CLAIR: CLIP-Aided Weakly Supervised Zero-Shot Cross-Domain Image Retrieval

Chor Boon Tan
tan.chor.boon@u.nus.edu
Conghui Hu
conghui@nus.edu.sg
Gim Hee Lee
gimhee.lee@nus.edu.sg

School of Computing
National University of Singapore

Abstract

The recent growth of large foundation models that can easily generate pseudo-labels for huge quantity of unlabeled data makes unsupervised Zero-Shot Cross-Domain Image Retrieval (UZS-CDIR) less relevant. In this paper, we therefore turn our attention to weakly supervised ZS-CDIR (WSZS-CDIR) with noisy pseudo labels generated by large foundation models such as CLIP. To this end, we propose CLAIR to refine the noisy pseudo-labels with a confidence score from the similarity between the CLIP text and image features. Furthermore, we design inter-instance and inter-cluster contrastive losses to encode images into a class-aware latent space, and an inter-domain contrastive loss to alleviate domain discrepancies. We also learn a novel cross-domain mapping function in closed-form, using only CLIP text embeddings to project image features from one domain to another, thereby further aligning the image features for retrieval. Finally, we enhance the zero-shot generalization ability of our CLAIR to handle novel categories by introducing an extra set of learnable prompts. Extensive experiments are carried out using TUBerlin, Sketchy, Quickdraw, and DomainNet zero-shot datasets, where our CLAIR consistently shows superior performance compared to existing state-of-the-art methods. We release our code at this https URL.

1 Introduction

Cross-Domain Image Retrieval (CDIR) is the task of retrieving images from a target domain that are pertinent to a query image from a source domain. The vanilla CDIR setting relies on supervised training with category labels and/or pairing annotations from both domains, limiting its real-world applicability due to high annotation costs. **Zero-shot CDIR** [16, 23] addresses this by training on a closed set of categories and generalizing to unseen ones at test time. However, it still necessitates an annotated training set. **Unsupervised CDIR** [9, 29] aims for label-free domain alignment, but its application is limited to a closed set of seen categories, and is increasingly becoming obsolete with the availability of large foundation models such as CLIP [21] that are capable of generating pseudo labels on large amounts of unlabeled data.

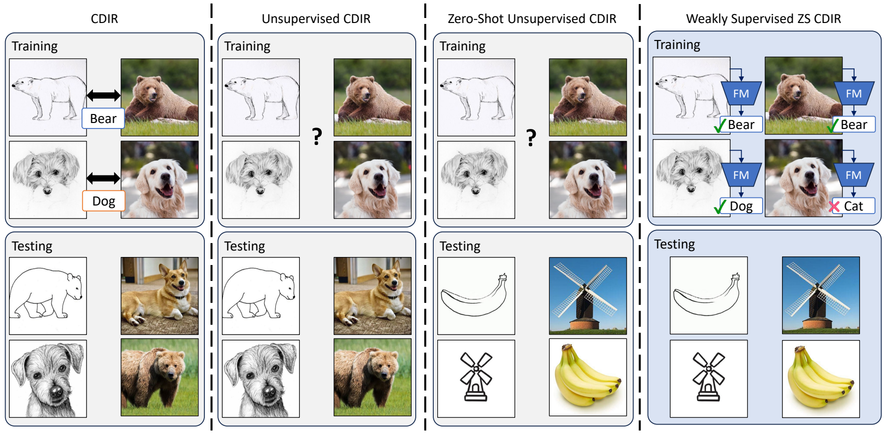

Figure 1: From left to right, **CDIR** assumes both class labels and paired images across domains are provided during training. In **UCDIR**, class labels and paired annotations are no longer available. **ZS-UCDIR** introduces new unseen classes during the test phase. The use of foundation model converts the problem into **WSZS-CDIR**, where the model is trained under weak supervision of noisy pseudo labels.

In view of these limitations, we investigate the problem of Weakly Supervised Zero-Shot CDIR (WSZS-CDIR). As illustrated in Figure 1, this task is a variant of the recently proposed Unsupervised Zero-Shot CDIR (UZS-CDIR) problem setting of image sketches [32] (also known as Zero-Shot Sketch-Based Image Retrieval, *i.e*. ZS-SBIR) with two important differences: 1) We aim to generalize our model across any domain instead of limiting to only sketch image domain. 2) The existence of many large foundation models has significantly reduced the difficulty in obtaining pseudo-labels for training. Consequently, we have converted UZS-CDIR into a WSZS-CDIR problem. Nonetheless, several challenges remain : 1) The pseudo-labels generated from large foundation models tend to be extremely noisy. A naive use of these noisy pseudo-labels to train the model cannot yield good performance. 2) A robust method to extract semantic-aware and domain-invariant features for effective cross-domain alignment is still needed. 3) A simple training on the pseudo-labels of the available data limits the model to the class categories in the training data and does not give the model zero-shot ability to generalize on unseen categories.

We propose CLAIR in this paper, which stands for **CL**IP-**A**ided Weakly Supervised framework for Zero-Shot Cross-Domain **I**mage **R**etrival. Our CLAIR leverages CLIP to generate noisy pseudo-labels of the unlabeled training data from any two domains. We mitigate the noisy pseudo-labels issue by computing the confidence scores of the pseudo-labels assigned to the training images in both query and target domains based on the similarity between their CLIP text and image features. To effectively align distinct image domains into a shared class-aware latent space, we design inter-instance and inter-cluster contrastive losses for encoding semantic features, and an inter-domain contrastive loss to alleviate domain discrepancies. We further improve domain alignment by deriving a closed-form mapping function using easily generated paired cross-domain CLIP text prompts and directly applying it to image features. Lastly, we enhance zero-shot generalization ability by concatenating a set of learnable prompts [30] with image features into the encoder for training.

Our CLAIR framework undergoes end-to-end training that alternates between optimizing the parameters of the domain alignment encoder network and the pseudo-labels in a refinement process. During inference, we pass images of both domains through the domain alignment encoder to obtain image features for ranking.

Our main contributions can be summarized as follows:

1. We introduce the problem setting of Weakly Supervised Zero-Shot Cross-Domain Image Retrieval (WSZS-CDIR), and develop a novel framework CLAIR that tackles all the challenges of this setting.
2. We propose a novel domain mapping function that requires no training images to reduce the domain gap.
3. We propose a multi-granular contrastive learning loss that minimizes feature distance between semantically similar samples across instances, clusters and domains.
4. We validate the efficacy of our proposed framework with extensive experiments on the TUBerlin Extended, Sketchy Extended, Quickdraw Extended and DomainNet Zero-Shot datasets. The latter is a modified train/test split of DomainNet, specifically tailored by us for the evaluation of zero-shot tasks spanning multiple domains.

2 Related Work

Unsupervised Cross-Domain Image Retrieval. [10] uses optimal transport to establish paired domain correspondence. [9] combines cluster-wise contrastive loss and a cross-domain distance of distance loss to reduce the domain gap, while [29] minimizes cross-domain discrepancy of both classifiers by enforcing consistent predictions of the same image. However, UCDIR approaches do not work under the zero-shot setting.

Zero-shot Sketch-based Image Retrieval. While early works relied on word embeddings [5, 35], later approaches have moved away from such semantics. [11, 22] adopts test-time training to improve domain alignment before inference. [16, 26] applied Vision Transformers (ViTs) [6] to train a Zero-shot model. [23] is the first to utilize CLIP for ZS-SBIR tasks. However, these approaches all require a training dataset with explicit categorical labels or paired images across domains. More recently, [32] proposes self-distillation with DINO [19] to tackle ZS-SBIR with an Unsupervised constraint (UZS-SBIR).

Contrastive Language Pre-training. CLIP [21] is a cross-modal visual-language foundation model trained on a substantial dataset comprising approximately 400 million paired samples. Its robustness and generalizability have been showcased in various downstream applications such as Image Classification [2], ZS-SBIR [23], Zero-shot Semantic Segmentation [36], prompt-guided semantic editing [1], *etc*. We adopt CLIP to generate pseudo labels for weak supervision and for zero-shot generalization across domains.

Prompt Learning. Fundamentally, the concept involves modifying inputs to add context to foundation models, improving performance. [30] optimized a frozen model with learnable prompts for specific tasks. [12] showed that adding a small amount of learnable prompts boosts ViT performance across various downstream tasks. [7, 25] demonstrated the efficiency of prompt-learning CLIP for Unsupervised Domain Adaptation (UDA). We integrate prompt learning to improve our model's ability to retrieve unseen classes.

3 Problem Definition

We follow the settings of UZS-CDIR. Given a query image $I_i^A \in \mathbb{R}^{h \times w \times c}$ from domain A, our task is to retrieve images of the same category in domain B, where i is the index of

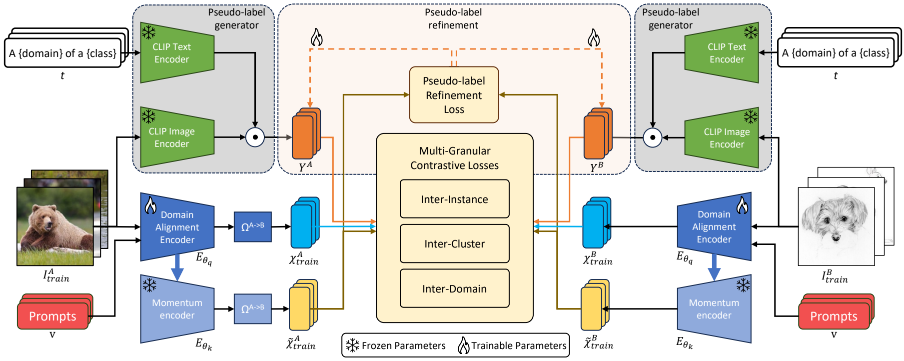

Figure 2: An overview of our proposed CLAIR. **Before training**, CLIP is used to initialize pseudo-labels. A Cross-domain mapping function $\Omega^{A \to B}$ provides initial feature alignment, with parameters updated during training. Learnable prompts v are input along with images to the domain alignment encoder E_{θ_q}, mirrored with a momentum encoder E_{θ_k} for training stability. **During training**, E_{θ_q} is optimized using contrastive losses on multiple granualities, guided by pseudo-labels and samples from the features χ_{train} generated by E_{θ_k}. Pseudo-labels are updated at every epoch end via a refinement loss supervised by χ_{train}, with gradient scaling biased to unconfident labels. The prompts v and both E_{θ_q} and E_{θ_k} are shared across domains. **During inference**, E_{θ_k} is used to produce image features. The framework is best viewed in color.

the image. Different from the traditional zero-shot setting where training images from both domains are labeled, our setting utilizes only the unlabeled training images $\mathcal{I}^A_{train} = \{I^A_i\}^N_{i=1}$ and $\mathcal{I}^B_{train} = \{I^B_j\}^M_{j=1}$, and the set of training categories names γ_{train}. As defined in [18] for CLIP-based zero-shot problems, we evaluate our model on novel categories, *i.e.*, $\gamma_{train} \cap \gamma_{test} = \varnothing$. Additionally, training images across the two domains are unpaired. We convert UZS-CDIR into WSZS-CDIR by generating noisy pseudo labels of the unlabeled training data using CLIP.

4 Our Methodology

Fig. 2 shows an illustration of our CLAIR for WSZS-CDIR. Formally, we train a domain alignment encoder $E_{\theta_q} : I_i \mapsto x_i \in \mathbb{R}^d$ that projects images from both domains to a common feature space. We denote $y^A_i \in \mathbb{R}^K$ as the pseudo-label vector of the embedding vector x^A_i and its corresponding image I^A_i, where K is the number of classes in the training dataset. E_{θ_q} is mirrored by a momentum encoder $E_{\theta_k} : I_i \mapsto \tilde{x}_i \in \mathbb{R}^d$, whose weights are updated at the end of every epoch by $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$, where m is a hyperparameter. It provides a stabler set of embeddings to be used as loss supervision [4]. During inference, image features are encoded using only E_{θ_k}.

4.1 Pseudo-Label Initialization

We employ CLIP text encoder E_t to generate text embeddings $\{w_j\}^K_{j=1}$ for the K training classes, using the prompts $\{$*"A {domain} of a {class j}"*$\}^K_{j=1}$. For each image in each dataset, its CLIP image embedding f_i is obtained. The pseudo-label weights y_i are then calculated by:

$$y_i = \left\{ \frac{f_i^\top w_j}{||f_i||||w_j||} \right\}_{j=1}^{K}, \text{ and } \quad \tilde{y}_i = \arg\max_j \sigma(y_i^j). \tag{1}$$

The j-th element of y_i is the cosine similarity score between the image I_i corresponding to the image embedding f_i and the class text embedding w_j. The set of vectors $Y^A = \{y_i^A\}_{i=1}^N \in \mathbb{R}^{N\times K}$ and $Y^B = \{y_j^B\}_{j=1}^M \in \mathbb{R}^{M\times K}$ thus forms the weights of the pseudo-labels,and the pseudo-label class assignment $\tilde{y}_i$ can be obtained by taking the argmax of $\sigma(y_i)$, where $\sigma(\cdot)$ is the Softmax operator.

4.2 Pseudo-Label Refinement

To refine the quality of pseudo-labels, we alternate training between updating the parameters of E_{θ_q} and the pseudo-label weights [2]. Every epoch end, we fix our network and update Y^A and Y^B for n_r steps using the Kullback-Leiber (KL) loss function. Specifically, we apply the gradients for every $y_i \in Y$, *i.e.*:

$$y_i \leftarrow y_i - \psi(y_i) \circ \nabla_{y_i} \mathrm{KL}(\sigma[\hat{x}_i] \mid \sigma[y_i]), \tag{2}$$

$$\text{where} \quad \hat{x}_i = \left\{ \frac{\tilde{x}_i^\top w_j}{||\tilde{x}_i||||w_j||} \right\}_{j=1}^{K}, \text{ and } \quad \psi(y_i) = \left\{ \frac{c_\psi}{\lambda_\psi\sqrt{2\pi}} \exp\left[-\frac{(y_{ij}-0.5)^2}{2\lambda_\psi^2} \right] \right\}_{j=1}^{K}. \tag{3}$$

We drop the domain superscripts in this section for brevity. In this loss function, $\sigma[\hat{x}_i]$ and $\sigma[y_i]$ are the target and predicted distributions respectively. $\hat{x}$ is obtained by a cosine similarity operation with the features $\tilde{x}_i$ in a memory bank encoded by E_{θ_k} and text embeddings w_j in a similar process used to initialize y_i. $\psi(y_i)$ is a normal distribution function of y_i centered at 0.5, used to scale the loss gradients and maximize updating of the least-confident parameters, *i.e.*, those with values around 0.5. c_ψ and λ_ψ are hyperparameters that correspond to the weight and the standard deviation of $\psi(y_i)$ respectively.

Subsequently, we fix Y^A and Y^B to use as supervision for training E_{θ_q}. For symmetry, we include $\mathcal{L}_{KL}$, which flips the inputs of the KL loss. This alternation of model training and pseudo-label refinement continues until convergence or upon reaching the maximum number of epochs.

4.3 Multi-Granular Contrastive Losses

We propose a multi-granular contrastive learning approach to align features on three levels of granularity using the pseudo-labels as supervision. On an instance-level, we apply the standard contrastive loss from [3]:

$$\mathcal{L}_{II} = \frac{-1}{|\mathbf{I}|} \sum_{i\in\mathbf{I}} \log\left[\frac{\exp(x_i^\top x_i'/\tau)}{\sum_{a\in\mathbf{I}} \exp(x_i^\top \tilde{x}_a/\tau)} \right], \tag{4}$$

where x_i and x_i' are features vectors of different augmentations of I_i encoded by E_{θ_q}, while $\tilde{x}_a$ represents feature vector in the memory bank encoded by E_{θ_k}. τ is an adjustable temperature hyperparameter. We denote $\mathbf{I}$ as indices of the entire training dataset for one domain, and $|\mathbf{I}|$ as its cardinality.

To ensure that our model learns more semantically meaningful features, we extend contrastive loss to the cluster level. We form positive pairs between instances of the same cluster within a domain, where each cluster contains samples of the same pseudo-labels. We define the inter-cluster loss on domain A as:

$$\mathcal{L}_{IC}^{A} = \frac{-1}{|\mathbf{I}^A|} \sum_{i \in \mathbf{I}^A} \frac{1}{|P^A(i)|} C^A(x_i^A, \tilde{x}_p^A, \tilde{x}_a^A), \tag{5a}$$

where

$$C^A = \sum_{p \in P^A(i)} \log \left[\frac{\exp(x_i^{A,\top} \tilde{x}_p^A / \tau)}{\sum_{a \in \mathbf{I}^A} \exp(x_i^{A,\top} \tilde{x}_a^A / \tau)} \right]. \tag{5b}$$

To help bridge the domain gap, we further extend the loss function to the domain level. We form positive pairs between instances of the same cluster across domains. We define the inter-domain loss on domain A as:

$$\mathcal{L}_{ID}^{A} = \frac{-1}{|\mathbf{I}^A|} \sum_{i \in \mathbf{I}^A} \frac{1}{|P^B(i)|} D^A(x_i^A, \tilde{x}_p^B, \tilde{x}_a^B) \tag{6a}$$

where

$$D^A = \sum_{p \in P^B(i)} \log \left[\frac{\exp(x_i^{A,\top} \tilde{x}_p^B / \tau)}{\sum_{a \in \mathbf{I}^B} \exp(x_i^{A,\top} \tilde{x}_a^B / \tau)} \right]. \tag{6b}$$

Here, $P(i)$ represents the indices for the set of samples belonging to the same assigned pseudo-class as I_i, and $|P(i)|$ as its cardinality. Specifically, for an image I_i^A from domain A, $P^A(i) = \{p \in \mathbf{I}^A : \tilde{y}_p^A = \tilde{y}_i^A\}$ and $P^B(i) = \{p \in \mathbf{I}^B : \tilde{y}_p^B = \tilde{y}_i^A\}$. $\mathcal{L}_{IC}^B$ and $\mathcal{L}_{ID}^B$ can be calculated in the same way.

For every loss objective, we average the loss between domains, *i.e.* $\mathcal{L} = 0.5(\mathcal{L}^A + \mathcal{L}^B)$, thus forming our full training objective:

$$\mathcal{L}_{total} = w_{KL}\mathcal{L}_{KL} + w_{II}\mathcal{L}_{II} + w_{IC}\mathcal{L}_{IC} + w_{ID}\mathcal{L}_{ID}, \tag{7}$$

where w_{KL}, w_{II}, w_{IC}, and w_{ID} are hyperparameters to balance the loss terms.

4.4 Cross-Domain Mapping Function

While paired cross-domain images are not available, it is trivial to generate paired text embeddings using prompt templates. We make the assumption that embeddings of domain-specific prompt templates are a good approximation for their image counterparts, and therefore a mapping function obtained from the text embeddings would work on the image embeddings:

$$\Omega w^A = w^B \implies \Omega x^A = x^B, \tag{8}$$

where Ω is a transformation matrix that maps features from one domain to another. Paired text embeddings w^A and w^B can be obtained from encoding the text template *"A {domain} of {object}"* using text encoder E_t. For example, *"An infograph of an airplane"* is paired with *"A photo of an airplane"*. Using internet-sourced object names, we create 3,000 text embeddings pairs for each domain pair. If Ω is assumed to be orthogonal, then the optimal solution $\text{argmin}_\Omega \|\Omega w^A - w^B\|^2$ can be solved in closed-form:

$$\Omega = UV^\top, \tag{9}$$

where $U\Sigma V^\top$ is the decomposition of $w^B w^{A\top}$ [24]. Ω can be directly inserted into our model even before training to improve performance. We found empirically that allowing the parameters of Ω update along with the rest of the model during training yields better results. It should be noted that the cross-domain mapping function improves cross-domain alignment without being adversely affected by noisy pseudo labels since it is learned fully from text embeddings.

4.5 Prompt Learning

To handle unseen categories and prevent *catastrophic forgetting*, we follow [23] to freeze all layers of our CLIP-pretrained domain alignment encoder except for the Layer Norm layers. A set of learnable prompts $\mathrm{v} \in \mathbb{R}^{n_p \times d_p}$ is then concatenated with image features from both domains for knowledge distillation and preserving the generalization ability of CLIP. n_p and d_p are hyperparameters that correspond to the number and dimensionality of prompts respectively.

The remaining trainable parameters of E_{θ_q} and v are updated using the loss functions described in the above sections. Limiting the training to a smaller subset of features has the added benefit of expediting convergence, leading to reduced overall training time.

In summary, we integrate CLIP extensively into our framework, encompassing pseudo-label initialization, multi-granular contrastive losses, the approximation of a no-train domain-aligning mapping function, and prompt learning. Detailed descriptions of our hyperparameter tuning strategy can be found in the supplementary material.

5 Experiments

5.1 Dataset

We evaluate on the established ZS-SBIR datasets **Sketchy Extended [17]**, **TUBerlin Extended [34]** and **QuickDraw Extended [8]**, and follow [31] by reporting Precision scores for Sketch to Image domain. Additionally, we propose **DomainNet Zero-Shot**, a subset of DomainNet [20] which spans across 6 domains (Clipart (C), Infograph (I), Sketch (S), Quickdraw (Q), Painting (P) and Real (R)) and shares the same 345 categories as the original Quickdraw dataset [8]. Thus, we follow [8] and split DomainNet into the same train/test split, resulting in over 200,000 samples across 6 domains. Following [9], we report results on 6 domain pairings: C-S, I-R, I-S, P-C, P-Q and Q-R, ensuring that every domain is experimented on twice, and retrieval is evaluated bidirectionally (*i.e.* C$\rightarrow$S and S$\rightarrow$C). We plot the distribution of images across classes and domains for visualization in the supplementary material. Finally, we randomly sample approximately 12 samples per training category for every dataset, which we set aside as a validation dataset.

5.2 Implementation Details

Our proposed approach is implemented in PyTorch on a 24GB NVIDIA RTX 4090 GPU. For both initializing the pseudo-labels and warm-starting our backbone encoder, we use the ViT encoder of CLIP with ViT-B/32 weights. Only the prompts and the weights of Layer Normalization layers in E_{θ_q} are set to update in training. These parameters are trained with a SGD optimizer with learning rate set to 0.0002, applied with a cosine learning rate schedule that gradually reduces it to 0, and with early stopping employed for regularization. Following [23], we inject our prompts $\mathrm{v} \in \mathbb{R}^{10 \times 768}$ into the first layer of the transformer, and use the same encoder and prompts for both domain inputs. Resizing all images to 224×224 pixels,

Setting	Method	Sketchy	TUBerlin	Quickdraw
		P@200	P@100	P@200
US ZS	CDS [13]	24.99	18.58	6.49
	PCS [33]	13.29	13.89	3.65
	CoDA [29]	17.02	15.85	5.15
	CLIP-UCDIR [9]	40.63	38.75	8.90
	Vanilla CLIP [21]	35.01	45.40	8.42
	AMA [32]	58.50	59.20	19.20
WS ZS	CLAIR-A (Ours-A)	64.64	71.17	21.49
	CLAIR-B (Ours-B)	**65.74**	**72.11**	**23.07**

Setting	Method	Sketchy	TUBerlin	Quickdraw
		P@200	P@100	P@200
FS ZS	TCN [27]	60.8	61.6	29.8
	TVT [26]	61.8	66.2	29.3
	PSKD [ViT] [28]	64.5	66.2	29.8
	Sketch3T [22]	64.8	67.1	-
	ZSE-SBIR [16]	62.40	65.70	21.60
	SketchLVM [23]	72.50	73.20	38.80
	Dr.CLIP [15]	70.60	76.30	31.20
	DCDL [14]	76.90	74.10	29.60

Table 1: Experimental results on the Zero-Shot test splits of Sketchy Extended, TUBerlin Extended and QuickDraw Extended datasets. **Bold** represent best results. **Unsupervised Zero-Shot (US ZS)**: The model is trained with unlabeled images. **Weakly Supervised Zero-Shot (WS ZS)**: The model is trained with noisy pseudo labels generated from CLIP. **Fully Supervised Zero-Shot (WS ZS)**: The model is trained with image labels as supervision. CLAIR-A uses the same hyperparameters across all experiments while CLAIR-B follows AMA to tune hyperparameters on Sketchy. More details on our baseline methods can be found in the supplementary section.

Method	DomainNet Zero-Shot												
	C→S	S→C	I→R	R→I	I→S	S→I	P→C	C→P	P→Q	Q→P	Q→R	R→Q	Avg.
CDS [13]	36.64	39.07	38.26	24.61	24.51	18.27	37.87	39.68	8.27	11.28	12.60	8.14	24.93
PCS [33]	34.37	30.71	17.85	34.37	17.99	17.59	35.72	26.35	8.38	6.55	13.82	11.95	21.30
CoDA [29]	30.65	32.85	35.31	18.22	19.67	15.06	29.37	33.30	5.99	7.22	7.99	6.49	20.18
CLIP-UCDIR [9]	42.43	41.06	21.02	24.73	28.35	35.02	43.25	42.32	9.31	10.27	10.68	9.78	26.52
Vanilla CLIP [21]	67.07	63.57	56.62	41.24	49.13	28.50	61.00	51.44	34.45	9.50	8.41	38.39	42.44
CLAIR-A (Ours)	**81.61**	**76.24**	**56.79**	**68.92**	**51.53**	**60.34**	**74.73**	**81.19**	**44.93**	**26.24**	**26.32**	**51.14**	**58.33**

Table 2: Precision at top 50 (%) on the Zero-shot test splits of all domain pairings in the DomainNet Zero-Shot dataset, compared against Unsupervised baselines. Domains: C - Clipart, S - Sketch, I - Infograph, P - Painting, Q - Quickdraw, R - Real.

we train our model with a batch size of 64 for 25 epochs, and it maps the input image into a feature space $x \in \mathbb{R}^{64}$. With the reduced learnable parameters, we observed quick training convergence: 9.5 GB RAM and 21 minutes across all DomainNet Zero-Shot pairings, and 2.1 GB RAM and 25 seconds for inference.

5.3 Evaluation Metrics

As our objective involves category-level image retrieval, we adopt the evaluation metrics from [9, 13] and calculate precision at the top k retrievals (P@k) for all samples across all classes. We report P@100 for TUBerlin Extended and P@200 for Sketchy Extended and Quickdraw Extended datasets, aligning with ZS-SBIR literature. For DomainNet Zero-Shot, we report P@50 scores bidirectionally for each domain pair. Specifically, for two domains A and B, scores are reported with A as the query domain and B as the target domain and vice versa. In a similar vein with recent UCDIR works, we provide additional P@100 and P@200 scores, as well as Precision-Recall curves for all DomainNet Zero-Shot experiments in the supplementary material.

Method	C → S	S → C	I → R	R → I
w/o mapping	67.07	63.57	56.61	41.24
w/ mapping	**70.36**	**69.46**	**59.48**	**55.81**
Method	I → S	S → I	P → C	C → P
w/o mapping	49.13	28.50	61.00	51.44
w/ mapping	**52.67**	**42.61**	**66.26**	**69.73**
Method	P → Q	Q → P	Q → R	R → Q
w/o mapping	34.45	9.50	8.42	38.39
w/ mapping	**36.71**	**12.87**	**10.78**	**40.48**
Average P@50	Without mapping		With mapping	
	42.44		**48.93**	

Ablation	Experiment	Avg. P@50
	Full model	**57.52**
(i) Leave one out	w/o prompt	53.40
	w/o mapping	54.68
	w/o refinement	51.62
	w/o freeze	50.64
	w/o $\mathcal{L}_{KL}$	56.51
	w/o $\mathcal{L}_{II}$	55.44
	w/o $\mathcal{L}_{IC}$	54.50
	w/o $\mathcal{L}_{ID}$	50.11
(ii) Expanded class list	Full expanded list	55.24
	Exclude training class names	54.56
(iii) In the wild dataset	Mismatched categories	57.09

Table 3: [Left] Effect of cross-domain mapping function on image retrieval using Vanilla CLIP. [Right] The combined table of three ablation studies: (i) significance of each component, (ii) effect of using training class names to initialize pseudo-labels, (iii) impact of having mismatched classes between domains. P@50 scores are averaged across all domain pairings from the DomainNet Zero-Shot dataset.

5.4 Results

Table 1 shows that our proposed CLAIR outperforms AMA on all comparable datasets, thus validating our methodology. Additionally, table 2 demonstrates the generalizability of our CLAIR, surpassing all Unsupervised baselines on every domain pair from the DomainNet Zero-Shot dataset. Notably, both our CLAIR and CLIP-UCDIR initiate training with identical weights inherited from Vanilla CLIP. The performance of CLIP-UCDIR declines after fine-tuning, reminiscent of catastrophic forgetting. In comparison, our model improves on Vanilla CLIP by 15.89% on average in terms of P@50 on all datasets. Some qualitative results (including failure cases) are illustrated in the supplementary material. We observe that most of the top images, even the false positives, retrieved by our method resemble the query image in likeliness. In fact, several false positives are due to images containing multiple objects (*e.g.* a monkey on a tree) but are only assigned one class, illustrating the limitation of the single-label nature of the dataset.

An interesting observation arises when considering the performance of CLIP in scenarios where Quickdraw serves as the query domain. Consisting of a large amount of rudimentary drawn doodles, the high level of abstraction presents a significant domain gap. Consequently, CLIP faces challenges in such contexts. In comparison, our CLAIR not only mitigates these challenges but also demonstrates substantial improvement over CLIP in such domain pairings by an average P@50 score of 14.65%, 16.7% and 17.9% on the QuickDraw extended dataset, and the Q → P and Q → R pairings of the DomainNet Zero-Shot dataset, respectively. This performance differential signifies the robustness of our training approach in addressing domain-specific intricacies and optimizing retrieval accuracy.

5.5 Ablation Study

Influence of the Cross-Domain Mapping Function. Table 3 shows the impact of the cross-domain mapping function Ω. The inclusion of Ω improves performance of Vanilla CLIP on all DomainNet Zero-Shot pairs by an average of 6.49%. This affirms that a mapping function derived solely from domain-specific text prompts without exposure to any training images enhances retrieval scores within their respective image domains, and aligns with our objective of initializing the model with a closed-form solution to mitigate the domain gap prior to training.

Performance contribution of components. Table 3 accesses the impact of each module on the performance of our model by systematically excluding one component at a time. The results validate our design choices in the model architecture, where every component contributed to its performance. Moreover, it also shows the removal of a single component does not make training fail catastrophically. Of all the components, the most significant ones are the inter-domain contrastive loss and the freezing of CLIP weights, elevating P@50 accuracy from 50.11% and 50.64% to 57.52%, respectively.

Impact of using training class names. CLAIR makes use of the training category names to initialize pseudo-labels. To assess the potential bias introduced by using the category names, we propose an alternate source and generate the pseudo labels using the 3000 object names used to solve the cross-domain mapping function. To ensure a zero-shot setting, the test category names are excluded from the expanded list. Additionally, we further exclude all training classes from this expanded list to evaluate the impact of its inclusion during training. Table 3 shows that having the names of training classes is helpful but not essential. Expanding the pseudo-label list of classes and excluding the training labels specifically only leads to a performance degradation of 2.28% and 2.96% respectively. Notably, these scores still outperform all other unsupervised baselines.

Assessment on mismatched categories. The datasets used in our evaluation all share the same classes across domains. This assumption may not hold for more realistic, or 'in-the-wild' datasets, especially for unsupervised tasks. To assess this impact on our proposed method, we further subset the train split of the DomainNet Zero-Shot dataset into two groups, with 60 classes in each group. Each domain consists of 20 unique classes not found in the other domain, with the remaining 40 classes shared. Naturally, these class distributions are not known by our model during training. We observe from table 3 that performance degradation is almost negligible (0.43%), which means that our model can perform well even if the categories in the query and target domains are mismatched.

6 Conclusion

In this paper, we propose the problem setting of Weakly Supervised Zero-Shot Cross-Domain Image Retrieval (WSZS-CDIR) to alleviate the burden of human annotation and enable retrieval on any novel categories, and introduce CLAIR to solve this task. Our CLAIR first uses CLIP to generate noisy pseudo-labels which we then refine iteratively alongside a domain-aligned feature encoder throughout the training process. To enhance cross-domain image alignment, we propose a novel contrastive learning strategy that strategically selects positive and negative pairs at various levels of granularity, ranging from inter-instance samples to inter-domain clusters. We also design an effective cross-domain mapping function that is unaffected by noisy pseudo-labels to further improve the alignment of features across domains. Experimental results on the Sketchy Extended, TUBerlin Extended, QuickDraw Extended and DomainNet Zero-Shot datasets demonstrate the superiority of our CLAIR over baseline approaches for every domain pairing.

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022.

[2] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1348–1357, 2023.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[5] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2179–2188, 2019.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022.

[8] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

[9] Conghui Hu and Gim Hee Lee. Feature representation learning for unsupervised cross-domain image retrieval. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022.

[10] Conghui Hu, Yongxin Yang, Yunpeng Li, Timothy M Hospedales, and Yi-Zhe Song. Towards unsupervised sketch-based image retrieval. *arXiv preprint arXiv:2105.08237*, 2021.

[11] Finlay GC Hudson and William AP Smith. If at first you don't succeed: Test time re-ranking for zero-shot, cross-domain retrieval. *arXiv preprint arXiv:2303.17703*, 2023.

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

[13] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cds: Cross-domain self-supervised pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9123–9132, 2021.

[14] Qiang Li, Shihao Wang, Wei Zhang, Shaojin Bai, Weizhi Nie, and Anan Liu. Dcdl: Dual causal disentangled learning for zero-shot sketch-based image retrieval. *IEEE Transactions on Multimedia*, 2025.

[15] Xue Li, Jiong Yu, Ziyang Li, Hongchun Lu, and Ruifeng Yuan. Dr. clip: Clip-driven universal framework for zero-shot sketch image retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9554–9562, 2024.

[16] Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, and Yonggang Qi. Zero-shot everything sketch-based image retrieval, and in explainable style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23349–23358, 2023.

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

[18] Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, et al. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *arXiv preprint arXiv:2407.21794*, 2024.

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

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

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

[22] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3t: Test-time training for zero-shot sbir. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7462–7471, 2022.

[23] Aneeshan Sain, Ayan Kumar Bhunia, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2023.

[24] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[25] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4355–4364, 2023.

[26] Jialin Tian, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Tvt: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2370–2378, 2022.

[27] Hao Wang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Transferable coupled network for zero-shot sketch-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9181–9194, 2021.

[28] Kai Wang, Yifan Wang, Xing Xu, Xin Liu, Weihua Ou, and Huimin Lu. Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 601–609, 2022.

[29] Xu Wang, Dezhong Peng, Ming Yan, and Peng Hu. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. *arXiv preprint arXiv:2302.06081*, 2023.

[30] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[31] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018.

[32] Zhihui Yin, Jiexi Yan, Chenghao Xu, and Cheng Deng. Asymmetric mutual alignment for unsupervised zero-shot sketch-based image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16504–16512, 2024.

[33] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021.

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

[35] Zhaolong Zhang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Zero-shot sketch-based image retrieval via graph convolution network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12943–12950, 2020.

[36] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.