ProS: Prompting-to-simulate Generalized knowledge for Universal Cross-Domain Retrieval

Kaipeng Fang[1] Jingkuan Song[2]* Lianli Gao[2] Pengpeng Zeng[2] Zhi-Qi Cheng[3]
Xiyao Li[4] Heng Tao Shen[5]

[1] School of Computer Science and Enginnering, University of Electronic Science and Technology of China
[2] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
[3] Carnegie Mellon University [4] Kuaishou Technology [5] Tongji University

fangkaipeng@126.com, jingkuan.song@gmail.com

Abstract

The goal of Universal Cross-Domain Retrieval (UCDR) is to achieve robust performance in generalized test scenarios, wherein data may belong to strictly unknown domains and categories during training. Recently, pre-trained models with prompt tuning have shown strong generalization capabilities and attained noteworthy achievements in various downstream tasks, such as few-shot learning and video-text retrieval. However, applying them directly to UCDR may not be sufficient to handle both domain shift (i.e., adapting to unfamiliar domains) and semantic shift (i.e., transferring to unknown categories). To this end, we propose ***Pro****mpting-to-****S****imulate (ProS), the first method to apply prompt tuning for UCDR. ProS employs a two-step process to simulate Content-aware Dynamic Prompts (CaDP) which can impact models to produce generalized features for UCDR. Concretely, in Prompt Units Learning stage, we introduce two Prompt Units to individually capture domain and semantic knowledge in a mask-and-align way. Then, in Context-aware Simulator Learning stage, we train a Content-aware Prompt Simulator under a simulated test scenario to produce the corresponding CaDP. Extensive experiments conducted on three benchmark datasets show that our method achieves new state-of-the-art performance without bringing excessive parameters. Code is available at https://github.com/fangkaipeng/ProS.*

1. Introduction

Cross-Domain Retrieval (CDR) [11, 51] is an important task in Information Retrieval (IR) systems that utilizes data from one domain (*e.g.* Infograph, Sketch) as a query to search semantically similar examples in another domain (*e.g.* natural image). Despite achieving promising results, existing methods [2, 4, 11, 38] heavily follow a close-set learning setting, where domains and categories in both training and testing data are pre-defined. Such evaluation limits the model to the training samples and unable

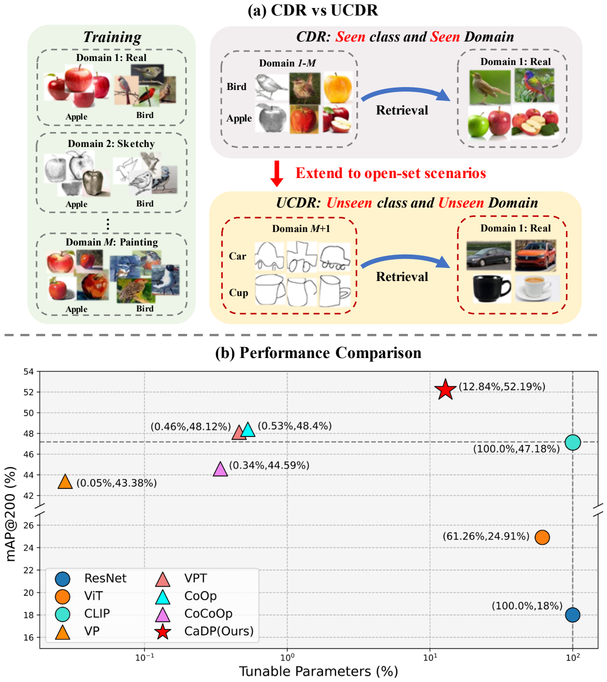


Figure 1. **(a)** Illustration of Cross-Domain Retrieval (CDR) and its generalized version (UCDR). **(b)** Comparison of our ProS ☆ with different backbones ○ and various prompt-based methods △ under UCDR protocol. All prompt-based methods use CLIP as the backbone. Our method yields solid improvement and achieves a better trade-off between performance and trainable parameters usage against state-of-the art.

*Corresponding author

to tackle the open-set applications such as e-commerce search [6, 10, 27] and recommendation [10, 19], where a searched product not only comes from various new domains (*e.g*. styles) but is also accompanied by the emergence of new categories. Towards this goal, one emerging research direction is proposed, namely Universal Cross-Domain Retrieval (UCDR) [32], shown in Fig. 1(a), for performing well under generalized test scenarios, where test data may belong to strictly unknown domains and categories.

The core challenge for UCDR is how to empower models with the capabilities of both domain shift (*i.e*. adapting to unfamiliar domains) and semantic shift (*i.e*. transferring to unknown categories). The mainstream approaches [32, 41] focus on fine-tuning visual models, *e.g*. ResNet [14] and ViT [8], with the guidance of expert knowledge from semantic information or pre-trained models. Compared with ResNet, ViT has stronger representational capability and contains more knowledge gained from larger pre-training data, which facilitates the generalization of the model for unknown images. This raises an intuitive question: *can the extensive general knowledge inherent in large-scale models enhance its ability to tackle the UCDR task where prior information about test samples is strictly unknown?*

With this question in mind, we empirically conduct a confirmatory experiment on the widely-recognized large-scale pre-trained model, CLIP [35] under UCDR. Specifically, we fully fine-tune the CLIP and compare its performance with ViT and ResNet, as shown in Fig. 1 (b). From the figure, we can see that CLIP yields tremendous gains, reaching a 22.27% improvement in mAP@200 compared with ViT. This result proves the effectiveness of common knowledge inherent in large-scale CLIP for UCDR task.

However, full fine-tuning strategy inevitably forgets the useful knowledge gained in the pre-training phase [20]. Besides, since the entire model is updated, this strategy brings significant computational costs. To alleviate above problems, we take the first step to leverage prompt tuning [17, 49] which is a more flexible and lightweight strategy than full fine-tuning. However, when directly applying existing prompt tuning methods [1, 17, 49, 50] to UCDR, *i.e*. △ in Fig. 1 (b), we find that these methods show a fine benefit in terms of computational cost without a strong advantage or even worse over full fine-tuning CLIP in terms of mAP@200. This may be attributed that these methods do not fulfill the properties of UCDR, contributing to poor generalization. Therefore, we pose a question: *how to design a more effective prompt tuning method for UCDR task?*

Towards above question, we propose a novel prompt tuning method in a simulated way, named **Pro**mpting-to-**S**imulate (ProS), to effectively leverage prompt tuning to mine the generalized knowledge from CLIP for UCDR. Specifically, ProS introduces a two-stage learning to simulate two content-aware dynamic prompts (CaDP) which reflect the input sample's domain and category, respectively. In Prompt Unit Learning (PUL), two groups of distinct learnable prompts are established: domain prompt units and semantic prompt units. These two units are employed to extract domain knowledge and semantic knowledge from source data in a mask-and-align strategy. Then, in Context-aware Simulator learning (CSL), we train a Content-aware Prompt Simulator (CaPS) under simulated test scenarios to dynamically generate content-aware dynamic prompts (CaDP) based on the learned prompt units. With the CaDP, we can impact CLIP's knowledge to gain a more generalizable representation and achieve better performance compared with existing prompt tuning methods in UCDR, as shown in Fig. 1 (b). In addition, our method uses considerably fewer learnable parameters compared to full fine-tuning, accounting for only 12.84% of the model parameters, and does not introduce excessive learnable parameters compared to prompt tuning methods.

Our main contributions can be summed as follows:

- We are the first to investigate how to adapt CLIP with prompts for UCDR.
- We propose a prompt-based method named Prompting-to-Simulate (ProS), which learns the generalized knowledge to deal with open-set scenarios.
- Extensive experiments on three benchmark datasets show that our ProS achieves new state-of-the-art results compared with prompt-based methods without bringing excessive parameters.

2. Related Work

2.1. Universal Cross-Domain Retrieval

UCDR requires solving domain and category generalization simultaneously which can be roughly regarded as a combination of Domain Generalization (DG) [25, 42] and Zero-Shot Learning (ZSL) [13, 43]. In DG, some approaches synthesize unseen samples using data augmentation strategy [31] or generative adversarial networks (GANs) [48] to handle the domain shift. Another technique [33] utilizes innovative category-level center and distribution alignments to address out-of-domain generalization. In ZSL, some researchers leverage the auxiliary data, including human-annotated attribute information [21, 22], text description[36] or knowledge graph [23] to learn the relationship between seen categories and pre-define unseen categories names. It should be clarified that UCDR does not require pre-defined unseen class names. Inspired by above approaches, SnMpNet [32] trains the model with the mix-up data augmentation strategy and align features with semantic information. 3DPC-GZSL [44] enhances generalization to previously unseen classes using an efficient visual-semantic synthesis. Additionally, SASA[41] deploys a ViT-based model and trains categorical prototypes to handle the semantic shift. Different from these works, we make the first attempt to apply a large-scale pre-trained model, *i.e.* CLIP, to the UCDR task, and propose a new prompt tuning method to improve the generalization ability of the CLIP.

2.2. Vision-Language Pre-training Models

Motivated by the success of self-supervised learning, Vision-Language Pre-training has emerged as a prominent topic. CLIP [35] initially demonstrated that equipped with large-scale image-text pairs, a contrastive learning framework can achieve a performance comparable to fully supervised baselines. Subsequently, ALIGN [16] scales the training dataset to billions and gains a better vision-language representation. Another line of work [5, 12] leverage pre-trained object detection models like Faster-RCNN [37] to extract image regional features offline for training multi-modal transformers. Recently, 3D-VLP [18] has innovatively generalized vision-language pre-training to 3D domain, achieving superior performances across various tasks. These models show impressive performance on various vision-language tasks, as they are capable of processing both single-modal and multi-modal inputs. In this paper, we make the first attempt to leverage the powerful CLIP for the UCDR task.

2.3. Prompt Tuning

Prompt Tuning [29, 47], as a new paradigm, initially surfaces in NLP for adapting pre-trained language models (PLMs) [3, 7] to downstream tasks, such as few-shot learning [30, 46] and video-text retrieval [26, 45, 52]. These prompts can either be handcrafted for specific tasks [40] or learned automatically via gradients which are termed as "Prompt Learning" [24]. Driven by the success of prompt paradigm in NLP, various studies like CoOP [49] and CoCoOp [50] apply text prompting in multi-modal scenarios, while VPT [17] and VP [1] introduce this paradigm into vision transformer. Despite the pioneering successes of VP and VPT in visual models, we find that they are static after training which is not flexible to handle the generalized scenarios. Our method proposes a Prompt-to-Simulate method to well address UCDR task.

3. Method

3.1. Preliminary

Problem Formulation. Universal Cross-Domain Retrieval (UCDR) [32] aims to accurately retrieve images with queries from unseen domains and classes. From a general perspective, we assume that the training dataset is composed of K-different ($K \geq 2$) source domains (Real, Sketch, Quickdraw, *etc.*) which can be formulated as $\mathcal{D}_{train} = \bigcup_{d \in \{1,\ldots,K\}} \{\boldsymbol{x}_i^d, y_i^{tr}\}_{i=1}^{N_d}$. The d^{th} domain has N_d images, where each image $\boldsymbol{x}_i^d$ belongs to class $y_i^{tr} \in \mathcal{C}_{train}$. In the testing phase, the gallery set and query set are needed to enable image retrieval. Specifically, the query set $\mathcal{D}_{query} = \{\boldsymbol{x}_i^q, y_i^{te}\}_{i=1}^{N_q}$ consists of N_q samples belonging to unseen classes $y_i^{\text{te}} \in \mathcal{C}_{\text{test}}$ from the query domain q, while the gallery set consists of real images. To simulate a more realistic retrieval scenario, the gallery set has two settings: 1) all samples belong to the unseen class, termed *Unseen Gallery*; (2) samples belong to both seen class and unseen class, termed *Mixed Gallery*.

The UCDR protocol can be separated into two sub-tasks, namely: (a) U^{c}CDR, where the query domain $q \in \{1, ..., K\}$ but the classes are held out, *i.e.* $\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} \equiv \phi$. (b) U^{d}CDR, where the query domain $q \notin \{1, ..., K\}$ but the classes are seen, *i.e.* $\mathcal{C}_{\text{train}} \equiv \mathcal{C}_{\text{test}}$. The goal of UCDR is to learn a mapping function that can place query and gallery images into a latent domain-independent subspace Φ , where images of the same label but different domains are clustered together. The most challenging part is images from unseen domains and unknown classes also need to be mapped correctly in Φ.

CLIP. Contrastive Language-Image Pre-training (CLIP) [35] is a multi-modality model pre-trained from 400 million image-text pairs. It aligns texts and images by an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. CLIP classifies images in a zero-shot way, based on the similarity between image features $f(\boldsymbol{x})$ and text features. The text features of various class-wise caption $\{\boldsymbol{t}_i\}_{i=1}^{|C_{train}|}$ can be formulated as $\{g(\boldsymbol{t}_i)\}_{i=1}^{|C_{train}|}$, which are produced manually by filling class names into a text template. For example, given a text template "a photo of a [CLASS].", then exchange [CLASS] with class names to construct class-wise caption, *e.g.*, [CLASS] → cat / dog / bike. Formally speaking, given an image $\boldsymbol{x}$ and class caption $\{\boldsymbol{t}_i\}_{i=1}^{|C_{train}|}$, CLIP output a prediction by:

$$\hat{y}_{clip} = \arg\max_i (f(\boldsymbol{x}) \otimes g(\boldsymbol{t}_i)), \tag{1}$$

where $\otimes$ is cosine similarity. Benefiting from large-scale pre-training and alignment of images and text, CLIP is robust for various visual appearance changes [15], which is suitable for UCDR. However, fine-tuning CLIP will distort pre-trained features and lead to poor out-of-distribution performance [20]. Therefore, how to effectively leverage this powerful foundation model for UCDR becomes an important problem.

Prompt Tuning. Compared with full fine-tuning, prompt tuning prevents pre-trained feature distortion by introducing task-specific trainable parameters termed prompts into the input while keeping the pre-trained model frozen. Specifically, [49] incorporate learnable text prompts $\mathbf{P_t} = \{p_t^i \in \mathbb{R}^\ell\}_{i=1}^{N_t}$ into the text template as the input of CLIP text encoder, where N_t and $\mathbb{R}^\ell$ indicate text prompts length and dimensions, respectively. Then the text template can be represented as "$\mathbf{P_t}$ a photo of a [CLASS]". Furthermore,

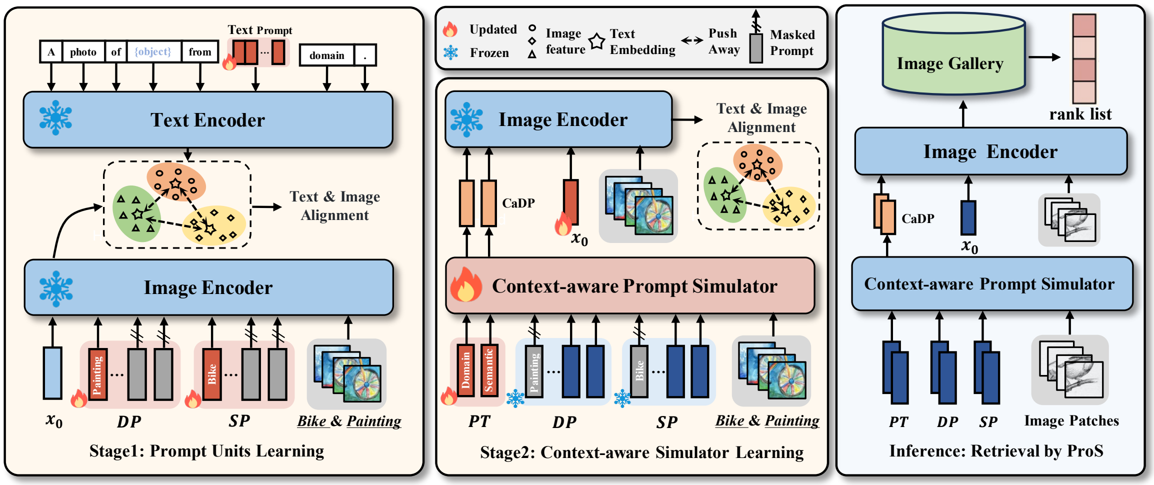


Figure 2. **Overview of our proposed ProS**. In Prompt Units Learning Stage, we capture knowledge from source data into domain prompts units DP and semantic prompts units SP by masking irrelevance prompts. In the Context-aware Prompt Simulation Stage, we train a Context-aware Prompt Simulator (CaPS) with a mask operation to dynamically convey prompt templates PT to two Content-aware Dynamic Prompts (CaDP) to simulate unknown domains and categories. In the retrieval phase, we employ CaPS to produce CaDP which impacts the CLIP image encoder to convert unseen samples into suitable embeddings for retrieval. The gray parts indicate masked prompts.

[17] proposed visual prompts for ViT which can be formulated as $\mathbf{P_v} = \left\{p_v^i \in \mathbb{R}^\ell\right\}_{i=1}^{N_v}$, where N_v represents the number of visual prompts and p_v is a learnable parameter with ℓ dimension. Then the visual prompts are inserted into the first Transformer layer, named VPT-Shallow in [17]:

$$\begin{aligned} [\mathbf{x}_1, \mathbf{Z}_1, \mathbf{E}_1] &= L_1\left([\mathbf{x}_0, \mathbf{P_v}, \mathbf{E}_0]\right), \\ [\mathbf{x}_i, \mathbf{Z}_i, \mathbf{E}_i] &= L_i\left([\mathbf{x}_{i-1}, \mathbf{Z}_{i-1}, \mathbf{E}_{i-1}]\right) \quad i = 2, 3, \ldots, N_L, \\ \textbf{Output} &= \text{Head}\left(\mathbf{x}_{N_L}\right), \end{aligned} \tag{2}$$

where $\mathbf{Z}_i \in \mathbb{R}^{N_v \times \ell}$ represents the prompt embeddings computed by the i^{th} Transformer layer in the CLIP image encoder and $\mathbf{E}_0$ means the image patch embeddings extracted from a CNN feature map [8]. In addition, $\mathbf{x_0}$ is *CLS* token, which produces $\mathbf{y}$ and serves as the final output of image representation.

3.2. Prompt-to-Simulate

Prompt Units Learning (PUL). To capture domain-specific and class-specific knowledge in PUL, we introduce two types of prompt units: domain prompt units DP and semantic prompt units SP, as shown in Fig. 2 (left). Specifically, we employ K learnable vectors as domain prompt units to represent K source domains in training datasets:

$$DP = \left\{dp_i \in \mathbb{R}^\ell\right\}_{i=1}^{K}. \tag{3}$$

Similarly, we introduce a set of learnable semantic prompts for $|\mathcal{C}_{\text{train}}|$ seen classes, which can be formulated as:

$$SP = \left\{sp_i \in \mathbb{R}^\ell\right\}_{i=1}^{|\mathcal{C}_{\text{train}}|}. \tag{4}$$

To ensure that each prompt only learns knowledge related to a specific domain or class, we iteratively update single dp and sp w.r.t. an input image. To achieve this, we mask the **i**rrelevant **d**omain and **s**emantic prompt units by $\mathbf{M_{id}} = \left\{\mathbf{m_{id}}^i \in [0,1]\right\}_{i=1}^{K}$ and $\mathbf{M_{is}} = \left\{\mathbf{m_{is}}^i \in [0,1]\right\}_{i=1}^{|\mathcal{C}_{\text{train}}|}$. Assume that, the input image is from domain $\mathbf{d}$ and class $\mathbf{c}$, than $\mathbf{m_{id}^d}$ and $\mathbf{m_{is}^c}$ are 1 and other entries are 0. After mask operation, two masked prompt units are then concatenated with other input tokens to formulate a prompted input:

$$\mathbf{x_p} = [\mathbf{x_0}, \mathbf{M_{id}} \odot DP, \mathbf{M_{is}} \odot SP, \mathbf{E_0}], \tag{5}$$

where $\odot$ discards prompts with corresponding mask of 0 preserves others of 1. Correspondingly, the image feature is obtained in the same way as Eq. (2). Since the text feature can be considered as the class prototype, we introduce the learnable text prompts $\mathbf{P_t}$ as we discussed in Sec. 3.1 to achieve more flexible training. Additionally, we design a new text template: "a photo of $[CLASS]^i$ from $\mathbf{P_t}$ domain.", to better suit the UCDR task. Given the prompted input $\mathbf{x_p}$ and the class caption $\{\boldsymbol{t}_i\}_{i=1}^{|C_{train}|}$ generated by above text template, the training loss can be formulated as:

$$\mathcal{L} = -\sum_{i=1}^{|\mathcal{C}_{\text{train}}|} y_i \log\left(f\left(\mathbf{x_p}\right) \otimes g\left(\boldsymbol{t}_i\right)\right), \tag{6}$$

where $\otimes$ is cosine similarity and y indicates the class label. The above procedure results in a mask-and-align objective. Our masked input ensures prompts to capture specific source knowledge from domains and categories. Then, the image features and text features are aligned among different domains in a union semantic space.

Context-aware Simulator Learning (CSL). In CSL, our goal is to train a module termed Content-aware Prompt Simulator (CaPS) to generate Content-aware Dynamic Prompts (CaDP) similar to Prompt Units which can impact CLIP to extract generalized features of unseen samples. Towards this goal, we concatenate two sets of Prompt Units with image patches as CaPS's input, which can be formulated as $[PT_d, PT_s, DP, SP, \mathbf{E_0}]$, where $PT_d \in \mathbb{R}^{\ell}$, $PT_s \in \mathbb{R}^{\ell}$ are domain and semantic prompt templates for CaDP generation and $\mathbf{E_0}$ are unseen sample's embedding.

Inspired by the idea of meta-learning, we train CaPS under a simulated test scenario by a relevant mask operation. Specifically, we mask **r**elevant **d**omain and **s**emantic prompt units w.r.t. input samples by $\mathbf{M_{rd}} = \left\{\mathbf{m_{rd}}^i \in [0,1]\right\}_{i=1}^{K}$ and $\mathbf{M_{rs}} = \left\{\mathbf{m_{rs}}^i \in [0,1]\right\}_{i=1}^{|\mathcal{C}_{\text{train}}|}$, where $\mathbf{m_{rd}^{d}} = 0$, $\mathbf{m_{rs}^{c}} = 0$ and others are 1 if the input sample is from domain $\mathbf{d}$ and class $\mathbf{c}$. Then we can hide the prompt units corresponding to input samples by multiplying them with $\mathbf{M_{rd}}$ and $\mathbf{M_{rs}}$ mask matrices. The whole process can be formulated as:

$$\begin{aligned} [\mathcal{P}_d, \mathcal{P}_s] &= \mathcal{M}([PT_d, PT_s, \mathbf{M_{rd}} \odot DP, \mathbf{M_{rs}} \odot SP, \mathbf{E_0}]), \\ \mathbf{x_p} &= [\mathbf{x_0}, \mathcal{P}_d, \mathcal{P}_s, \mathbf{E_0}], \end{aligned} \tag{7}$$

where $\mathcal{P}_d$ and $\mathcal{P}_s$ represent CaDP for domain and semantic which is converted by PT_d and PT_s respectively.

Similarly, we use the same objective as Eq. (6) to train the newly introduced PT_d, PT_s and $\mathcal{M}$ while the remaining part would be fixed in this stage.

3.3. Retrieval by ProS

After two-stage training, now we can generate CaDP by CaPS, and extract image features by CLIP with the help of CaPS. Fig. 2 (right) shows how to employ our proposed ProS for retrieval. Since we only need image features for retrieval, CLIP text encoder g is discarded. We first use CaPS $\mathcal{M}$ to generate two CaDP, and then use CLIP Image Encoder f to get well-generalized image features $\mathbf{y}$:

$$\begin{aligned} [\mathcal{P}_d, \mathcal{P}_s] &= \mathcal{M}([PT_d, PT_s, DP, SP, \mathbf{E_0}]), \\ \mathbf{x} &= f\left([\mathbf{x_0}, \mathcal{P}_d, \mathcal{P}_s, \mathbf{E}_0]\right), \\ \mathbf{y} &= \text{Head}\left(\mathbf{x}\right). \end{aligned} \tag{8}$$

To perform retrieval, all images in $\mathcal{C}_{\text{test}}$ are indexed by Eq. (8) to formulate the gallery. Then, given a query, we do the same thing to obtain the query feature and produce a rank list by sorting similarities between the query and all features in the gallery. In the next section, we will give a comprehensive study of the retrieval results of our ProS.

4. Experiment

4.1. Experimental setting

Datasets. Following the existing works [32, 41], we evaluate the effectiveness of our ProS on three popular datasets, including DomainNet [34], Sketchy [28, 39], and TU-Berlin [9, 28]) under three different cross-domain retrieval settings, *i.e.*, UCDR, U^{c}CDR and U^{d}CDR as we introduced in Sec. 3.1.

DomainNet [34] contains 596,006 images collected in 6 domains (*Real, Sketch, Quickdraw, Infograph, Clipart, Painting*) from 345 categories and split into 245, 55, 45 categories for training, validation, and testing respectively. To satisfy the unseen domain requirement in both U$^{\text{d}}$CDR and UCDR protocol, *leave-one-out* evaluation protocol is applied, where we iteratively select one domain as unseen query set and use other domains for training. Additionally, for U$^{\text{d}}$CDR evaluation, we select 45 categories from training set and choose 25% samples from unseen domains as queries (except for Quickdraw, which is 10% due to its large size). The gallery set is constructed with *Real* images from *unseen* categories or mixed with *seen* categories, termed *Unseen Gallery* and *Mixed Gallery* respectively.

Sketchy [28, 39] contains 75,471 Sketches and 73,002 images from 125 categories where the train, validation, and test splits contain 93, 11, and 21 categories [32, 41], respectively. **TU-Berlin** [9, 28] has 20,000 Sketches and 204,489 images which are separated into 200 categories for training, 20 categories for validation, and 30 categories for testing [32, 41]. Both Sketchy and TU-Berlin are used for U$^{\text{c}}$CDR.

Evaluation Metrics. For a fair comparison, we utilize the same evaluation metrics following [32, 41]. For Sketchy and DomainNet, we evaluate precision and mean Average Precision w.r.t. top-200 candidates (Prec@200 and mAP@200). For TU-Berlin, we use Prec@100 and mAP@all as evaluation metrics instead.

Implementation Details. In all our experiments, we use a fixed image encoder and text encoder initialized with pre-trained CLIP (ViT-B/32). For Context-aware Prompt Simulator $\mathcal{M}$, we adopt a two-layer ViT with random initialization. Following [49], the length of text prompt $\mathbf{P_t}$ is set to 16 with 512 dimensions. In addition, the dimensions of other prompts including prompt units and Content-aware Dynamic Prompts are set to 768. The training epochs are set to 10 with early stopping of 2 epochs based on the evaluation performance for all the datasets with a batch size of 50. We use Adam optimizer with a learning rate of 1e-3 and adopt a cosine decay strategy for the learning rate.

4.2. Main Results

Baselines. We compare ProS with two groups of methods: two existing UCDR methods, *i.e.*, SnMpNet [32] which

Query Domain	Method	UCDR				U^dCDR	
		Unseen Gallery		Mixed Gallery			
		mAP@200	Prec@200	mAP@200	Prec@200	mAP@200	Prec@200
Sketch	SnMpNet [32]	0.3007	0.2432	0.2624	0.2134	0.3529	0.1657
	SASA [41]	0.5262	0.4468	0.4732	0.4025	0.5733	0.5290
	CLIP-Full	0.5367	0.4666	0.4788	0.4136	0.6128	0.3806
	CoOp [49]	0.5512	0.4947	0.4995	0.4479	0.6374	0.4245
	VPT [17]	0.6216	0.5676	0.5609	0.5130	0.6769	0.4405
	ProS (Ours)	**0.6457**	**0.6001**	**0.5843**	**0.5463**	**0.7385**	**0.4911**
Quickdraw	SnMpNet [32]	0.1736	0.1284	0.1512	0.1111	0.1077	0.0509
	SASA [41]	0.2564	0.1970	0.2116	0.1651	0.1805	0.1549
	CLIP-Full	0.2011	0.1522	0.1622	0.1196	0.1820	0.0723
	CoOp [49]	0.1484	0.1237	0.1183	0.0961	0.1834	0.084
	VPT [17]	0.2467	0.2092	0.1953	0.1688	0.2367	0.0982
	ProS (Ours)	**0.2842**	**0.2544**	**0.2318**	**0.2127**	**0.2889**	**0.1186**
Painting	SnMpNet [32]	0.4031	0.3332	0.3635	0.3019	0.4808	0.4424
	SASA [41]	0.5898	0.5188	0.5463	0.4804	0.5596	0.5178
	CLIP-Full	0.6558	0.5926	0.6083	0.5478	0.6189	0.3688
	CoOp [49]	0.6886	0.6207	0.6509	0.5884	0.6625	0.4128
	VPT [17]	0.7138	0.6503	0.6752	0.6153	0.6618	0.4105
	ProS (Ours)	**0.7516**	**0.6955**	**0.7120**	**0.6612**	**0.7227**	**0.4615**
Infograph	SnMpNet [32]	0.2079	0.1717	0.1800	0.1496	0.1957	0.1764
	SASA [41]	0.2823	0.2425	0.2491	0.2113	0.2340	0.2093
	CLIP-Full	0.5332	0.4893	0.4718	0.4309	0.5311	0.3330
	CoOp [49]	0.5285	0.4807	0.4820	0.4390	0.5530	0.3546
	VPT [17]	0.5434	0.4957	0.4870	0.4468	0.5690	0.3566
	ProS (Ours)	**0.5798**	**0.5442**	**0.5219**	**0.4956**	**0.6056**	**0.3962**
Clipart	SnMpNet [32]	0.4198	0.3323	0.3765	0.2959	0.552	0.5074
	SASA [41]	0.4397	0.3670	0.3940	0.3295	0.684	0.6361
	CLIP-Full	0.6880	0.6200	0.6423	0.5755	0.6922	0.4174
	CoOp [49]	0.7025	0.6414	0.6648	0.6068	0.7495	0.4776
	VPT [17]	0.7344	0.6785	0.6942	0.6409	0.7536	0.4770
	ProS (Ours)	**0.7648**	**0.7186**	**0.7228**	**0.6815**	**0.8105**	**0.5298**
Average	SnMpNet [32]	0.3010	0.2418	0.2667	0.2144	0.3378	0.2686
	SASA [41]	0.4189	0.3544	0.3748	0.3178	0.4463	0.4094
	CLIP-Full	0.5229	0.4641	0.4727	0.4175	0.5274	0.3144
	CoOp [49]	0.5238	0.4722	0.4831	0.4356	0.5572	0.3507
	VPT [17]	0.5720	0.5203	0.5225	0.4770	0.5796	0.3566
	ProS (Ours)	**0.6052**	**0.5626**	**0.5546**	**0.5195**	**0.6332**	**0.3994**

Table 1. UCDR and U^dCDR evaluation results on DomainNet. UCDR has two different gallery settings, *i.e*. the gallery set consists of (1) only unseen class images (Unseen Gallery) or (2) both seen and unseen images from Real domain (Mixed Gallery).

adopt ResNet as backbone and SASA [41], which implement with ViT. For fair evaluation, we further build three CLIP-based baselines, including CLIP-full and two prompt tuning methods, *i.e*., CoOp [49] and VPT [17] where CLIP-full indicates fine-tuning full CLIP.

Results on UCDR and U^dCDR. We first evaluate our ProS methods on DomainNet under UCDR and U^dCDR. The UCDR results in Tab. 1 show that our ProS method outperforms all baselines across all domains, demonstrating the superiority of ProS in terms of generalization on unseen domains and categories. Furthermore, we have highlighted the following conclusions: **First** (CLIP *vs*. ViT): all CLIP-based methods outperform ViT-based method which proves our assumption that general knowledge gained in large-scale pre-training can enhance the model's ability to handle the UCDR task. Specifically, CLIP-Full achieves 9.79% mAP@200 improvement than SASA. **Second** (Prompt Tuning *vs*. Fine-Tuning): VPT and CoOp both outperform CLIP-Full, demonstrating the effectiveness of prompt tuning to adapt CLIP. **Third** (ProS *vs*. Prompt Tuning): compared with VPT, our ProS further improves the performance by 3.21% mAP@200 on average, indicating that ProS effectively applies prompt tuning for UCDR.

From reported **U^dCDR** results in Tab. 1, we can draw the conclusion that our ProS consistently outperforms all competitors in terms of cross-domain alignment and generalization. Specifically, comparing with average results, our method surpasses VPT by 5.36% in mAP@200.

Results on U^cCDR. In Tab. 2, we explore the ability of ProS in handling semantic shift under U^cCDR setting on Sketchy and TU-Berlin. We observe that: **First**, our ProS shows consistent improvement compared to CLIP-

Method	Sketchy		TU-Berlin	
	mAP@200	Prec@200	mAP@all	Prec@100
SnMpNet [32]	0.5781	0.5155	0.3568	0.5226
SASA [41]	0.6910	0.6090	0.4715	0.6682
CLIP-Full	0.6553	0.6145	0.6076	0.7158
CoOp [49]	0.5074	0.4659	0.5585	0.6759
VPT [17]	0.6588	0.6105	0.5574	0.6815
ProS (Ours)	**0.6991**	**0.6545**	**0.6675**	**0.7442**

Table 2. $\mathrm{U^cCDR}$ evaluation results on Sketchy and TU-Berlin. Consistent with [32, 41], we use Prec@200 and mAP@200 in Sketchy, and Prec@100 and mAP@all in TU-Berlin as evaluation metrics.

Full and VPT, revealing that our ProS can enhance the ability of CLIP to manage semantic shifts and generalize to unseen categories. **Second,** compared to SASA, our ProS has a larger improvement on TU-Berlin than that on Sketchy(19.6% *vs.* 0.81% in mAP@200). We could see CLIP, CoOp, and VPT, all of which use CLIP as the backbone, exhibit relatively low mAP on Sketchy, potentially caused by poor performance of the visual backbone when handling this dataset. However, our method still surpasses all competitors and achieves the best performance.

4.3. Ablation Study

In this section, we conduct extensive ablation experiments to reveal the contributions of each component. For all ablations, we select Infograph as an unseen query set from DomainNet and evaluate ProS under UCDR setting with two gallery configurations.

Impact of Each Components. To validate the effectiveness of each component in our ProS, we individually remove each component and experiment with the following five settings: 1) "w/o SP" (remove Semantic Prompt Units). 2) "w/o" DP (remove Domain Prompt Units). 3) "w/o Mask" (remove mask operation in training). 4) "w/o $\mathcal{M}$" (remove Content-aware Prompt Simulator). 5) "w/o CLS" (remove CLS token). The experiment results are shown in Tab. 3. We observe that: **First,** removing either SP or DP led to a performance degradation of -2.65% and -2.56% in mAP@200 under mixed gallery settings, indicating their essential roles in Content-aware Prompt Simulation stage. **Second,** without the mask operation, mAP@200 has decreased by 1.27% which proves the effectiveness of mask operation. **Third,** when we remove CaPS $\mathcal{M}$ from our model, it leads to worse performance than full ProS, highlighting the significant benefits of $\mathcal{M}$ for the UCDR task. **Forth,** training CLS token is absolutely necessary as its omission caused a performance reduction of 6.18

Impact of Transformer Layer in $\mathcal{M}$. We further investigate the impact of Transformer layers in CaPS $\mathcal{M}$. From the Tab. 4, it can be found that: 2 layer ViT yields the best performance. Although employing only 1 layer can slightly reduce computational costs, it results in a worse performance. Hence, for an optimal trade-off between accuracy and efficiency, a 2-layer ViT is chosen to implement $\mathcal{M}$.

Method	Unseen Gallery		Mixed Gallery	
	mAP@200	Prec200	mAP200	Prec200
ProS	**0.5798**	**0.5442**	**0.5219**	**0.4956**
w/o SP	0.5529	0.5142	0.4954	0.4641
w/o DP	0.5566	0.5190	0.4963	0.4680
w/o Mask	0.5692	0.5324	0.5092	0.4809
w/o $\mathcal{M}$	0.5573	0.5128	0.5053	0.4685
w/o CLS	0.5241	0.4846	0.4601	0.4260

Table 3. Different components evaluated on DomainNet under UCDR protocol with Infograph as the unseen domain for queries.

$\mathcal{M}$'s Layer	Unseen Gallery		Mixed Gallery	
	mAP@200	Prec@200	mAP@200	Prec@200
1	0.5752	0.5379	0.5177	0.4900
2	0.5798	0.5442	0.5219	0.4956
3	0.5790	0.5429	0.5210	0.4942
6	0.5743	0.5388	0.5160	0.4893
12	0.5598	0.5228	0.5036	0.4742

Table 4. Ablation study on the effectiveness of different transformer layers in Context-aware Prompt Simulator $\mathcal{M}$.

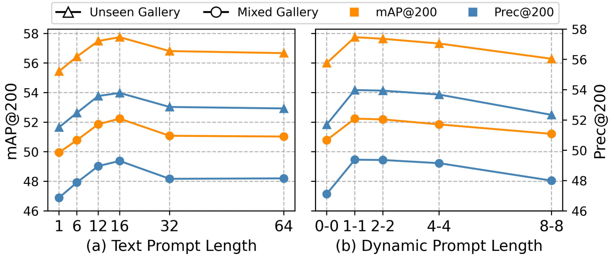


Figure 3. Evaluation results of two prompts length. (a) investigate the impact of text prompt length. (b) analyze Content-aware Dynamic Prompt length generated by CaPS $\mathcal{M}$, where 0-0 represents VPT and 1-1 means one CaDP for domain and one for semantic.

Impact of two prompts length. We further conduct experiments to analyze the impact of two key prompt lengths in ProS. *i.e.*, text prompts and Content-aware Dynamic Prompt (CaDP). Fig. 3(a) shows that when text prompts length increases, the performance improves to its peak when the length is 16 and then gradually decreases. Therefore, we set the text prompt length to 16 for all datasets. In Fig. 3(b), we find that ProS achieves the state-of-the-art with only 1-1 CaDP. We speculate that this might be consistent with Prompt Units, which also use a single prompt to represent one domain or category.

4.4. Qualitative Analysis

We visualize the feature space construct by CLIP-Full, VPT and ProS for 10 randomly selected unseen categories from *Infograph* Query and *Real* Gallery sets, as shown in

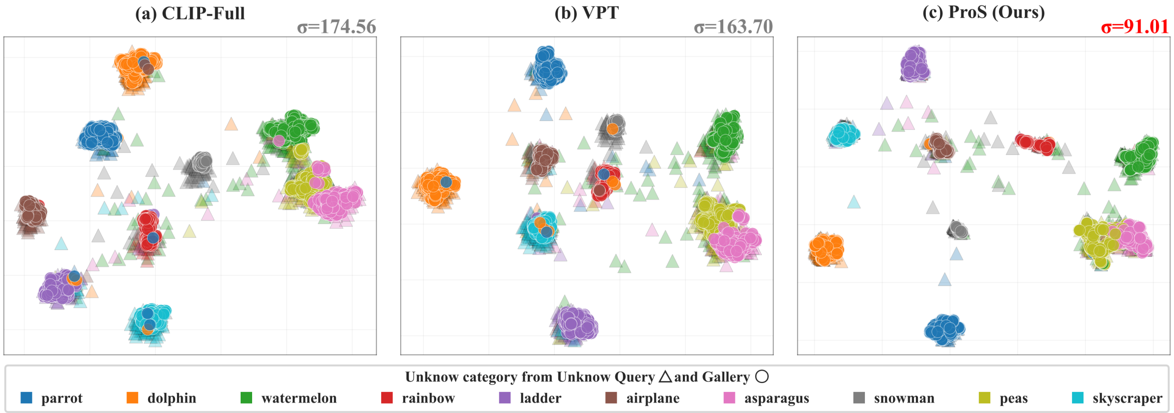
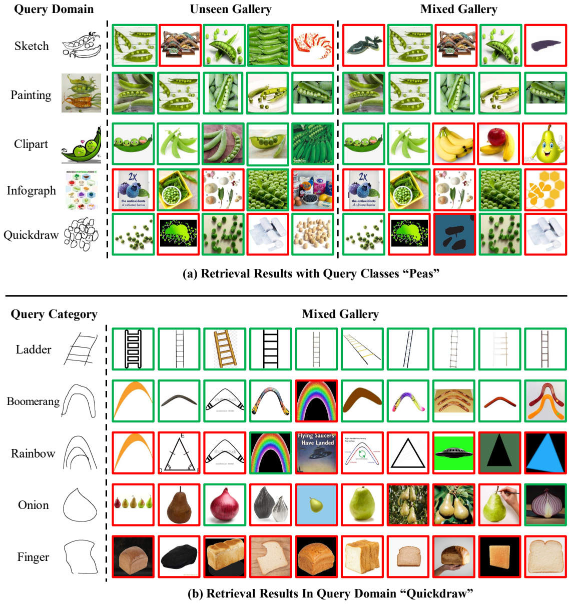


Figure 4. Visualization of image features from 10 randomly selected unseen classes of *Real* Query and unseen *Infograph* Gallery set. Different colors represent different categories while ◯ and △ represent samples from *real* and *Infograph* domains, respectively. We further evaluate performance by metric from [53], *i.e.*, $\sigma = \frac{\max \mathcal{D}_{intra}}{\min \mathcal{D}_{inter}}$ (lower is better).



(a) Retrieval Results with Query Classes "Peas"

(b) Retrieval Results In Query Domain "Quickdraw"

Figure 5. Retrieval Results under UCDR protocols on DomainNet. **(a)** displays the retrieval results of "Peas" by the query from an unseen query domain. **(b)** shows the retrieval results of a few queries from Quickdraw. True positives and false positives are shown with green and red borders, respectively.

Fig. 4. To further compare the feature space constructed by three methods, we evaluate the inter-class distinctiveness and intra-class compactness of feature space by the metric $\sigma = \frac{\max \mathcal{D}_{intra}}{\min \mathcal{D}_{inter}}$ from [53] where $\mathcal{D}_{intra}$ means intra-class distance and $\mathcal{D}_{inter}$ means inter-class distance. We can observe that: **First,** the features obtained by fine-tuned CLIP are not sufficiently separated and are not effectively aligned gallery and query domain, while VPT can partially alleviate this issue. **Second,** our ProS can extract more clustered features of the same classes compared with both fine-tuned CLIP and VPT, demonstrating the superiority of our approach. **Third,** comparing the extracted features using metrics from [53], we draw the consistent conclusion that CLIP $<$ VPT $\ll$ ProS.

To further demonstrate the effectiveness of our ProS, we visualize a few retrieval results produced by ProS. Fig. 5 (a) shows the top five retrieved candidates of peas with query images from different domains, while Fig. 5 (b) shows the top ten retrieved candidates with queries from Quickdraw. From the figure, we can see that Sketch and Quickdraw images lack enough information such as color, texture, and details while Inforgraph has an object co-occurrence problem. They all obtain worse performance compared with other domains. This phenomenon will be exacerbated when the gallery contains both seen and unseen classes due to the interference caused by seen classes. Towards this phenomenon, we can conclude that: due to the quality issues of the dataset, the performance of our APL in certain domains may be relatively poor.

5. Conclusion

In this paper, to the best of our knowledge, we take the first attempt to deploy CLIP for Universal Cross-Domain Retrieval (UCDR) with a prompt tuning paradigm and propose a more generalized prompt method named Prompt-to-Simulate (ProS). Specifically, ProS can dynamically fit unknown domain and category distribution with guidance of source knowledge by a two-stage training paradigm following a mask-and-align objective. With the above approach, our model obtains quite strong performance under UCDR compared to the state-of-the-art.

Acknowledgement

This study is supported by grants from the National Natural Science Foundation of China (Grant No. 62122018, No. 62020106008, No. U22A2097, No. U23A20315), Kuaishou, and SongShan Laboratory YYJC012022019.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2, 3

[2] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, pages 4247–4256, 2021. 1

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NIPS*, 2020. 3

[4] Abhra Chaudhuri, Ayan Kumar Bhunia, Yi-Zhe Song, and Anjan Dutta. Data-free sketch-based image retrieval. In *CVPR*, pages 12084–12093, 2023. 1

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *CoRR*, abs/1909.11740, 2019. 3

[6] Ayush Chopra, Abhishek Sinha, Hiresh Gupta, Mausoom Sarkar, Kumar Ayush, and Balaji Krishnamurthy. Powering robust fashion retrieval with information rich feature embeddings. In *CVPR*, pages 326–334, 2019. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 3

[8] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 4

[9] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Comput. Graph.*, 34(5):482–498, 2010. 5

[10] Anibal Fuentes and Jose M. Saavedra. Sketch-qnet: A quadruplet convnet for color sketch-based image retrieval. In *CVPR*, pages 2134–2141, 2021. 2

[11] Bojana Gajic and Ramón Baldrich. Cross-domain fashion image retrieval. In *CVPR*, pages 1869–1871, 2018. 1

[12] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 3

[13] Jiannan Ge, Hongtao Xie, Shaobo Min, Pandeng Li, and Yongdong Zhang. Dual part discovery network for zero-shot learning. In *ACM MM*, pages 3244–3252, 2022. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[15] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *ACL*, pages 2744–2751, 2020. 3

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 3

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 2, 3, 4, 6, 7

[18] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3D-language pre-training. In *CVPR*, pages 10984–10994, 2023. 3

[19] Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, and Julian J. McAuley. Complete the look: Scene-based complementary product recommendation. In *CVPR*, pages 10532–10541, 2019. 2

[20] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022. 2, 3

[21] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36 (3):453–465, 2014. 2

[22] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, pages 646–651, 2008. 2

[23] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *CVPR*, pages 1576–1585, 2018. 2

[24] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, 2021. 3

[25] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. In *ICCV*, pages 1446–1455, 2019. 2

[26] Hao Li, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Haonan Zhang, and Gongfu Li. A differentiable semantic metric approximation in probabilistic embedding for cross-modal retrieval. *NeurIPS*, 35:11934–11946, 2022. 3

[27] Haoyuan Li, Hao Jiang, Tao Jin, Mengyan Li, Yan Chen, Zhijie Lin, Yang Zhao, and Zhou Zhao. DATE: domain adaptive product seeker for e-commerce. In *CVPR*, pages 19315–19324, 2023. 2

[28] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2298–2307, 2017. 5

[29] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55:195:1–195:35, 2023. 3

[30] Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. In *ICML*, pages 23103–23123, 2023. 3

[31] Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *ECCV*, pages 466–483, 2020. 2

[32] Soumava Paul, Titir Dutta, and Soma Biswas. Universal cross-domain retrieval: Generalizing across classes and domains. In *ICCV*, pages 12036–12044, 2021. 2, 3, 5, 6, 7

[33] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *CVPR*, pages 2594–2605, 2022. 2

[34] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1406–1415, 2019. 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 3

[36] Scott E. Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016. 2

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 3

[38] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting unlabelled photos for stronger fine-grained SBIR. In *CVPR*, pages 6873–6883, 2023. 1

[39] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 35(4):119:1–119:12, 2016. 5

[40] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, pages 4222–4235, 2020. 3

[41] Jialin Tian, Xing Xu, Kai Wang, Zuo Cao, Xunliang Cai, and Heng Tao Shen. Structure-aware semantic-aligned network for universal cross-domain retrieval. In *SIGIR*, pages 278–289, 2022. 2, 5, 6, 7

[42] Mengzhu Wang, Jianlong Yuan, Qi Qian, Zhibin Wang, and Hao Li. Semantic data augmentation based distance metric learning for domain generalization. In *ACM MM*, 2022. 2

[43] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2251–2265, 2019. 2

[44] Yuwei Yang, Munawar Hayat, Zhao Jin, Hongyuan Zhu, and Yinjie Lei. Zero-shot point cloud segmentation by semantic-visual aware synthesis. In *CVPR*, pages 11586–11596, 2023. 2

[45] Haonan Zhang, Lianli Gao, Pengpeng Zeng, Alan Hanjalic, and Heng Tao Shen. Depth-aware sparse transformer for video-language learning. In *ACM MM*, pages 4778–4787, 2023. 3

[46] Ji Zhang, Lianli Gao, Xu Luo, Hengtao Shen, and Jingkuan Song. Deta: Denoised task adaptation for few-shot learning. In *ICCV*, pages 11541–11551, 2023. 3

[47] Ji Zhang, Shihan Wu, Lianli Gao, Hengtao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *CVPR*, 2024. 3

[48] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, pages 561–578, 2020. 2

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 2, 3, 5, 6, 7

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16795–16804, 2022. 2, 3

[51] Xiaoping Zhou, Xiangyu Han, Haoran Li, Jia Wang, and Xun Liang. Cross-domain image retrieval: methods and applications. *Int. J. Multim. Inf. Retr.*, 11(3):199–218, 2022. 1

[52] Jinkuan Zhu, Pengpeng Zeng, Lianli Gao, Gongfu Li, Dongliang Liao, and Jingkuan Song. Complementarity-aware space learning for video-text retrieval. *TCSVT*, 2023. 3

[53] Xiaosu Zhu, Jingkuan Song, Yu Lei, Lianli Gao, and Heng Tao Shen. A lower bound of hash codes' performance. In *NeurIPS*, 2022. 8