UCDR-Adapter: Exploring Adaptation of Pre-Trained Vision-Language Models for Universal Cross-Domain Retrieval

Haoyu Jiang[1*] Zhi-Qi Cheng[2†] Gabriel Moreira[2] Jiawen Zhu[3]
Jingdong Sun[2] Bukun Ren[2] Jun-Yan He[4] Qi Dai[5] Xian-Sheng Hua[1]
[1]Zhejiang University [2]Carnegie Mellon University
[3]Dalian University of Technology [4]DAMO Academy, Alibaba Group [5]Microsoft Research
{zhiqic, gmoreira, jingdons}@cs.cmu.edu, Jiawen@mail.dlut.edu.cn, gid@microsoft.com,
{jianghaoyu0608, bukunren46, junyanhe1989, huaxiansheng}@gmail.com

Abstract

Universal Cross-Domain Retrieval (UCDR) retrieves relevant images from unseen domains and classes without semantic labels, ensuring robust generalization. Existing methods commonly employ prompt tuning with pre-trained vision-language models but are inherently limited by static prompts, reducing adaptability. We propose UCDR-Adapter*, which enhances pre-trained models with adapters and dynamic prompt generation through a two-phase training strategy. First,* Source Adapter Learning *integrates class semantics with domain-specific visual knowledge using a* Learnable Textual Semantic Template *and optimizes* Class and Domain Prompts *via momentum updates and dual loss functions for robust alignment. Second,* Target Prompt Generation *creates dynamic prompts by attending to masked source prompts, enabling seamless adaptation to unseen domains and classes. Unlike prior approaches, UCDR-Adapter dynamically adapts to evolving data distributions, enhancing both flexibility and generalization. During inference, only the image branch and generated prompts are used, eliminating reliance on textual inputs for highly efficient retrieval. Extensive benchmark experiments show that UCDR-Adapter consistently outperforms ProS in most cases and other state-of-the-art methods on UCDR, U^{c}CDR, and U^{d}CDR settings.*[1]

1. Introduction

Universal Cross-Domain Retrieval (UCDR) [26] retrieves relevant images from unseen domains and classes without relying on semantic labels, ensuring robust generalization across diverse and dynamic scenarios. It addresses the challenge of mismatched training and testing data distributions [34], common in real-world deployments where new domains emerge during inference. Achieving effective UCDR requires learning image representations that are both domain-agnostic and semantically discriminative [25]. Balancing these objectives is critical for enabling models to generalize to unseen domains and categories.

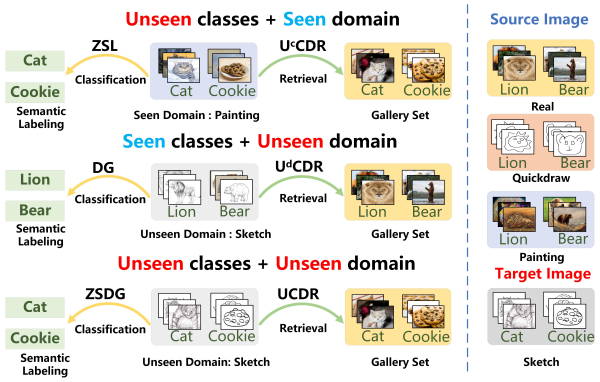

Figure 1. Overview of UCDR settings. Training involves seen categories (e.g., Lion) and domains (e.g., Real). Testing includes unseen domains (e.g., Sketch) and categories (e.g., Cookie) using U^{d}CDR and U^{c}CDR principles. Unlike Zero-Shot Domain Generalization (ZSDG) [2, 24, 25], UCDR does not rely on true labels for unseen data, aligning better with real-world scenarios.

Recent approaches have leveraged pre-trained vision-language models such as CLIP [30] and BLIP [19] to infuse semantic priors that enhance generalization [27, 32]. The rich world knowledge in these models helps distinguish fine-grained categories critical for retrieval. However, fine-tuning on the target dataset is often ineffective, as performance remains limited by the scarcity and lack of diversity in training data. Such data lacks sufficient coverage of the complex visual world, leading to poor generalization. Effective strategies are needed to fully leverage the knowledge in pre-trained models without being restricted by target training data limitations. This motivates advanced ap-

*Part of this work was completed at Carnegie Mellon University.
†Corresponding Author. Now Asst. Prof., University of Washington.
[1]Project: https://github.com/fine68/UCDR2024.

proaches to adapt pre-trained model priors for the UCDR task while overcoming training data constraints.

To address these limitations, we propose *UCDR-Adapter*, a comprehensive framework that enhances pre-trained models with adapter modules and dynamic prompt generation through a novel two-phase training strategy. In the first phase, *Source Adapter Learning*, we integrate class semantics with domain-specific visual knowledge using a *Learnable Textual Semantic Template*. We further optimize *Class and Domain Prompts* via momentum-based updates and dual loss functions to achieve robust and consistent multimodal alignment. In the second phase, *Target Prompt Generation*, we dynamically generate adapted prompts by attending over masked source prompts, effectively simulating adaptation to unseen domains and classes.

Unlike existing methods such as ProS [8], which rely on static prompts, UCDR-Adapter dynamically adjusts to evolving data distributions, improving both flexibility and generalization. By leveraging dynamic momentum-based updates, our approach effectively captures and adapts to diverse retrieval scenarios, avoiding the constraints of fixed prompt configurations. During inference, only the image branch and generated target prompts are used, eliminating reliance on textual inputs and ensuring efficient retrieval. This design enables the model to fully utilize the rich knowledge encoded in pre-trained models while overcoming the limitations of static prompts and data scarcity.

Our main contributions are summarized as follows:

1. We propose *UCDR-Adapter*, a framework that enhances pre-trained vision-language models with *adapter modules* and *dynamic prompt generation*, adapting model knowledge for universal cross-domain retrieval tasks.
2. We introduce a *two-phase training strategy* that integrates *class semantics* with *domain-specific visual knowledge* and enables dynamic prompt generation for effective adaptation to unseen domains and classes.
3. To achieve *robust multimodal alignment*, we employ *momentum-based updates* and *dual loss functions*, addressing challenges such as data scarcity and limited adaptability to diverse data distributions.
4. Extensive experiments on benchmark datasets show that UCDR-Adapter achieves superior performance over state-of-the-art methods in most scenarios, while maintaining efficiency and offering a more robust, adaptable solution for diverse retrieval challenges.

2. Related Work

UCDR Overview: Universal Cross-Domain Retrieval (UCDR) [26] retrieves relevant images from unseen domains and classes without semantic labels, addressing mismatched training and testing distributions [34]. This challenge is common in real-world deployments where new domains emerge during inference. Effective UCDR requires learning image representations that are both domain-agnostic and semantically discriminative [25], ensuring robust generalization to unseen scenarios.

Table 1. Comparison of ZSL, DG, U^{c}CDR, U^{d}CDR, ZSDG, and UCDR task settings for cross-domain retrieval.

Task	Query (Domain)	Query (Class)	Objective
ZSL [3, 35, 39, 43]	✓	✗	Classify
DG [5, 10, 31, 38]	✗	✓	Classify
ZSDG [2, 24, 25]	✗	✗	Classify
U^{c}CDR [1, 26, 34]	✓	✗	Retrieve
U^{d}CDR [1, 26, 34]	✗	✓	Retrieve
UCDR [1, 26, 34]	✗	✗	Retrieve

Existing UCDR Methods: Various methods address UCDR. SnMpNet [26] uses Mixup [42] to blend training data at image and feature levels to simulate unseen domains and classes but requires extensive hyperparameter tuning [11, 40]. SCNNet [1] combines local feature vectors to characterize unseen data [17], yet both SnMpNet and SCNNet enforce similar representations for dissimilar domains, limiting generalization. Additionally, these methods neglect global structural information in cross-domain alignment, potentially compromising performance. SASA [34] leverages semantic knowledge from a pre-trained visual transformer to guide a student network's fine-tuning but risks overfitting to training data, hindering generalization. Recently, ProS [8] extended prompt tuning to UCDR by introducing Content-aware Dynamic Prompts adapting pre-trained vision-language models like CLIP to new domains and classes. However, the reliance on fixed-distribution data generation prompts constrains its adaptability to varying data distributions, thereby diminishing its performance across diverse scenarios.

Vision and Language Pretraining: Contrastive Vision-Language Pre-training (VLP) [20, 36, 41] enhances visual understanding by linking text and image modalities, enabling robust semantic representations and zero-shot generalization. Early works [6, 21] utilized object detection models aligned with text encoders like BERT [14]. The CLIP architecture [30] further advanced this by training separate visual and text encoders through contrastive learning, achieving strong adaptation to unseen classes.

Adapters and Prompts in VLP: Adapter-based and prompt-tuning methods effectively adapt pre-trained vision-language models to new tasks [9, 37, 45]. Prompt tuning [29] introduces lightweight, task-specific parameters in the input space while keeping the backbone frozen, enhancing downstream performance [18, 22]. Zhou *et al.* [45] developed learnable continuous vectors to model contextual words and obtain task-related adapters. Visual Prompt Tuning (VPT) [13] extended this to vision tasks by introducing visual adapters within the ViT architecture. Recent works like STYLIP [4] and CoCoOp [44], along with others [15, 16], achieve domain generalization through style- and content-

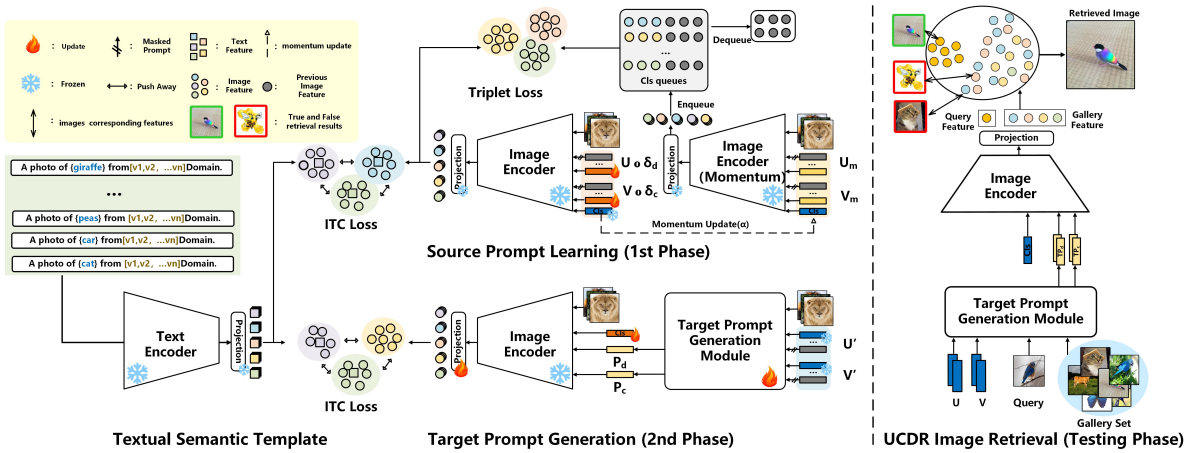

Figure 2. UCDR-Adapter architecture. In Phase 1 (top), *Source Adapter Learning* optimizes class and domain prompts via a momentum encoder and dual loss functions for aligned multimodal representations. Only relevant prompts are activated based on the input image. In Phase 2 (bottom), the *Target Prompt Generation module* generates adapted prompts by attending over masked source prompts, simulating adaptation to unseen domains and classes. At test time (right), only the image branch is utilized with the generated target prompts for effective retrieval without textual cues.

conditional prompt learning. However, these approaches lack class-specific specialization, leading to semantic ambiguities and inadequate handling of domain gaps.

3. Problem Definition

The Universal Class-Domain Retrieval (UCDR) task aims to retrieve images from new domains and categories not seen during training. It consists of two key phases:

Training on *Seen* Domains & Classes: In the training phase, we have a set of domains $\mathcal{D}_{tr}$ with $|\mathcal{D}_{tr}|$ domains. Within each domain $d \in \mathcal{D}_{tr}$, there are N_d image samples. We also have a discrete set of classes $\mathcal{C}_{tr}$ with $|\mathcal{C}_{tr}|$ categories. Typically, the training data $\mathcal{S}_{tr}$ contains images across these domains and classes. The training data $\mathcal{S}_{tr}$ is $\{(\text{IMG}_i^{c,d}, \text{TEP}_i^{c,d}, \text{CLS}_i^{c,d})\}_{i=1}^{N_d}$, where i ranges from 1 to N_d, domain $d \in \mathcal{D}_{tr}$ and class $c \in \mathcal{C}_{tr}$. Here, $\text{IMG}_i^{c,d}$ is i^{th} image in domain d of class c, $\text{TEP}_i^{c,d}$ is the corresponding text template, and $\text{CLS}_i^{c,d}$ is the category label, respectively.

Testing on *Unseen* Domains & Classes: At test time, the model sees new queries from a test domain set $\mathcal{D}_{te}$ not encountered during training. The test set $\mathcal{S}_{te}$ contains N_t tuples $\{(\text{IMG}_i^{c,d}, \text{CLS}_i^{c,d})\}_{i=1}^{N_t}$, where $c \in \mathcal{C}_{te}$ and $d \in \mathcal{D}_{te}$. The key challenge is to retrieve images from a gallery matching the query image categories.

Particularly, the UCDR task has two scenarios:

1. **U$^{\text{d}}$CDR**: The test domain $d \in \mathcal{D}_{te}$ is novel ($\mathcal{D}_{te} \notin \mathcal{D}_{tr}$), but the classes are seen in training ($\mathcal{C}_{tr} = \mathcal{C}_{te}$).

2. **U$^{\text{c}}$CDR**: The test domain is seen ($d \in \mathcal{D}_{tr}$), but the classes $\mathcal{C}_{te}$ are new ($\mathcal{C}_{te} \cap \mathcal{C}_{tr} = \varnothing$).

The key challenge is when both domains and classes are new, testing the limits of retrieval generalization.

4. Proposed UCDR-Adapter

The UCDR-Adapter enables universal cross-domain image retrieval through a three-phase approach utilizing prompt optimization: 1) In the Source Prompt Learning phase (Sec. 4.1), class- and domain-specific prompts are optimized via dual losses to align multimodal representations. This equips the model with source knowledge. 2) The Target Prompt Generation phase (Sec. 4.2) produces adapted prompts by attending over masked source prompts. This simulates generalization to new domains and classes. 3) At test time (Sec. 4.3), only the image branch is used with generated target prompts for retrieval without textual cues. By unifying strategic prompt optimization and adaptive prompt generation, UCDR-Adapter provides an efficient and effective approach for generalized UCDR.

4.1. Source Prompt Learning (1st Phase)

The first phase is designed to optimize class- and domain-specific prompts within the image encoder. This enables focused representation learning on the seen classes and domains encountered during training. The key innovations in the first phase are listed as:

Textual Semantic Template: We propose a learnable semantic template to enhance cross-domain generalization, as shown in Figure 2. The template is inspired by Zhou *et al.* [45] and formulated as '*A photo of* $\underline{c_i}$ *from the* $\underline{v_1, \ldots, v_N}$ *domain*', consisting of:

1. **Class Strings** c_i**:** Precisely defined strings representing each class c_i (e.g. 'dog' for a image from the dog class). This integrates class semantics.

2. **Domain Vectors** v_j**:** N trainable d-dimensional vec-

tors that capture nuanced visual characteristics of each domain. They are initialized randomly and optimized during training to represent domain-specific details (e.g. textures for the real domain).

Generally, image features from the encoder are projected into the v_j space and used to update the vectors via cross-modal alignment (as Eq. 3). This guides them to capture domain visual details. Integrating v_j into the template input enables the text encoder to generate domain-aware embeddings, improving generalization. By jointly modeling class semantics and domain visual knowledge, the template equips the model with top-down and bottom-up representations. This unified dual knowledge enhances cross-domain generalization, allowing transfer to new classes.

Domain & Class Prompts: We introduce learnable domain and class-specific prompts to incorporate domain and class semantics into the image encoder:

1. **Domain Prompts** $U \in \mathbb{R}^{D_{tr} \times m}$: U is a learnable prompt matrix with $|\mathcal{D}_{tr}|$ rows corresponding to the number of training domains, and m columns representing the prompt dimensionality.

2. **Class Prompts** $V \in \mathbb{R}^{C_{tr} \times m}$: V is a learnable prompt matrix with $|\mathcal{C}_{tr}|$ rows corresponding to the number of training classes, and m columns representing the prompt dimensionality.

Particularly, for an image $\text{IMG}_i^{c,d}$ belonging to class c and domain d, the masks δ_c and δ_d are applied to V and U respectively to extract the c^{th} and d^{th} rows. The masked row vectors are concatenated channel-wise into a single prompt vector $p = [V \circ \delta_c; U \circ \delta_d]$. This prompt vector p is then projected to the input dimensionality of the vision transformer encoder and add to the input image features.

This allows the vision transformer to adapt its representations specifically based on the domain and class via the prompt, enhancing domain and class discrimination. The prompts are optimized via backpropagation along with the vision transformer encoder parameters to learn effective domain- and class-specific representations. By learning dedicated prompts for each domain and class, the model gains specialized representation capabilities targeted for cross-domain recognition.

Momentum Updated Prompts: To improve sample diversity and representation learning of the class-specific prompts, we incorporate momentum updated prompts inspired by MoCo [12]: the queue acts as a large dictionary, whose size is decoupled from the batch size, allowing it to be large. Given more samples, we ensure a better coverage of the data distribution, hence the sample diversity argument. Specifically, the momentum prompts U_m, V_m are separate matrices for domain and class that are updated with momentum using the current prompt weights U, V as follows:

$$\theta_M(t+1) = (1-\alpha)\theta_M(t) + \alpha\theta_C(t), \tag{1}$$

where θ_M, θ_C are parameters of momentum (U_m, V_m) and current prompts (U, V). α is the momentum update rate, set to 0.001. And t represents the number of iteration rounds. Image features which extracted using U_m, V_m, are stored in class-specific queues $\mathcal{Q} = \{q_i\}_{i=1}^{|\mathcal{C}_{tr}|}$ (see Figure 2).

For image $\text{IMG}_i^{c,d}$, positive I_p and negative I_n pairs are sampled from positive queue q_c and negative queues q_{-c}. With the image features I_f^c, extracted using p. And r is the number of ternary loss pairs matched for $\text{IMG}_i^{c,d}$, and b is margin, we set it to 0.5. The triplet loss is:

$$\mathcal{L}_{\text{Triplet}} = \frac{1}{r}\sum_{i=1}^{r} \max(0, \|I_f^c - I_p\|^2 - \|I_f^c - I_n\|^2 + b), \tag{2}$$

where the updated prompts increase the diversity of class representations. In short, the storing momentum features in class queues allows the sampling of hard positives and negatives for robust learning. $\mathcal{L}_{\text{Triplet}}$ loss on these samples enhances inter-class discrepancy and intra-class similarity.

Cross-Modal Alignment: Cross-modal alignment is used to align image and text representations. This process helps to improve the domain invariance of image features and optimize the prompts. We use a contrastive loss to align the encoded image features (I_f^c) for class c with the encoded text features (T_f^c) from the semantic template. The $\mathcal{L}_{\text{ITC}}$ loss is defined as:

$$\mathcal{L}_{\text{ITC}} = -\log \frac{\exp(s(I_f^c, T_f^c)/\tau)}{\sum_{k=1}^{|\mathcal{C}_{tr}|} \exp(s(I_f^c, T_f^k)/\tau)}, \tag{3}$$

where s represents cosine similarity and τ is temperature, set to 0.07. The $\mathcal{L}_{\text{ITC}}$ function brings the matched image and text pairs closer and pushes non-matched pairs apart.

This alignment between image and text features helps the model learn domain-agnostic representations. By aligning image and text features across domains, the model learns domain-agnostic representations. The $\mathcal{L}_{\text{ITC}}$ loss finetunes the prompts and semantic templates to align with cross-modal semantics.

$$\mathcal{L}_{\text{Phase1}} = \mathcal{L}_{\text{Triplet}} + \mathcal{L}_{\text{ITC}}, \tag{4}$$

the overall loss in Phase1($\mathcal{L}_{\text{Phase1}}$) combines $\mathcal{L}_{\text{ITC}}$ and $\mathcal{L}_{\text{Triplet}}$. In summary, cross-modal alignment provides dual-supervision from both visual and semantic modalities to improve domain invariance and prompt optimization.

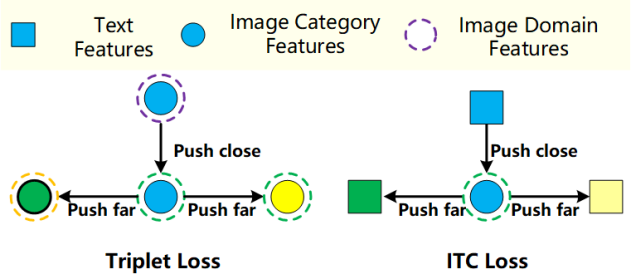

Figure 3. Phase 1 loss function. Right: ITC loss. Left: Triplet Loss. Solid circles are image features, boxes represent the semantic features of each class, and the dashed circles correspond to the image domain.

4.2. Target Prompt Generation (2nd Phase)

In the second phase, we simulate a scenario where the class and domain of an image are unknown, to generate target prompts to achieve generalization to new domains and classes. As illustrated in Figure 4, the Target Prompt Generation (TPG) module leverages the learned prompts from phase 1. We mask out the rows of U, V corresponding to the known domains and classes using $1 - \delta_c$ and $1 - \delta_d$:

$$U' = U \circ (1 - \delta_d), \tag{5}$$
$$V' = V \circ (1 - \delta_c), \tag{6}$$

this leaves us with prompts with masked-out familiar domains (U') and classes (V').

Particularly, the image $\text{IMG}_i^{c,d}$ is passed through the Feature Encoder $g()$ to extract features $I_g \in \mathbb{R}^D$. And Attn() represent Soft-Attention. We compute attention weights over the masked prompts:

$$w_d = \text{Attn}(I_g, U'), \tag{7}$$
$$w_c = \text{Attn}(I_g, V'), \tag{8}$$

this aligns the image features with the unfamiliar prompt rows. The attention weights (w_d and w_c) are used to compute target prompts as weighted combinations of the masked rows as $P_d = V'w_d$ and $P_c = U'w_c$. Here the target prompts are concatenated and projected to the encoder input dimension, then add to the input image features for enhanced feature extraction. The extracted features I_f^c are aligned with T_f^c from phase 1 via $\mathcal{L}_{\text{ITC}}$ to learn generalized representations. By generating target prompts via attention over masked prompts, the model learns to produce prompts for new domains and classes.

4.3. UCDR Image Retrieval (Testing Phase)

In the testing phase for UCDR task, textual semantic information is unavailable for new classes. Thus, we discard the textual components and only utilize the image branch. The key components are:

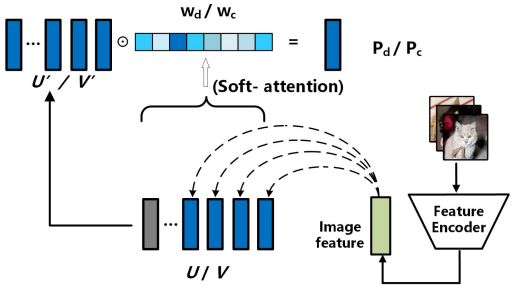

Figure 4. Target Prompt Generation process. Gray indicates that the class and domain prompts are masked. Where P_d and P_c are target prompts generated for unseen domains and classes.

1. **Domain & Class Prompts** from phase 1 - the domain and class prompts for seen classes/domains.

2. **Target Prompt Generation Module** from phase 2 to generate adapted prompts.

3. **Frozen Image Encoder** from the pre-trained CLIP model does not require any extra modifications.

For a query image and each gallery image, the TPG module is applied to create adapted target prompts based on the image content. These target prompts are input to the image encoder to extract enhanced representations for each image. The query embedding is then compared to the gallery embeddings via Euclidean distance, and the gallery images are ranked based on distance to retrieve the final results. The same process is applied for the U^{c}CDR and U^{d}CDR sub-tasks without modification, utilizing the TPG capability to generalize. By retaining only the crucial image branch with the prompt generation, the model is able to perform retrieval for new classes and domains at test time without textual cues.

5. Experiments

5.1. Experimental Setup

Datasets: Following the benchmarks in [25, 26, 34], we evaluate the performance of the UCDR-Adapter model on three widely used datasets (DomainNet [28], Sketchy [23, 33], TU-Berlin [7, 23]) for three retrieval tasks, namely UCDR, U^{d}CDR and U^{c}CDR. DomainNet contains 596,006 images from 345 classes and 6 domains (real, Sketch, Quickdraw, Infograph, Clipart, Painting) for U^{d}CDR and UCDR experiments. and used 245 classes as a training set, 55 classes as a validation set, and 45 classes as a test set as previously set [26]. In training, we use images from five domains, to satisfy the unseen setting. Considering that the training data for the CLIP model consists of image-text pairs downloaded from the web, we do not use the Real as the unseen domain. In addition, we have two search settings: 1) Use an unseen class image belonging to the real

domain as the gallery set; 2) Considering more realistic scenarios, the gallery set is a combination of visible and invisible classes from the real domain. Sketchy contains 75,471 sketches and 73,002 images from 125 categories. Based on the settings in [25, 26, 34], the training set, validation set, and test set consist of 93, 11, and 21 categories, respectively. Sketchy and TU-Berlin are both used for U^{c}CDR.

Evaluation Metrics: To ensure a fair comparison, we use the same assessment metrics as [25, 26, 34]. For Sketchy and DomainNet, we evaluate the top 200 candidates' precision and mean average precision scores (Prec@200 and mAP@200). For TU-Berlin, we used Prec@100 and mAP@all as evaluation metrics.

Implementation Details: We implement UCDR-Adapter in PyTorch and train on a single Nvidia A100 GPU. ViT-B/32 is used for both image and text encoders. The Domain Prompts (U) and Class Prompts (V) are 768-d vectors. The momentum encoder has a momentum update rate of 1e-3 and a queue length of 20 per class. Training runs for 50 epochs with early stopping if validation accuracy does not improve for 2 epochs. Stage 1 uses a batch size of 400, and stage 2 uses 50. Adam optimization is used with initial LR 1e-3 decayed to 1e-6 over 20 epochs. For Zero-Shot CLIP, we use pre-trained VIT-B/32 Image Encoder for retrieval.

5.2. Comparison with State-of-the-art

UCDR Results: We evaluate UCDR-Adapter on UCDR using DomainNet. Following the protocol in [26], we select unseen samples from five domains for training, and use one holdout domain for querying unseen classes. We consider two galleries: 1) unseen classes only; 2) both seen and unseen classes. The results are in Table 2. UCDR-Adapter improves the mAP@200 over SASA [34] by +16.84% and +17.2% on the two galleries. It also outperforms Zero-Shot-CLIP, which is comparable to SASA but degrades on less detailed domains. Compared to Zero-Shot-CLIP, UCDR-Adapter achieves gains of +16.36% and +19.23% respectively. This shows our training strategy enhances generalization over the pre-trained features for unseen classes and domains. The consistent improvements show UCDR-Adapter's effectiveness on the UCDR task by adapting the pre-trained knowledge through prompt optimization.

U^{d}CDR Results: We evaluate UCDR-Adapter's domain generalization ability on U^{d}CDR using DomainNet. As shown in Table 3, we select 25% queries per seen class from the unseen domain (10% for Quickdraw given its large size). The gallery contains seen class images from the Real domain. UCDR-Adapter improves mAP@200 over SASA by +19.08%, indicating our domain-level adapter fine-tuning allows CLIP to focus on class-discriminative features, mitigating the impact of domain shift. The boost shows UCDR-Adapter's ability to learn domain-invariant representations, crucial for generalization in U^{d}CDR.

U^{c}CDR Results: We evaluate UCDR-Adapter's generalization to unseen classes on U^{c}CDR using Sketchy and TU-Berlin. As shown in Table 4, queries are unseen classes from the seen training domain, while the gallery contains all classes. On Sketchy, UCDR-Adapter improves mAP@200 and prec@200 over SASA by +3.75% and +7.98% respectively. On TU-Berlin, it achieves gains of +18.66% in mAP@200 and +6.35% in Allprec@100. This demonstrates the TPG module can effectively leverage acquired knowledge to infer representations of unseen classes and domains. These consistent improvements verify UCDR-Adapter's ability to generalize to new categories, which is crucial for the U^{c}CDR task.

5.3. Ablation Studies

We conduct ablation experiments on Sketchy as the unseen domain, with unseen classes for both queries and gallery. The results are in Table 5. A comparison of training parameters is shown in Table 6.

Fine-tuning vs Adapters: Fine-tuning the entire CLIP model [30] is computationally expensive as it requires updating the whole pre-trained network. More critically, fine-tuning often leads to catastrophic forgetting, degrading the original capabilities of CLIP. An alternative is a linear probe, where only a classifier layer is trained on top of frozen CLIP features. As shown in Table 5, the linear probe improves over zero-shot CLIP but underperforms our proposed adapter approach, which achieves a significant gain of +5.69% in mAP. This verifies that adapters can efficiently tune CLIP for downstream tasks through the targeted updates of small adapter modules, without interfering with the frozen pre-trained weights. Adapters provide an effective balance between transferability and adaptation. Our designed prompts and training strategy take advantage of this to unlock CLIP's knowledge for UCDR.

Two Phases vs One Phase: Comparing adapters trained end-to-end versus generated by the TPG module, TPG improves mAP by +6.82%. This validates our two-stage approach, where TPG harnesses knowledge from seen classes effectively through masking.

Learnable Semantic Template: Replacing the fixed CLIP text encoding with our learnable semantic template improves mAP by +2.87%, confirming the benefits of optimizing domain information into category supervision.

Momentum & Triplet Loss: The momentum encoder and triplet loss help enhance sample diversity and representation learning of class/domain knowledge in the adapters. The momentum encoder maintains a diverse set of class-specific samples in queues for mining hard positives/negatives. Combined with the triplet loss, this strengthens inter-class discrepancy and intra-class similarity in adapter representations. When integrated in phase 2 training, the improved adapters lead to better generalization. Our full UCDR-

Table 2. Evaluation results on DomainNet for UCDR using two different gallery settings: (1) Only unseen class samples from the held-out domain; (2) Both seen and unseen class samples from the held-out domain. The results demonstrate UCDR-Adapter's effectiveness in retrieving unseen classes under domain shift.

Training Domains	Query Domain	Method	Unseen Class Gallery		Seen+Unseen Class Gallery	
			mAP@200	Prec@200	mAP@200	Prec@200
Real, Quickdraw, Infograph, Painting, Clipart	Sketch	SnMpNet [ICCV 2021] [26]	0.3007	0.2432	0.2624	0.2134
		SCNNet [WACV 2023] [1]	0.4075	0.4120	0.3422	0.2534
		SASA [SIGIR 2022] [34]	0.5262	0.4468	0.4732	0.4025
		Zero-Shot CLIP	0.4222	0.3529	0.3760	0.3081
		ProS [CVPR 2024] [8]	0.6457	0.6001	0.5843	0.5463
		UCDR-Adapter(Ours)	**0.6591**	**0.6142**	**0.6073**	**0.5707**
Real, Sketch, Infograph, Painting, Clipart	Quickdraw	SnMpNet [ICCV 2021] [26]	0.1736	0.1284	0.1512	0.1111
		SCNNet [WACV 2023] [1]	0.1998	0.1580	0.1698	0.1411
		SASA [SIGIR 2022] [34]	0.2564	0.1970	0.2116	0.1651
		Zero-Shot CLIP	0.0744	0.0561	0.0607	0.0386
		ProS [CVPR 2024] [8]	**0.2842**	**0.2544**	**0.2318**	0.2127
		UCDR-Adapter(Ours)	0.2794	0.2534	0.2317	**0.2154**
Real, Sketch, Quickdraw, Infograph, Clipart	Painting	SnMpNet [ICCV 2021] [26]	0.4031	0.3332	0.3635	0.3019
		SCNNet [WACV 2023] [1]	0.4242	0.4409	0.3731	0.3964
		SASA [SIGIR 2022] [34]	0.5898	0.5188	0.5463	0.4804
		Zero-Shot CLIP	0.6169	0.5508	0.5755	0.5085
		ProS [CVPR 2024] [8]	0.7516	0.6955	0.7120	0.6612
		UCDR-Adapter(Ours)	**0.7538**	**0.6974**	**0.7203**	**0.6693**
Real, Sketch, Quickdraw, Painting, Clipart	Infograph	SnMpNet [ICCV 2021] [26]	0.2079	0.1717	0.1800	0.1496
		SCNNet [WACV 2023] [1]	0.2737	0.2476	0.2369	0.1983
		SASA [SIGIR 2022] [34]	0.2823	0.2425	0.2491	0.2113
		Zero-Shot CLIP	0.5007	0.4474	0.4501	0.3990
		ProS [CVPR 2024] [8]	**0.5798**	**0.5442**	0.5219	0.4956
		UCDR-Adapter(Ours)	0.5714	0.5364	**0.5315**	**0.5022**
Real, Sketch, Quickdraw, Infograph, Painting	Clipart	SnMpNet [ICCV 2021] [26]	0.4198	0.3323	0.3765	0.2959
		SCNNet [WACV 2023] [1]	0.4843	0.4664	0.4322	0.4016
		SASA [SIGIR 2022] [34]	0.5392	0.4300	0.4902	0.3886
		Zero-Shot CLIP	0.6037	0.5131	0.5700	0.4768
		ProS [CVPR 2024] [8]	0.7648	0.7186	0.7228	0.6815
		UCDR-Adapter(Ours)	**0.7718**	**0.7263**	**0.7391**	**0.6979**
Average		SnMpNet [ICCV 2021] [26]	0.3010	0.2418	0.2667	0.2144
		SCNNet [WACV 2023] [1]	0.3579	0.3449	0.3108	0.2981
		SASA [SIGIR 2022] [34]	0.4387	0.3670	0.3940	0.3295
		Zero-Shot CLIP	0.4435	0.3840	0.3737	0.3462
		ProS [CVPR 2024] [8]	0.6052	0.5626	0.5546	0.5195
		UCDR-Adapter(Ours)	**0.6071**	**0.5655**	**0.5660**	**0.5311**

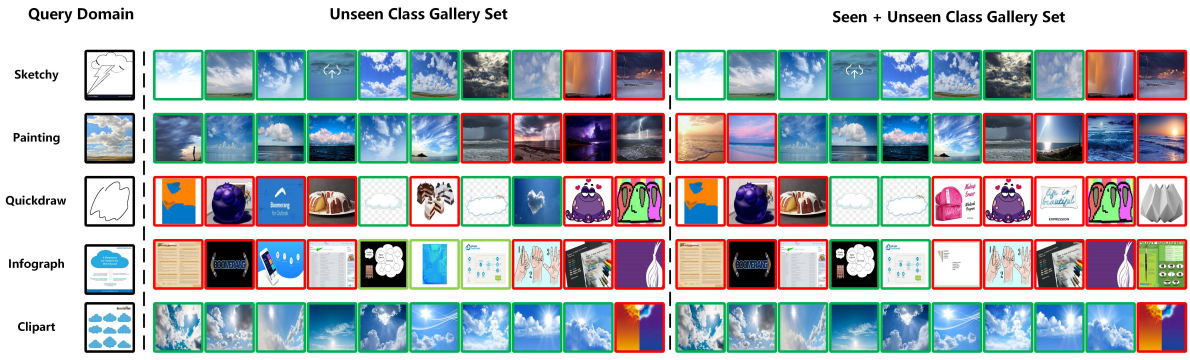

Figure 5. The results of using UCDR-Adapter for the UCDR on DomainNet. The 'Cloud' class from the holdout domain as the query.

Adapter achieves significant gains over zero-shot CLIP while only adding minimal parameters in the adapters.

Retrieval Qualitative Analysis: Figure 5 shows the top 10 retrieval results for a 'cloud' query from the Sketchy, Painting, Clipart, and Quickdraw domains under the U^{c}CDR set. For domains with more distinctive features like Sketchy,

Table 3. U^{d}CDR evaluation results on DomainNet, where queries from holdout domain during training.

Query Domain	Methods	mAP@200	Prec@200
Sketch	SnMpNet [26]	0.3529	0.1657
	SASA [34]	0.5733	0.5290
	ProS [8]	**0.7385**	**0.4911**
	UCDR-Adapter(Ours)	0.7332	0.4893
Quickdraw	SnMpNet [26]	0.1077	0.0509
	SASA [34]	0.1805	0.1549
	ProS [8]	0.2889	**0.1186**
	UCDR-Adapter(Ours)	**0.2900**	0.1181
Painting	SnMpNet [26]	0.4808	0.4424
	SASA [34]	0.5596	0.5178
	ProS [8]	0.7227	0.4615
	UCDR-Adapter(Ours)	**0.7306**	**0.4634**
Infograph	SnMpNet [26]	0.1957	0.1764
	SASA [34]	0.2340	0.2093
	ProS [8]	0.6056	**0.3962**
	UCDR-Adapter(Ours)	**0.6064**	0.3922
Clipart	SnMpNet [26]	0.5520	0.5074
	SASA [34]	0.6840	0.6361
	ProS [8]	0.8105	0.5298
	UCDR-Adapter(Ours)	**0.8251**	**0.5392**
Average	SnMpNet [26]	0.3378	0.2685
	SASA [34]	0.4462	0.4094
	ProS [8]	0.6332	0.3994
	UCDR-Adapter(Ours)	**0.6370**	**0.4004**

Table 4. U^{c}CDR evaluation results on Sketchy and TU-Berlin.

Method	Sketchy		TU-Berlin	
	mAP@200	prec@200	mAP@All	prec@100
SnMpNet [26]	0.5781	0.5155	0.3568	0.5226
SASA [34]	0.6910	0.6090	0.4715	0.6682
ProS [8]	0.6991	0.6545	**0.6675**	**0.7442**
UCDR-Adapter(Ours)	**0.7285**	**0.6888**	0.6581	0.7317

Table 5. UCDR ablation experiments, with Sketchy as the holdout domain. $\text{Prompt}_{\text{vision}}$ and $\text{Prompt}_{\text{text}}$ are image and text prompts, respectively. U and V are domain and category prompts. Mask uses two-phases masking operation, and TST is a training Textual Semantic Template. $\mathcal{L}_{\text{Triple}}$ (pair = i) is the use of i ternary loss pairs.

Tuning Phases	Training Task	mAP@200	Prec@200
One Phase	Fine-Tuning CLIP	0.2674	0.2063
	Zero-Shot CLIP	0.4222	0.3529
	Linear Probing(LP) CLIP	0.4936	0.4365
	LP + $\text{Prompt}_{\text{vision}}$	0.5505	0.4917
	LP + $\text{Prompt}_{\text{vision}}$ + $\text{Prompt}_{\text{text}}$	0.5763	0.5202
Two Phases	U + V + TPG	0.6187	0.5753
	U + V + TPG + Mask	0.6253	0.5807
	Full (U + V + TPG + Mask + TST)	0.6540	0.6062
	Full + $\mathcal{L}_{\text{Triple}}$ (pair 1)	0.6560	0.6107
	Full UCDR-Adapter (Ours)	**0.6591**	**0.6142**

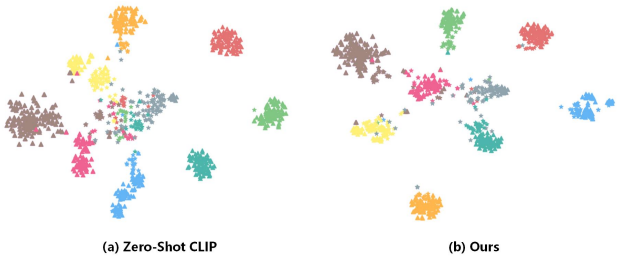

Figure 6. t-SNE visualization of features for 10 random unseen classes from Real (stars) and Sketchy (triangles) domains. Sketchy was used as the unseen holdout domain during training. Colors indicate different classes. (a) CLIP features show entanglement between some classes. (b) Our adapted model shows improved inter-class separation and intra-class clustering, indicating enhanced discrimination of unseen classes.

Table 6. Comparison of the number of parameters and mean average precision for the three tuning CLIP methods.

Methods	Training Parameter(M)	mAP@200
Fine-Tuning (Image encoder)	87.84	0.2674
Linear Probing	0.39	0.4936
UCDR-Adapter(Ours)	**2.36**	**0.6591**

Painting, and Clipart, the model retrieves fewer incorrect candidates compared to Quickdraw which lacks detail. In some cases, errors occur due to background similarities, as with the incorrect lightning' retrieval.

We also analyze the feature distributions of CLIP and our adapted model in Figure 6. Using 10 random unseen classes from Real (seen) and Sketchy (unseen) domains, CLIP fails to differentiate some classes in Figure 6(a). In contrast, our adapted CLIP in Figure 6(b) shows improved intra-class clustering and inter-class separation. This indicates our approach enhances CLIP's discriminative power while retaining its representation capabilities, crucial for the differentiation of unseen classes.

6. Conclusion

We introduce *UCDR-Adapter* for Universal Cross-Domain Retrieval (UCDR), enhancing pre-trained models through adapter-based prompt optimization and a two-phase training strategy. First, class- and domain-specific prompts are optimized with momentum encoders and dual loss functions for aligned multimodal representations. Then, target prompts are generated by attending to masked source prompts, facilitating adaptation to unseen domains and classes. Experiments demonstrate that UCDR-Adapter outperforms state-of-the-art methods on UCDR and its subtasks, ensuring effective generalization. This framework advances prompt-based tuning for UCDR and extends to other cross-domain tasks, offering an efficient solution for generalized visual retrieval.

Acknowledgments

This work was partially supported by the Air Force Research Laboratory under agreement number FA8750-19-2-0200, and by grants from the Defense Advanced Research Projects Agency (DARPA) under the GAILA program (award HR00111990063) and the AIDA program (FA8750-18-20018). Zhi-Qi Cheng also acknowledges support from the University of Washington startup fund, the Intel Ph.D. Fellowship, and the IBM Outstanding Student Scholarship. Portions of this research were funded by these grants. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions expressed in this work are those of the authors and do not necessarily reflect the official policies or endorsements of the Air Force Research Laboratory, DARPA, or the U.S. Government.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, Balaji Vasan Srinivasan, and Biplab Banerjee. Contrastive learning of semantic concepts for open-set cross-domain retrieval. In *WACV*, pages 4115–4124, 2023. 2, 7

[2] Ahmad Arfeen, Titir Dutta, and Soma Biswas. Handling class-imbalance for improved zero-shot domain generalization. In *BMVC*, page 728, 2022. 1, 2

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *NeurIPS*, 24, 2011. 2

[4] Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *WACV*, pages 5542–5552, 2024. 2

[5] Weipeng Cao, Yuhao Wu, Yixuan Sun, Haigang Zhang, Jin Ren, Dujuan Gu, and Xingkai Wang. A review on multimodal zero-shot learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1488, 2023. 2

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020. 2

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

[8] Kaipeng Fang, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Zhi-Qi Cheng, Xiyao Li, and Heng Tao Shen. Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17292–17301, 2024. 2, 7, 8

[9] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[10] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *WACV*, pages 434–443, 2023. 2

[11] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *AAAI*, pages 3714–3722, 2019. 2

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 4

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

[14] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 2

[15] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2

[16] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 2

[17] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, pages 2668–2677, 2018. 2

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, 2021. 2

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 1

[20] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, volume 34, pages 9694–9705, 2021. 2

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137, 2020. 2

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, pages 4582–4597, 2021. 2

[23] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, pages 2862–2871, 2017. 5

[24] Puneet Mangla, Shivam Chandhok, Vineeth N Balasubramanian, and Fahad Shahbaz Khan. Cocoa: Context-conditional adaptation for recognizing unseen classes in unseen domains. In *WACV*, pages 865–874, 2022. 1, 2

[25] Biswajit Mondal and Soma Biswas. Seic: Semantic embedding with intermediate classes for zero-shot domain generalization. In *ACCV*, pages 789–806, 2022. 1, 2, 5, 6

[26] Soumava Paul, Titir Dutta, and Soma Biswas. Universal cross-domain retrieval: Generalizing across classes and domains. In *ICCV*, pages 12056–12064, 2021. 1, 2, 5, 6, 7, 8

[27] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *CVPR*, pages 18983–18992, 2023. 1

[28] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019. 5

[29] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *EMNLP*, pages 2463–2473, 2019. 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2, 6

[31] Wenqi Ren, Yang Tang, Qiyu Sun, Chaoqiang Zhao, and Qing-Long Han. Visual semantic segmentation based on few/zero-shot learning: An overview. *IEEE/CAA Journal of Automatica Sinica*, 2023. 2

[32] Aneeshan Sain, Ayan Kumar Bhun0ia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *CVPR*, pages 2765–2775, 2023. 1

[33] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *Transactions on Graphics*, 35(4):1–12, 2016. 5

[34] 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. 1, 2, 5, 6, 7, 8

[35] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

[36] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186, 2023. 2

[37] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *CVPR*, pages 19265–19274, 2023. 2

[38] Guo-Sen Xie, Zheng Zhang, Huan Xiong, Ling Shao, and Xuelong Li. Towards zero-shot learning: A brief review and an attention-based embedding network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1181–1197, 2023. 2

[39] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021. 2

[40] Hao Yu, Huanyu Wang, and Jianxin Wu. Mixup without hesitation. In *ICIG*, pages 143–154, 2021. 2

[41] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. In *TMLR*, 2022. 2

[42] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[43] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(04):4396–4415, 2023. 2

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 2

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 3