Retrieval Across Any Domains via Large-scale Pre-trained Model

Jiexi Yan [1] **Zhihui Yin** [2] **Chenghao Xu** [2] **Cheng Deng** [2] **Heng Huang** [3]

Abstract

In order to enhance the generalization ability towards unseen domains, universal cross-domain image retrieval methods require a training dataset encompassing diverse domains, which is costly to assemble. Given this constraint, we introduce a novel problem of data-free adaptive cross-domain retrieval, eliminating the need for real images during training. Towards this goal, we propose a novel Text-driven Knowledge Integration (TKI) method, which exclusively utilizes a pre-trained vision-language model to implement an "aggregation after expansion" training strategy. Specifically, we extract diverse implicit domain-specific information through a set of learnable domain word vectors. Subsequently, a domain-agnostic universal projection, equipped with a non-Euclidean multi-layer perceptron, can be optimized using these assorted text descriptions through the text-proxied domain aggregation. Leveraging the cross-modal transferability phenomenon of the shared latent space, we can integrate the trained domain-agnostic universal projection with the pre-trained visual encoder to extract the features of the input image for the following retrieval during testing. Extensive experimental results on several benchmark datasets demonstrate the superiority of our method.

1. Introduction

With the rapid growth of data uploaded and shared through the Internet in diverse forms or modalities (*e.g.*, viewpoints, lightning, artistic styles, and photographs), retrieval across different domains (Huang et al., 2015; Paul et al., 2021; Hu & Lee, 2022; Hu et al., 2023) has attracted significant attention and has been applied to a wide range of applications, such as e-commerce and surveillance, in recent years. Such cross-domain retrieval (CDR) aims to retrieve relevant instances from a domain (*e.g.*, photo) when giving a query belonging to a different domain (*e.g.*, sketch, quickdraw, etc.).

Owing to the significant distribution shifts observed when training and test data lack alignment, the cross-modal retrieval (CDR) task is often constrained to a predefined retrieval domain, such as sketch-based image retrieval (Liu et al., 2017; Yelamarthi et al., 2018; Dey et al., 2019; Chaudhuri et al., 2023) or infrared-visible person re-identification (Li et al., 2020; Huang et al., 2021; Yu et al., 2023). However, the development of domain-specific retrieval models for each practical CDR task entails substantial training and maintenance costs. In response to this challenge, researchers have sought to establish a universal CDR model capable of direct generalization across diverse domains, introducing a series of methods aligned with this objective (Paul et al., 2021; Tian et al., 2022; Agarwal et al., 2023). To enhance the generalization ability of the CDR model to previously unseen domains, these methodologies leverage well-annotated training data from multiple domains to learn domain-agnostic feature embeddings (Li et al., 2019; Wang et al., 2020; Yan et al., 2024) in the latent space. However, given the need to apply among arbitrary domains, the optimal selection of training domains remains ambiguous. Furthermore, it is extremely expensive and even impossible to label multiple domains in the real world.

To this end, we pose the inquiry of whether it is feasible to employ a pre-trained model for the task of *data-free adaptive cross-domain retrieval* (DFACDR), which entails straightly utilizing a pre-trained model to retrieve instances across arbitrary domains. Such a DFACDR problem, representing the retrieval challenge without access to any realistic training data, has received limited attention in the existing literature, to the best of our knowledge.

In this paper, we investigate the effective employment of a large-scale pre-trained model (Jia et al., 2021; Radford et al., 2021; Yang et al., 2022; Chen et al., 2022; Cho et al., 2023; Yin et al., 2024) in the absence of training data, with the aim of leveraging its potential capabilities for retrieval across arbitrary domains. Since the large-scale pre-trained

[1]School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China [2]School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China [3]Department of Computer Science, University of Maryland College Park, USA. Correspondence to: Cheng Deng <chdeng.xd@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

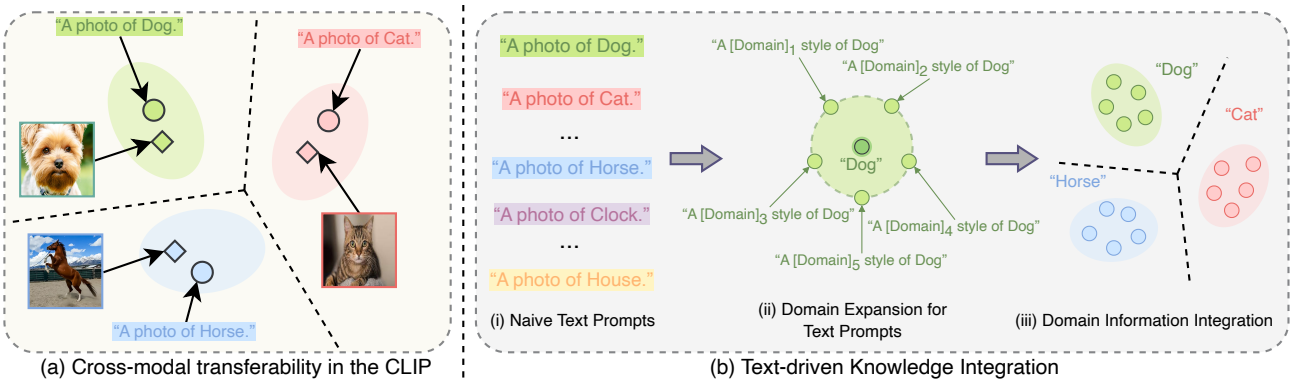


Figure 1. (a) **Cross-modal transferability in CLIP:** Text features have the potential to effectively embody their corresponding visual features within a shared cross-modal latent space. Consequently, it becomes feasible to manipulate visual features through text prompts, by capitalizing on their shared cross-modal latent space. (b) Simple illustration of our **text-driven knowledge integration** method.

model has already seen diverse domains, an instinctive approach for the data-free adaptative cross-domain retrieval task is to extract implicit domain-related information and employ it to acquire domain-agnostic feature embeddings in the latent space, all without the need for additional training data. We posit that leveraging large-scale vision-language models has the potential to provide insights and solutions to the challenging DFACDR task. Given the reciprocal representation capabilities of text and visual feature embeddings in a universal cross-modal latent space (Zhang et al., 2023; Cho et al., 2023) shown in Figure 1 (a), the existing textual representations can be harnessed for the proxies of visual encoders during fine-tuning.

Motivated by this, we propose a text-driven knowledge integration method, dubbed TKI, that takes an "aggregation after expansion" strategy to extract and integrate rich domain-related knowledge for learning domain-agnostic feature embeddings. In the conventional large-scale vision-language model, *e.g.*, CLIP (Radford et al., 2021), the textual category labels are combined with a pre-fixed prompt, such as "A photo of" to form text descriptions (*e.g.*, " A photo of dog."), which cannot reflect the domain-specific information. To explore the implicit domain knowledge, we introduce a prompt-based domain expansion module where a learnable domain placeholder `[domain]` is inserted into the pre-fixed prompt, *i.e.* " A `[domain]` style of `[class]`". In this way, we can make the model aware diverse domain knowledge guided by the learned domain-related prompts. To ensure the effective synthesis of the domain-related prompts, a domain expansion module with the "min-max" strategy is exploited shown in Figure 1 (b). And then, a text-proxied domain aggregation module is proposed to integrate the extracted domain knowledge and take full advantage of it to learn domain-agnostic feature embeddings via a universal projection. Specifically, we force the text features originating from different domains but pertaining to the same category closer in the latent space through the utilization of a contrastive-based loss. Due to the cross-modal transferability in the latent space of CLIP, the text embeddings can be regarded as proxies for image embeddings. Therefore, the image embeddings after the learned domain-agnostic universal projection also have the capability of unified representing data belonging to different domains.

In summary, the main contributions of this work include:

- We explore the possibility of exploiting the pre-trained vision-language model to effectively address domain-free cross-domain retrieval.
- We propose a new text-driven knowledge integration method that fully exploits the cross-modal semantic guidance ability of CLIP to enable the model to understand more domain-related knowledge. In our method, a two-stage learning strategy is adopted to optimize learnable domain-related prompts and capture unified domain-agnostic semantics, respectively.
- Extensive experiments on several cross-domain datasets are conducted to analyze our TKI. We show that the proposed method can capture more informative domain-related semantics, thereby significantly improving performance over state-of-the-art methods.

2. Related Work

Cross-Domain Retrieval. The field of cross-domain retrieval has received considerable attention in recent years due to the increasing demand for integrating information from various sources. This interdisciplinary field has broad practical applications in e-commerce, multimedia, and web

search. A significant body of research has focused on developing methods for cross-domain retrieval. For example, sketch-based image retrieval (Yelamarthi et al., 2018; Liu et al., 2019; Xu et al., 2020; Dutta et al., 2020; Wang et al., 2021) aims to retrieve natural images for sketch queries, where the domains are pre-defined. While substantial progress has been made in cross-domain retrieval, further research is needed to address ongoing challenges and improve the effectiveness of retrieval across diverse domains without pre-definition.

In cases where the test instances originate from a new domain, the network necessitates re-training with data pertinent to the corresponding domains. This not only mandates that training be executed for each domain pair with an adequate volume of data, but it also requires the domain from which retrieval will be conducted to be known a-priori. To tackle this issue, universal cross-domain retrieval (Paul et al., 2021) and its variants (Paul et al., 2022; Tian et al., 2022; Fang et al., 2023) have been proposed to exploit training data from multiple domains to optimize a cross-domain retrieval model, which can deal with the test samples from unseen domains. However, when there is a need for application across arbitrary domains, the optimal selection of training domains remains uncertain (Agarwal et al., 2023). Additionally, the task of labeling multiple large-scale domains in the real world can be exceedingly costly and at times, even unfeasible. Therefore, in this paper, we turn our focus to a more demanding scenario, that is, data-free cross-domain image retrieval, where no actual training instances are within reach.

Large-scale Pre-trained Vison-Language Model. Large-scale pre-training enables models to acquire generalizable features and representations. These learned characteristics can then be fine-tuned to cater to specific tasks, significantly enhancing the model's performance and accuracy (Zhang et al., 2022; Zhou et al., 2022b;a; Gao et al., 2023). Large-scale pre-trained vision-language models, such as Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) and ALIGN (Jia et al., 2021), are a specialized type of pre-trained model that have been trained on a massive amount of visual and linguistic data. These models are engineered to comprehend and generate meaningful outputs from both visual (images, videos) and linguistic (text) inputs, thereby understanding the complex interplay between visual and textual data. This capability proves to be invaluable for tasks including image captioning (Mokady et al., 2021), visual questioning answering (Parelli et al., 2023), and visual dialogue (Kang et al., 2023). CLIP, for instance, benefits from the richness of semantic information and the large-scale availability of images, allowing it to learn generic feature representations within the same cross-modal embedding space. This facilitates the model's generalization ability in downstream tasks without labels, known as the zero-shot setting (Romera-Paredes & Torr, 2015; Zhang & Saligrama, 2016; Xie et al., 2020). Building on CLIP, several works, such as ALBEF (Li et al., 2021) and SimVLF (Wang et al., 2022), have been introduced to encompass a broader variety of downstream tasks.

3. Preliminaries

Problem Definition. In this paper, we focus on a more realistic yet challenging problem, *i.e.*, data-free adaptative cross-domain retrieval where well-annotated multi-domain data is not available for training a cross-domain retrieval model. Furthermore, we hope the learned cross-domain retrieval model can effectively retrieve similar instances between arbitrary domains rather than predefined domains. It seems to be an impossible task that has not attracted enough attention.

Motivation Illustration. Large-scale pre-trained models, trained on extensive data across diverse domains, have acquired significant domain-related knowledge, which facilitates effective cross-domain retrieval. However, this valuable information is implicitly stored within the pre-trained model. Direct deployment of these large-scale pre-trained models, such as the visual encoder in CLIP, for cross-domain retrieval does not optimally stimulate and utilize the domain-related knowledge for effective retrieval across various domains. To address this, we propose leveraging large-scale vision-language models like CLIP, using the readily available text representations as a surrogate for actual training data to tap into domain-related knowledge. This approach is grounded in the understanding that text features can effectively represent their associated image features within a shared vision-language space.The primary task of our work is efficiently utilizing the available textual representations as the proxy to learn a domain-agnostic universal projection in the cross-modal latent space that can be straightly combined with the pre-trained visual encoder for cross-domain retrieval.

4. Text-driven Knowledge Integration

In this paper, we propose a new text-driven knowledge integration method (TKI) that can exploit textual descriptions as proxies to explore implicit domain-relevant knowledge with a pre-trained large-scale vision-language model such as CLIP (Radford et al., 2021). During fine-tuning, we adopt an "aggregation after expansion " strategy to effectively excavate and utilize intrinsic diverse domain information. The overall framework is shown in Figure 2.

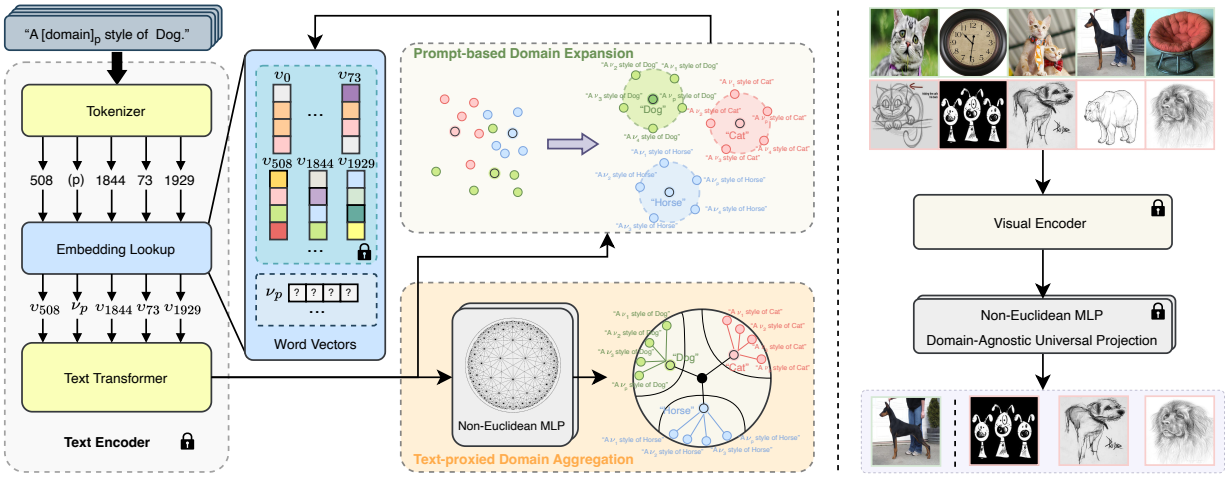


Figure 2. The overall framework of our proposed method. (Left) The training stage: We first expand the pre-set word vector set in the embedding lookup process with the domain-relevant embeddings $\{\boldsymbol{\nu}_p\}_{p=1}^P$, which correspond to the domain tokens $\{\texttt{[DOMAIN]}_p\}_{p=1}^P$, respectively. Then, these diverse text prompts are utilized to optimize a non-Euclidean MLP for acquiring a domain-agnostic universal projection within the shared cross-modal latent space. (Right) The test stage: Test images are fed into the visual encoder of CLIP combined with the learned non-Euclidean MLP to derive visual features for retrieval.

4.1. Prompt-based Domain Expansion

Though diverse domain-relevant knowledge is stored in the pre-trained CLIP, straightly employing the visual encoder of CLIP to extract image features cannot excavate such implicit domain-relevant knowledge and further utilize it. Without any realistic data for training, we turn to exploit the text representation via the text encoder of CLIP to assist in exploring implicit domain-relevant knowledge. To this end, we first explicitly exhibit the domain diversity information in the shared cross-modal space via the domain expansion process with a series of learnable domain prompts.

In downstream tasks utilizing the CLIP model, the textual representation is usually obtained from a pre-set natural language prompt such as "A photo of `[CLASS]`.", where the category placeholder `[CLASS]` is substituted by the relevant textual label, such as "dog" or "cat". Given that this form of text representation is domain-independent, its text embedding aligns with the visual embeddings of all images that fall under the corresponding category, but diverse domain styles, in a shared cross-modal space. Consequently, it is not feasible to directly utilize these simplistic text prompts as substitutes to uncover domain-relevant knowledge for the purpose of learning a universal projection in the shared cross-modal space.

Specifically, we insert a series of learnable domain tokens $\{\texttt{[DOMAIN]}_p\}_{p=1}^P$ into the original text prompts and derive the synthesized domain-related text prompts "A $\texttt{[DOMAIN]}_p$ style of $\texttt{[CLASS]}_i$.", which can effectively capture the diverse domain information. Here, $\texttt{[CLASS]}_i$ represents the text label of i-th category. During the embedding lookup procedure, the domain placeholder $\texttt{[DOMAIN]}_p$ will be replaced by the p-th domain-specific word vector $\boldsymbol{\nu}_p$. In an endeavor to proficiently imitate the distribution shifts among diverse domains in the latent space, it becomes imperative to adaptively acquire P unique domain-specific word vectors $\{\boldsymbol{\nu}_p\}_{p=1}^P$. To ensure a distinct representation of diverse domains, these vectors should exhibit discernible differences from one another, while simultaneously maintaining independence from any category-specific information.

For a complete text prompt with the p-th domain and i-th class label denoted as Λ_{pi} (*i.e.*, "A $\texttt{[DOMAIN]}_p$ style of $\texttt{[CLASS]}_i$."), we simply divide it into two parts, *i.e.*,domain-specific prompt Λ_p^d ("A $\texttt{[DOMAIN]}_p$ style of ") and category-specific prompt Λ_i^c ("$\texttt{[CLASS]}_i$"). Upon introducing these prompts into the pre-trained text encoder $\mathcal{T}(\cdot)$ incorporated in CLIP, we can extract their corresponding text features denoted as $\boldsymbol{t}_{pi} = \mathcal{T}(\Lambda_{pi}) \in \mathbb{R}^d$, $\boldsymbol{t}_p^d = \mathcal{T}(\Lambda_p^d) \in \mathbb{R}^d$ and $\boldsymbol{t}_i^c = \mathcal{T}(\Lambda_i^c) \in \mathbb{R}^d$, respectively. To efficiently learn unique domain-specific word vectors $\{\boldsymbol{\nu}_p\}_{p=1}^P$, we adopt a "min-max" learning strategy utilizing these text features in the shared cross-modal space.

First, we propose a **domain-wise diversity maximum loss** to encourage the diversity between every two domain-specific word vectors, where the formula is

$$\mathcal{L}_{max} = \frac{1}{P(P-1)} \sum_{p=1}^{P} \sum_{q \neq p} \log\left(1 + e^{\alpha(s_{pq}^d - \mu)}\right), \quad (1)$$

where s_{pq}^d denotes the cosine similarity between p-th and

q-th domain-specific text features, which is computed by

$$s_{pq}^{d} = \frac{\boldsymbol{t}_p^d \cdot \boldsymbol{t}_q^d}{\|\boldsymbol{t}_p^d\|_2 \|\boldsymbol{t}_q^d\|_2}. \tag{2}$$

Note that α and μ are hyperparameters. This loss aims to spread the distance between different domain-specific text features in the latent space, which can effectively ensure the diversity of the learned domain-specific word vectors.

Furthermore, in order to avoid being confounded by undesirable category-wise knowledge during the learning process, a constraint, *i.e.*, **category-wise correlation minimum loss**, is enforced to the domain-specific text features so that all domain-specific word vectors are independent of categories. Specifically, text features $\{\boldsymbol{t}_{1i}, \cdots, \boldsymbol{t}_{pi}, \cdots, \boldsymbol{t}_{Pi}\}$ synthesized via different domain-specific word vectors but the same class label should be pushed together in the latent space. To this end, we can employ a contrastive loss, where text features $\{\boldsymbol{t}_{1i}, \cdots, \boldsymbol{t}_{pi}, \cdots, \boldsymbol{t}_{Pi}\}$ are regarded as different augmentations of the i-th category. Nonetheless, while the contrastive-based loss capitalizes on rich sample-to-sample relationships, it is encumbered by the elevated training complexity required for optimizing dense sample-to-sample relationships. Regrettably, certain intricate relationships may impede performance. Therefore, we resort to employing a prototype-based contrastive loss to construct our category-wise correlation minimum loss as follows:

$$\mathcal{L}_{min} = -\frac{1}{PC}\sum_{p=1}^{P}\sum_{i=1}^{C}\log\left(\frac{\exp(s_{pii}^{c}/\tau)}{\sum_{j=1}^{C}\exp(s_{pij}^{c}/\tau)}\right), \tag{3}$$

where C is the number of categories while τ is the temperature coefficient. Here, the category-specific feature $\boldsymbol{t}_i^c$ is regarded as the i-th class prototype, and s_{pij} denotes the cosine similarity between $\boldsymbol{t}_{pi}$ and $\boldsymbol{t}_j^c$, which is computed by

$$s_{pij}^{c} = \frac{\mathbf{t}_{pi} \cdot \mathbf{t}_j^c}{\|\mathbf{t}_{pi}\|_2 \|\mathbf{t}_j^c\|_2}. \tag{4}$$

The overall loss for learning P domain-specific word vectors during the domain expansion process is summarized as:

$$\mathcal{L}_{exp} = \mathcal{L}_{min} + \lambda\mathcal{L}_{max}, \tag{5}$$

where λ is a trade-off hyperparameter.

4.2. Text-proxied Domain Aggregation

Owing to the application of the prompt-based domain expansion, an abundance of domain-specific information can be distinctly captured by the diverse domain word vectors that have been learned. Consequently, these text representations, fortified with diverse domain word vectors, can be employed as proxies for the purpose of optimizing a non-Euclidean multi-layer perceptron (MLP) $h(\cdot)$ to obtain a domain-agnostic universal projection within the shared cross-modal latent space.

In comparison with the conventional MLP, our modified version only replaces the Euclidean layer with a non-Euclidean layer of equivalent size, while the remainder of the structure remains unaltered. The sole additional hyperparameter introduced is the radius of the Poincaré ball. It is noteworthy that the backbone structure produces Euclidean representations. Consequently, the intermediate representations necessitate transformation prior to their input into the non-Euclidean layer. It is worth noting that the backbone structure produces Euclidean representations; hence, the intermediate representations require transformation prior to being input into the non-Euclidean layer. Specifically, given a Poincaré ball $\mathbb{D}^r$ of radius r, we assume the intermediate representation is in the tangent space of the Poincaré ball at the origin, Ω_0^r. Owing to the advanced principles of the hyperbolic Poincaré ball, the acquired domain-agnostic universal projection is better equipped to capture the implicit hierarchical structure prevalent among cross-domain image data. In this structure, images of identical categories from disparate domains exhibit diversified characteristics, yet they should possess heightened similarity within the latent space.

In pursuit of this objective, it is essential to aggregate the text features that are associated with the same class, yet encompass diverse domain-specific information. Specifically, given the $P \times C$ embedded diverse text features $\{\boldsymbol{t}_{pi}\}$, we can train the non-Euclidean MLP by a supervised contrastive loss as follows:

$$\mathcal{L}_{agg} = -\frac{1}{PC}\sum_{p=1}^{P}\sum_{i=1}^{C}\log\left(\frac{\sum_{q\neq p}\exp(h(\boldsymbol{t}_{pi})\cdot h(\boldsymbol{t}_{qi})/\tau)}{\sum_{q,j=1}^{P,C}\exp(h(\boldsymbol{t}_{pi})\cdot h(\boldsymbol{t}_{qj})/\tau)}\right). \tag{6}$$

4.3. Retrieval During Testing

Thanks to the cross-modal transferability in the CLIP model, text features can effectively represent their relevant visual features in the shared cross-modal latent space. Therefore, we can combine the trained domain-agnostic universal projection $h(\cdot)$ with the visual encoder $\mathcal{I}(\cdot)$ to extract the image features during testing. Specifically, given an input image $\boldsymbol{x}_n$, we can obtain the corresponding image feature $\boldsymbol{z}_n = h(\mathcal{I}(\boldsymbol{x}_n))$ for downstream retrieval task.

5. Experiments

5.1. Experimental Settings

Datasets. To comprehensively evaluate the effectiveness of our method, we conduct experiments on three cross-

domain benchmarks, *i.e.*, DomainNet (Peng et al., 2019), PACS (Li et al., 2017), and Office-Home (Venkateswara et al., 2017). Within the challenging data-free cross-domain retrieval task, we do not exploit any actual data for training. The actual images in datasets are only used for test. The statistics about these three datasets are summarized in Table 1.

Table 1. Statistics details of the employed datasets in experiments.

	DomainNet	PACS	Office-Home
# Domains	6	4	4
# Classes	345	7	65
# Samples	596006	9991	15588

Evalutation Metrics. To more effectively illustrate the efficacy of our proposed method, we employ mean average precision (mAP), top-200 mean average precision (mAP@200), and top-200 precision (Prec@200) as the evaluation metrics. The term "top-k" represents the evaluation exclusively based on the first k retrieved samples.

Experimental Implementation. In our experimental setup, we utilize CLIP as the large-scale vision-language model. Throughout the training process, the text and visual encoders within our comprehensive framework remain static. We leverage the publicly accessible pre-trained model, in which the Transformer serves as the backbone of the text encoder, and ViT-B/32 (Dosovitskiy et al., 2021) functions as the visual encoder. Additionally, we incorporate two other architectures for the visual encoder, specifically ResNet-50 (He et al., 2016), and ViT-L/14 (Dosovitskiy et al., 2021), to facilitate a more in-depth evaluation. For benchmarking purposes, we employ the pre-trained visual encoder in CLIP (referred to as Vanilla CLIP) as the baseline.

Our approach is implemented in PyTorch and trained with an NVIDIA A6000 GPU. The input images are resized to 224×224 during testing. We train our model by Adam (Kingma & Ba, 2015) optimizer with the same hyper-parameters (learning rate, τ, and λ are set as 0.005, 0.1, and 1, respectively) in all experiments.

5.2. Performance Evaluation

Quantitative Results. We present the results of cross-domain retrieval on the DomainNet, PACS, and Office-Home datasets, delineated in Tables 2 and 3. In contrast to the baseline, our proposed method demonstrates a significant enhancement in performance across all evaluative measures. This evidence corroborates the efficacy of our introduced text-driven approach in augmenting the generalization capacity of the pre-trained model, *i.e.,* CLIP. Importantly, this improvement is achieved without the necessity of any training images, highlighting the potential for retrieval across an unrestricted range of domains.

Qualitative Results. We undertake a qualitative assessment of the diverse text prompts on the PACS dataset, utilizing t-SNE visualization for this analysis. As illustrated in Figure 3, our methodology engenders a multitude of domains, while concurrently maintaining the integrity of content information. The diverse text features that are derived from an identical class name exhibit similar semantics, albeit with varied variations. This outcome substantiates our ability to efficaciously emulate a range of distribution shifts in the latent space of a large-scale vision-language model, by synthesizing diverse domains through the application of learnable, domain-specific word vectors.

Furthermore, to better demonstrate the effectiveness of the learned domain-specific text words, we conduct a text-to-image synthesis visualization. In Figure 4 4, we translate the learned text prompts ("A $[\text{DOMAIN}]_p$ style of house/Dog.") via a pre-trained Stable Diffusion v1.4 (Rombach et al., 2022). Here, we select two labels, i.e., house and dog and randomly select 4 different domain word vectors, where the word vectors are learned for the PACS dataset. To match the pre-trained stable diffusion model, the text prompts are learned with the CLIP ViT-L/14 as the backbone.

5.3. Ablation Study

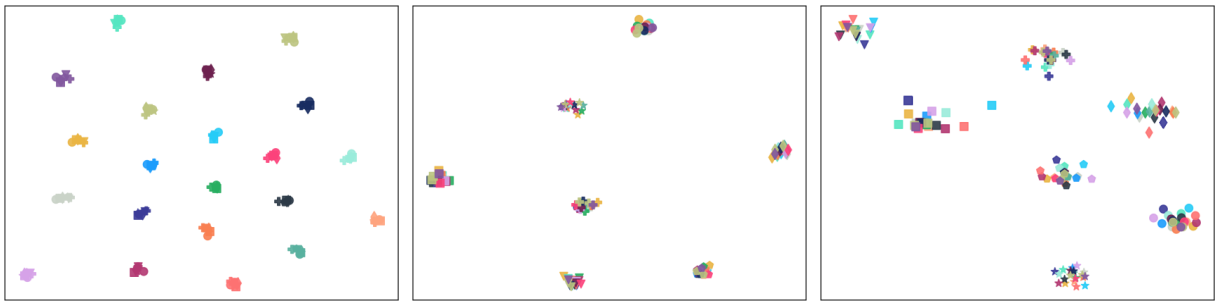
The effect of different compositional losses in domain expansion. To better measure the effectiveness of the proposed losses in our method, we conduct an ablation study on the PACS dataset (Li et al., 2017). The experimental results using different combinations of losses on PACS are shown in Table 5. Upon integrating both losses, we are able to generate a diversity of domains without compromising the integrity of the content information. We can see that the generated images show good diversity.

The effect of different backbones. To comprehensively evaluate our method, we conduct experiments with different backbones of the visual encoder. The experimental results on the PACS dataset (Li et al., 2017) using different backbones of the visual encoder are reported in Table 4. We can see that our method can achieve consistently superior performance when using different backbones of the visual encoder, which can further demonstrate the effectiveness of our method in a large range.

The effect of the number of domain-specific word vectors. To measure the effect of domain-specific word vectors, we conduct the ablation study on the PACS datset (Li et al., 2017) and show the experimental results with different numbers of domain-specific word vectors in Figure 5. For a fair comparison, we set $P = 20$ in all experiments.

Table 2. Experimental results (%) on the DomainNet dataset.

Method	Query Domain	Gallery Domain: Real mAP	Real Prec	Sketch mAP	Sketch Prec	Quickdraw mAP	Quickdraw Prec	Infograph mAP	Infograph Prec	Painting mAP	Painting Prec	Clip-art mAP	Clip-art Prec	Avg. mAP	Avg. Prec
Vanilla CLIP	Real	-	-	21.03	26.39	6.80	13.62	8.11	11.87	19.41	24.18	23.46	24.39	15.76	20.09
Our TKI	Real	-	-	**27.36**	**31.60**	**7.81**	**15.02**	**14.07**	**16.96**	**27.04**	**30.57**	**30.27**	**28.83**	**21.31**	**24.60**
Vanilla CLIP	Sketch	18.94	25.90	-	-	6.63	13.05	3.75	6.51	12.08	16.84	15.30	17.88	11.34	16.04
Our TKI	Sketch	**28.96**	**35.83**	-	-	**7.78**	**14.66**	**7.77**	**11.24**	**18.03**	**22.86**	**21.20**	**22.41**	**16.75**	**21.40**
Vanilla CLIP	Quickdraw	1.00	2.27	1.15	2.91	-	-	0.21	0.83	0.35	1.25	1.23	2.52	0.80	1.96
Our TKI	Quickdraw	**2.69**	**4.70**	**2.57**	**4.92**	-	-	**0.66**	**1.89**	**1.24**	**2.79**	**2.68**	**4.25**	**1.97**	**3.71**
Vanilla CLIP	Infograph	17.44	23.90	9.85	14.59	4.38	8.51	-	-	8.28	11.83	11.13	13.85	10.22	14.54
Our TKI	Infograph	**21.95**	**28.15**	**13.07**	**17.64**	**5.05**	**9.48**	-	-	**11.42**	**14.93**	**14.24**	**16.16**	**13.15**	**17.27**
Vanilla CLIP	Painting	25.97	33.58	16.12	21.55	6.03	11.85	4.57	7.58	-	-	16.14	18.39	13.77	18.59
Our TKI	Painting	**34.55**	**41.38**	**21.34**	**26.38**	**6.99**	**13.04**	**8.55**	**11.92**	-	-	**21.30**	**22.01**	**18.55**	**22.95**
Vanilla CLIP	Clip-art	22.81	29.93	16.00	21.15	7.49	14.21	4.15	6.87	10.79	14.75	-	-	12.25	17.38
Our TKI	Clip-art	**34.10**	**41.30**	**21.97**	**26.89**	**9.13**	**16.41**	**8.42**	**11.93**	**18.04**	**21.98**	-	-	**18.33**	**23.70**



(a) $\mathcal{L}_{max}$ (b) $\mathcal{L}_{min}$ (c) $\mathcal{L}_{max} + \mathcal{L}_{min}$

Figure 3. t-SNE visualization results on the PACS dataset utilizing learned diverse text features. Different colors denote features obtained from different domain-specific word vectors, and different shapes indicate features obtained from different class text labels.

Figure 4. Prompt-based image generation results utilizing the learned diverse text prompts.

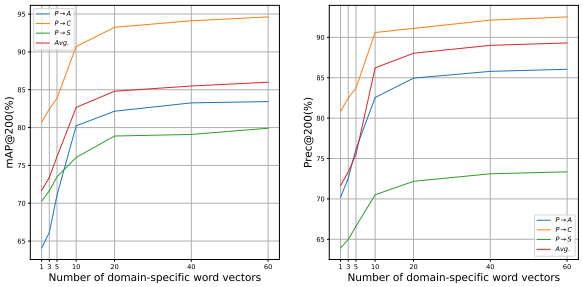


Figure 5. mAP@200 and Prec@200 results on PACS with regard to the number of learnable domain-specific word vectors P.

Table 3. Experimental results (%) on PACS and Office Home datasets. In PACS, "A", "C", "P", and "S" represent Art Painting, Cartoon, Photo, and Sketch, respectively. In Office Home, "A", "C", "P", and "R" represent Art, Clipart, Product, and Real, respectively.

	PACS								Office-Home							
	A → C		A → P		A → S		Avg.		A → C		A → P		A → R		Avg.	
Method	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Vanilla CLIP	67.46	71.44	50.41	56.83	65.40	62.39	61.09	63.55	28.26	16.92	29.45	17.07	32.80	19.18	30.17	17.72
Our TKI	**80.94**	**82.86**	**66.40**	**70.73**	**72.01**	**67.52**	**73.12**	**73.70**	**33.27**	**18.65**	**37.79**	**20.05**	**40.26**	**21.95**	**37.11**	**20.22**
	C → A		C → P		C → S		Avg.		C → A		C → P		C → R		Avg.	
Method	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Vanilla CLIP	48.71	58.74	52.58	59.43	74.23	71.84	58.51	63.34	16.97	8.51	22.65	14.64	21.74	14.58	20.45	12.58
Our TKI	**70.23**	**76.07**	**73.66**	**77.30**	**78.85**	**75.59**	**74.25**	**76.32**	**22.70**	**9.92**	**31.17**	**18.24**	**29.97**	**17.75**	**27.95**	**15.30**
	P → A		P → C		P → S		Avg.		P → A		P → C		P → R		Avg.	
Method	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Vanilla CLIP	64.09	70.21	80.73	80.84	70.27	63.93	71.70	71.66	24.80	9.94	26.52	16.49	39.75	21.00	30.36	15.81
Our TKI	**82.16**	**84.95**	**93.35**	**91.11**	**78.88**	**72.18**	**84.80**	**88.03**	**32.10**	**11.55**	**34.09**	**19.14**	**47.82**	**23.68**	**38.00**	**18.12**
	S → A		S → C		S → P		Avg.		R → A		R → C		R → P		Avg.	
Method	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Vanilla CLIP	44.02	56.14	56.87	61.83	47.18	55.77	49.36	57.91	30.74	11.88	31.51	18.12	47.48	24.03	36.58	18.01
Our TKI	**65.01**	**71.59**	**74.51**	**75.37**	**66.06**	**70.85**	**68.53**	**72.60**	**37.33**	**13.30**	**37.56**	**20.13**	**54.82**	**26.23**	**43.24**	**19.89**

Table 4. Experimental results (%) exploiting different visual encoder backbones on PACS. "A", "C", "P", and "S" represent Art Painting, Cartoon, Photo, and Sketch, respectively.

	ResNet-50								ViT-L/14							
	A → C		A → P		A → S		Avg.		A → C		A → P		A → S		Avg.	
Method	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Vanilla CLIP	52.53	61.04	42.34	53.43	45.31	49.88	46.73	54.78	76.09	77.42	56.51	61.07	68.97	67.13	67.19	68.54
Our TKI	**78.07**	**81.14**	**71.63**	**76.78**	**63.86**	**63.26**	**71.19**	**73.73**	**87.29**	**87.08**	**75.63**	**76.74**	**77.89**	**72.09**	**80.27**	**78.64**
	C → A		C → P		C → S		Avg.		C → A		C → P		C → S		Avg.	
Method	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Vanilla CLIP	32.90	45.24	46.94	57.84	58.24	61.44	46.03	54.84	51.80	59.24	52.84	59.04	77.95	75.03	60.62	64.44
Our TKI	**62.86**	**71.07**	**79.02**	**81.47**	**72.56**	**71.81**	**71.48**	**74.78**	**80.87**	**83.92**	**80.30**	**81.90**	**87.88**	**81.19**	**83.02**	**82.34**
	P → A		P → C		P → S		Avg.		P → A		P → C		P → S		Avg.	
Method	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Vanilla CLIP	53.57	61.79	77.37	78.43	58.63	57.28	63.19	65.83	61.94	67.53	83.53	82.69	68.89	63.39	71.45	71.20
Our TKI	**77.41**	**81.34**	**95.77**	**93.06**	**75.93**	**70.91**	**83.04**	**81.77**	**81.43**	**84.95**	**94.30**	**91.63**	**78.10**	**70.40**	**84.61**	**82.33**
	S → A		S → C		S → P		Avg.		S → A		S → C		S → P		Avg.	
Method	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Vanilla CLIP	25.63	40.25	38.75	48.30	34.93	47.76	33.10	45.44	56.29	64.49	78.80	78.43	52.23	58.49	62.44	67.14
Our TKI	**48.40**	**59.06**	**61.38**	**66.21**	**58.99**	**65.56**	**56.26**	**63.61**	**83.65**	**85.81**	**89.93**	**87.07**	**76.38**	**77.47**	**83.32**	**83.45**

Effect of the text label information. During training, we need to use the text label information as prior to conduct domain expansion. To further evaluate the zero-shot generalization ability of our method, we conduct an ablation study on the usage of the text label information. We compare the two different settings on the PACS dataset (Li et al., 2017): (1) Using the corresponding label information in the dataset; (2) Randomly selecting text labels that don't overlap with the actual label in the dataset. The experimental results are shown in Table 6. We can see that although using the randomly selected text label information, our method can also improve the performance of the baseline, which can effectively demonstrate the zero-shot generalization ability of our method.

Table 5. Experimental results (%) with regard to different losses. "A", "C", "P", and "S" represent Art Painting, Cartoon, Photo, and Sketch, respectively.

		A → C		A → P		A → S		Avg.	
$\mathcal{L}_{max}$	$\mathcal{L}_{min}$	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
-	-	67.46	71.44	50.41	56.83	65.40	62.39	61.09	63.55
✓	-	77.44	80.02	61.15	65.45	70.04	66.40	69.54	70.62
-	✓	74.01	77.06	61.45	65.99	69.93	65.16	68.46	69.40
✓	✓	**80.94**	**82.86**	**66.40**	**70.73**	**72.01**	**67.52**	**73.12**	**73.70**

Table 6. Experimental results (%) using different label information. "A", "C", "P", and "S" represent Art Painting, Cartoon, Photo, and Sketch, respectively.

	A → C		A → P		A → S		Avg.	
	mAP	Prec	mAP	Prec	mAP	Prec	mAP	Prec
Baseline	67.46	71.44	50.41	56.83	65.40	62.39	61.09	63.55
Strategy 2	68.74	74.88	56.36	61.75	68.42	64.62	64.51	67.08
Strategy 1	**80.94**	**82.86**	**66.40**	**70.73**	**72.01**	**67.52**	**73.12**	**73.70**

6. Conclusion

We introduce a new data-free fine-tuning approach, TKI, which amalgamates a range of domains within a shared cross-modal space using learnable domain words. This approach, which does not rely on any images, is designed to address arbitrary cross-domain retrieval. TKI emulates a variety of distribution shifts within the latent space of a large-scale pre-trained model, thereby enhancing its generalization capability across disparate domains. This proposed methodology achieves state-of-the-art retrieval results on multiple cross-domain benchmarks, all without the utilization of any actual training data.

Acknowledgement

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600), Joint Fund of Ministry of Education of China (8091B022149, 8091B02072404), National Natural Science Foundation of China (62132016, 62171343, 62071361, and 62302372), and Fundamental Research Funds for the Central Universities (ZDRC2102, XJSJ23036).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Karanam, S., Srinivasan, B. V., and Banerjee, B. Contrastive learning of semantic concepts for open-set cross-domain retrieval. In *WACV*, pp. 4115–4124, 2023.

Chaudhuri, A., Bhunia, A. K., Song, Y.-Z., and Dutta, A. Data-free sketch-based image retrieval. In *CVPR*, pp. 12084–12093, 2023.

Chen, S., Gong, C., Li, J., Yang, J., Niu, G., and Sugiyama, M. Learning contrastive embedding in low-dimensional space. *NeurIPS*, 35:6345–6357, 2022.

Cho, J., Nam, G., Kim, S., Yang, H., and Kwak, S. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *ICCV*, pp. 15702–15712, 2023.

Dey, S., Riba, P., Dutta, A., Llados, J., and Song, Y.-Z. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, pp. 2179–2188, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Dutta, T., Singh, A., and Biswas, S. Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir. In *ECCV*, pp. 349–364, 2020.

Fang, K., Song, J., Gao, L., Zeng, P., Cheng, Z.-Q., Li, X., and Shen, H. T. Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. *arXiv preprint arXiv:2312.12478*, 2023.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, pp. 1–15, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Hu, C. and Lee, G. H. Feature representation learning for unsupervised cross-domain image retrieval. In *ECCV*, pp. 529–544. Springer, 2022.

Hu, C., Zhang, C., and Lee, G. H. Unsupervised feature representation learning for domain-generalized cross-domain image retrieval. In *ICCV*, pp. 11016–11025, 2023.

Huang, J., Feris, R. S., Chen, Q., and Yan, S. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, pp. 1062–1070, 2015.

Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, P., and Zhang, Z. Alleviating modality bias training for infrared-visible person re-identification. *IEEE TMM*, 24:1570–1582, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916. PMLR, 2021.

Kang, G.-C., Kim, S., Kim, J.-H., Kwak, D., and Zhang, B.-T. The dialog must go on: Improving visual dialog via generative self-training. In *CVPR*, pp. 6746–6756, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *ICCV*, pp. 5542–5550, 2017.

Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.-Z., and Hospedales, T. M. Episodic training for domain generalization. In *ICCV*, pp. 1446–1455, 2019.

Li, D., Wei, X., Hong, X., and Gong, Y. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, volume 34, pp. 4610–4617, 2020.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021.

Liu, L., Shen, F., Shen, Y., Liu, X., and Shao, L. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, pp. 2862–2871, 2017.

Liu, Q., Xie, L., Wang, H., and Yuille, A. L. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, pp. 3662–3671, 2019.

Mokady, R., Hertz, A., and Bermano, A. H. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., and Hofmann, T. Clip-guided vision-language pre-training for question answering in 3d scenes. In *CVPR*, pp. 5606–5611, 2023.

Paul, S., Dutta, T., and Biswas, S. Universal cross-domain retrieval: Generalizing across classes and domains. In *ICCV*, pp. 12056–12064, 2021.

Paul, S., Saha, A., and Samanta, A. Ttt-ucdr: Test-time training for universal cross-domain retrieval. *arXiv preprint arXiv:2208.09198*, 2022.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *ICCV*, pp. 1406–1415, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Romera-Paredes, B. and Torr, P. An embarrassingly simple approach to zero-shot learning. In *ICML*, pp. 2152–2161. PMLR, 2015.

Tian, J., Xu, X., Wang, K., Cao, Z., Cai, X., and Shen, H. T. Structure-aware semantic-aligned network for universal cross-domain retrieval. In *ACM SIGIR*, pp. 278–289, 2022.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pp. 5018–5027, 2017.

Wang, S., Yu, L., Li, C., Fu, C.-W., and Heng, P.-A. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, pp. 159–176. Springer, 2020.

Wang, Z., Wang, H., Yan, J., Wu, A., and Deng, C. Domain-smoothing network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:2106.11841*, 2021.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022.

Xie, G.-S., Liu, L., Zhu, F., Zhao, F., Zhang, Z., Yao, Y., Qin, J., and Shao, L. Region graph embedding network for zero-shot learning. In *ECCV*, pp. 562–580, 2020.

Xu, X., Yang, M., Yang, Y., and Wang, H. Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval. In *IJCAI*, pp. 984–990, 2020.

Yan, J., Deng, C., Huang, H., and Liu, W. Causality-invariant interactive mining for cross-modal similarity learning. *IEEE TPAMI*, 2024.

Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., and Huang, J. Vision-language pre-training with triple contrastive learning. In *CVPR*, pp. 15671–15680, 2022.

Yelamarthi, S. K., Reddy, S. K., Mishra, A., and Mittal, A. A zero-shot framework for sketch based image retrieval. In *ECCV*, pp. 300–317, 2018.

Yin, Z., Yan, J., Xu, C., and Deng, C. Asymmetric mutual alignment for unsupervised zero-shot sketch-based image retrieval. In *AAAI*, volume 38, pp. 16504–16512, 2024.

Yu, H., Cheng, X., Peng, W., Liu, W., and Zhao, G. Modality unifying network for visible-infrared person re-identification. In *ICCV*, pp. 11185–11195, 2023.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, pp. 493–510. Springer, 2022.

Zhang, Y., HaoChen, J. Z., Huang, S.-C., Wang, K.-C., Zou, J., and Yeung, S. Diagnosing and rectifying vision models using language. *arXiv preprint arXiv:2302.04269*, 2023.

Zhang, Z. and Saligrama, V. Zero-shot learning via joint latent similarity embedding. In *CVPR*, pp. 6034–6042, 2016.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *CVPR*, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022b.

A. Other Experimental Results

We add the ablation study about the non-Euclidean MLP on the DomainNet dataset and report the Prec@1 results in Table 7.

Table 7. Performance Comparison on Different Domains

		Gallery Domain						
Method	Query Domain	Real	Sketch	Quickdraw	Infograph	Painting	Clipart	Average
Vanilla CLIP	Real	–	53.56	32.77	39.11	55.22	58.94	47.92
Ours w/ non-Euclidean MLP	Real	**–**	**59.75**	**35.97**	**50.50**	**60.01**	**63.82**	**54.01**
Ours w/ Euclidean MLP	Real	–	57.64	34.22	46.24	58.38	61.54	51.60
Vanilla CLIP	Sketch	39.83	–	27.26	23.36	35.91	41.17	33.51
Ours w/ non-Euclidean MLP	Sketch	**46.36**	**–**	**29.42**	**32.22**	**40.95**	**46.31**	**39.05**
Ours w/ Euclidean MLP	Sketch	41.16	–	28.54	28.40	38.22	43.34	35.93
Vanilla CLIP	Quickdraw	4.57	6.49	–	2.90	2.14	5.69	4.36
Ours w/ non-Euclidean MLP	Quickdraw	**6.89**	**9.07**	**–**	**4.69**	**4.89**	**7.98**	**6.70**
Ours w/ Euclidean MLP	Quickdraw	5.25	8.13	–	3.16	3.64	6.72	5.38
Vanilla CLIP	Infograph	30.23	25.92	17.64	–	22.84	27.37	24.80
Ours w/ non-Euclidean MLP	Infograph	**35.38**	**30.25**	**19.03**	**–**	**27.31**	**31.10**	**28.61**
Ours w/ Euclidean MLP	Infograph	33.80	28.98	18.23	–	25.28	29.34	27.13
Vanilla CLIP	Painting	45.20	41.29	26.06	24.86	–	38.85	35.25
Ours w/ non-Euclidean MLP	Painting	**51.34**	**45.59**	**27.83**	**34.78**	**–**	**44.25**	**40.76**
Ours w/ Euclidean MLP	Painting	49.67	43.43	27.21	28.95	–	41.79	38.21
Vanilla CLIP	Clipart	52.28	52.33	30.99	26.13	39.71	–	40.29
Ours w/ non-Euclidean MLP	Clipart	**57.79**	**56.28**	**34.60**	**36.43**	**46.19**	**–**	**46.26**
Ours w/ Euclidean MLP	Clipart	55.08	53.67	33.31	31.62	44.11	–	43.56