing as the length increases. Based on these findings, we set $M_c = 4$ and $M_t = 1$ for all experiments.

The Impact of ASP Depth. To investigate the impact of ASP depth, we modify the number of Transformer layers in ASP (ℓ) and evaluate the resulting performance. The results, shown in **Fig. 6-c**, demonstrate that 2 layers offer the optimal configuration. This setting achieves the best trade-off between efficiency and accuracy, balancing computational efficiency with performance.

4.5 Visualization

In this section, by default, the experiments are conducted under UCDR with the `UnseenGallery`, using *Clipart* as the query domain.

T-SNE Visualization. As shown in **Fig. 7**, we visualize the image features extracted from frozen CLIP, ProS, and our DePro method for 10 randomly selected unseen classes of the *Real* and unseen *Clipart* domains using t-SNE [33]. We can observe that our DePro is able to better align the feature representations with the same classes between the two domains. Moreover, the features extracted by our DePro exhibit better separation between different classes.

Top-10 Visualization. As shown in **Fig. 8**, we compare the number of correct images retrieved among the top-10 images for the total of 6394 queries between DePro and ProS. DePro consistently outperforms ProS, achieving higher accuracy across more queries, highlighting its superior performance. **Fig. 9-a** further illustrates selected retrieval results for both models. Notably, existing models struggle when using the *Quickdraw* or *Infograph* as query domains. **Fig. 9-b** showcases some failure cases, attributing the challenges to *query ambiguity*, which arises from: 1) insufficient detail within a query, leading to vague inputs, and 2) too many objects within a query, causing confusion over which object to retrieve.

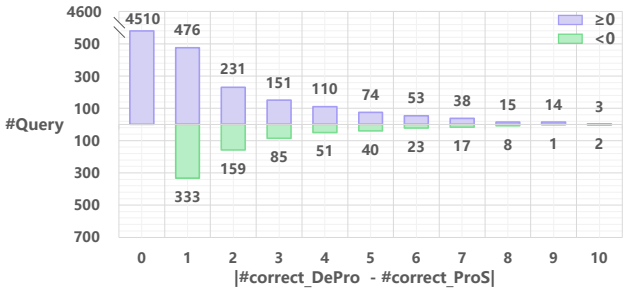


Figure 8: The distribution of differences in the number of correctly retrieved in top-10 images between DePro and ProS.

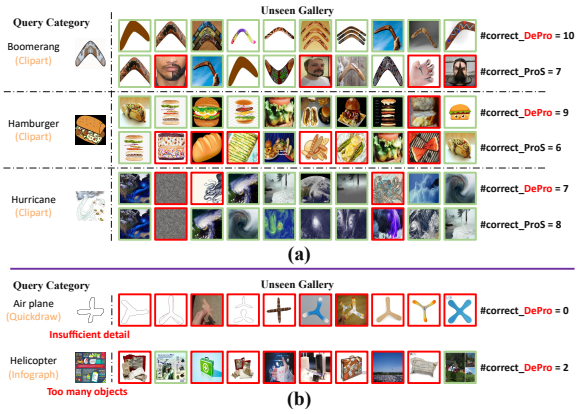


Figure 9: (a) Comparison of retrieval results between DePro and ProS. (b) The query ambiguity problem when using *Quickdraw* or *Infograph* as the query domain.

5 Conclusion

This paper presents DePro, a framework designed to address the universal cross-domain retrieval (UCDR) problem through a novel prompt decoupling strategy. By separating prompts into universal domain prompts (UDPs) and class prompts (CPs), DePro effectively manages both domain and semantic shifts inherent in UCDR. The incorporation of decoupling and regulation losses ensures domain-agnostic CPs and coherent feature integration. Additionally, the domain-aware triplet-hard (DaTri) loss further enhances retrieval accuracy by reducing the risk of class collapse and optimizing the synergy between the triplet-hard loss and the DPK domain sampler. Extensive experiments validate DePro's superior generalization across unseen domains and classes.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62476056 and 62306070) and the Social Development Science and Technology Project of Jiangsu Province (No. BE2022811). This work was also supported in part by the Southeast University Start-Up Grant for New Faculty under Grant 4009002309. Furthermore, the work was also supported by the Big Data Computing Center of Southeast University and the SEU Innovation Capability Enhancement Plan for Doctoral Student.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, Balaji Vasan Srinivasan, and Biplab Banerjee. 2023. Contrastive Learning of Semantic Concepts for Open-set Cross-domain Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4115–4124.

[2] Samyadeep Basu, Shell Hu, Daniela Massiceti, and Soheil Feizi. 2024. Strong Baselines for Parameter-Efficient Few-Shot Fine-Tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 11024–11031.

[3] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 35–44.

[4] Yongjuan Che, Yuexuan An, and Hui Xue. 2023. Boosting Few-Shot Open-Set Recognition with Multi-Relation Margin Loss. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. 3505–3513.

[5] Kaixiang Chen, Pengfei Fang, Zi Ye, and Liyan Zhang. 2024. Multi-Scale Explicit Matching and Mutual Subject Teacher Learning for Generalizable Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 9 (2024), 8881–8895.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[7] Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Transactions on graphics (TOG)* 31, 4 (2012), 1–10.

[8] Kaipeng Fang, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Zhi-Qi Cheng, Xiyao Li, and Heng Tao Shen. 2024. Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17292–17301.

[9] Jonathan Frankle, David J Schwab, and Ari S Morcos. 2020. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *arXiv preprint arXiv:2003.00152* (2020).

[10] Bojana Gajic and Ramon Baldrich. 2018. Cross-domain fashion image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1869–1871.

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* 132, 2 (2024), 581–595.

[12] Ce Ge, Jingyu Wang, Qi Qi, Haifeng Sun, Tong Xu, and Jianxin Liao. 2023. Scene-level sketch-based image retrieval with minimal pairwise supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 650–657.

[13] Andrew Graves and Mounia Lalmas. 2002. Video retrieval using an MPEG-7 based inference network. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 339–346.

[14] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. 2023. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 746–754.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[17] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. 2015. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*. 1062–1070.

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.

[19] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19113–19122.

[20] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15190–15200.

[21] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2024. How to Handle Sketch-Abstraction in Sketch-Based Image Retrieval?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16859–16869.

[22] Siyuan Li, Li Sun, and Qingli Li. 2023. CLIP-ReID: Exploiting Vision-Language Model for Image Re-identification without Concrete Text Labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1405–1413.

[23] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. 2021. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2898–2907.

[24] Hai Liu, Shuai Fang, Zhaoli Zhang, Duantengchuan Li, Ke Lin, and Jiazhang Wang. 2021. MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation. *IEEE Transactions on Multimedia* 24 (2021), 2449–2460.

[25] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2862–2871.

[26] Soumava Paul, Titir Dutta, and Soma Biswas. 2021. Universal cross-domain retrieval: Generalizing across classes and domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12056–12064.

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1406–1415.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[30] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. 2023. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2765–2775.

[31] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–12.

[32] Jialin Tian, Xing Xu, Kai Wang, Zuo Cao, Xunliang Cai, and Heng Tao Shen. 2022. Structure-aware semantic-aligned network for universal cross-domain retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 278–289.

[33] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[35] Junsheng Wang, Tiantian Gong, and Yan Yan. 2024. Semi-supervised Prototype Semantic Association Learning for Robust Cross-modal Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 872–881.

[36] Qiang Wang, Junlong Du, Ke Yan, and Shouhong Ding. 2023. Seeing in flowing: Adapting clip for action recognition with motion prompts learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5339–5347.

[37] Hantao Yao, Rui Zhang, and Changsheng Xu. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6757–6767.

[38] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. 2016. Sketchnet: Sketch classification with web images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1105–1113.

[39] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. 2023. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15211–15222.

[40] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*. Springer, 493–510.

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.