Asymmetric Mutual Alignment for Unsupervised Zero-Shot Sketch-Based Image Retrieval

Zhihui Yin[1], Jiexi Yan[2*], Chenghao Xu[1], Cheng Deng[1*]

[1]School of Electronic Engineering, Xidian University, China
[2]School of Computer Science and Technology, Xidian University, China
yzh.xdu@gmail.com, yanjiexi@xidian.edu.cn, chx@stu.xidian.edu.cn, chdeng.xd@gmail.com

Abstract

In recent years, many methods have been proposed to address the zero-shot sketch-based image retrieval (ZS-SBIR) task, which is a practical problem in many applications. However, in real-world scenarios, on the one hand, we can not obtain training data with the same distribution as the test data, and on the other hand, the labels of training data are not available as usual. To tackle this issue, we focus on a new problem, namely unsupervised zero-shot sketch-based image retrieval (UZS-SBIR), where the available training data does not have labels while the training and testing categories are not overlapping. In this paper, we introduce a new *asymmetric mutual alignment* method (AMA) including a self-distillation module and a cross-modality mutual alignment module. First, we conduct self-distillation to extract the feature embeddings from unlabeled data. Due to the lack of available information in an unsupervised manner, we employ the cross-modality mutual alignment module to further excavate underlying intra-modality and inter-modality relationships from unlabeled data, and take full advantage of these correlations to align the feature embeddings in image and sketch domains. Meanwhile, the feature representations are enhanced by the intra-modality clustering relations, leading to better generalization ability to unseen classes. Moreover, we conduct an asymmetric strategy to update the teacher and student networks, respectively. Extensive experimental results on several benchmark datasets demonstrate the superiority of our method.

Introduction

Sketch-based Image Retrieval (SBIR) (Eitz et al. 2010; Qi et al. 2016; Liu et al. 2017) is a practical problem that addresses retrieving relevant images from a gallery using sketches as queries. Under the framework of deep neural networks, SBIR focuses on learning better representations for abstract sketches (Eitz, Hays, and Alexa 2012) and the domain gap between sketches and images (Yu et al. 2016; Song et al. 2017). Since sketches can effectively convey the shape, pose, and fine-grained details of relevant objects, SBIR provides a superior alternative to text-based methods for retrieving images in situations where language cannot express visual characteristics accurately.

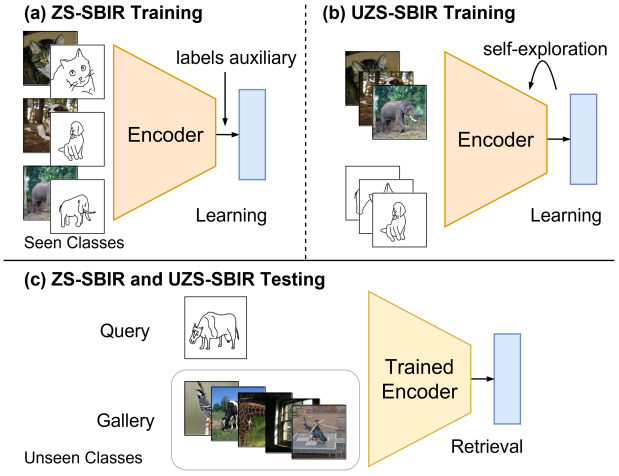


Figure 1: The setting of ZS-SBIR and UZS-SBIR. (a) ZS-SBIR typically trains on seen classes and supplemented with labels auxiliary such as word embeddings. (b) UZS-SBIR also trains on seen classes. However, due to the absence of labels, it is impossible to establish a one-to-one correspondence between training samples of sketches and images. (c) Both of them test on unseen classes for retrieval.

The application scenarios where SBIR is feasible are closer to the zero-shot setting, requiring the training and testing categories to be non-overlapping. ZS-SBIR (Yelamarthi et al. 2018; Liu et al. 2019; Tian et al. 2022; Sain et al. 2023; Lin et al. 2023), which combines zero-shot learning (Lampert, Nickisch, and Harmeling 2013; Xian et al. 2018; Chen et al. 2019; Yan et al. 2022) and SBIR as a single task, has become a research problem that has garnered increased attention recently. These works leverage labeled data and external auxiliary knowledge to achieve cross-category adaptation between sketches and images. However, the origin of the ZS-SBIR problem stems from the scarcity of annotated sketches. Therefore, we approach the essence of the problem and investigate Unsupervised Zero-shot Sketch Retrieval (UZS-SBIR) which eliminates the necessity of any annotations.

A core challenge in UZS-SBIR is how to effectively leverage unlabeled cross-modality data. To tackle this is-

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

sue, we propose a novel method, *asymmetric mutual alignment* (AMA), that separately extracts features of image and sketch samples in an unsupervised manner. After obtaining the feature embeddings from the self-distillation module, we implement a cross-modality mutual alignment module to further make full use of available unlabeled cross-modality data to capture the modality-related correlations and achieve good generalization ability to unseen classes. Specifically, we utilize stable clustering results of the teacher network as a reference to guide the student network with a well-designed contrastive clustering loss. This process is conducted on image and sketch modalities, respectively. And then, we introduce mutual alignment to enhance the consistency of their distributions, which can effectively structure the modality-aware relationship. This is essential for image-to-sketch retrieval. Furthermore, we adopt an asymmetric strategy to update the teacher and student networks, *i.e.*, the student network is updated by backpropagation while we iteratively update the teacher network with EMA (Caron et al. 2021).

We summarize the contributions as follows:

- We introduce a more challenging yet realistic problem, *i.e.*, UZS-SBIR task.
- We propose a simple yet effective method, namely AMA, that takes full advantage of unlabeled cross-modality data to derive discriminative feature representations, which also have good generalization ability to unseen classes.
- Extensive experiments on several widely-used datasets demonstrate the state-of-the-art retrieval performance of our method.

Related Work

Self-Supervised Learning

Self-supervised learning aims to acquire intermediate features of superior quality that can be effectively applied to various downstream tasks. The existing approaches in self-supervised learning have predominantly revolved around contrastive method (Chen et al. 2022; Yan et al. 2023a). MoCo (He et al. 2020) built a dynamic dictionary with a queue and a momentum encoder to enhance the training efficacy. SimCLR (Chen et al. 2020a) showed the non-necessity of the memory bank and illustrates the importance of data augmentation. SwAV (Caron et al. 2020) achieved a more memory-efficient method by employing online clustering for contrastive learning. Dino (Caron et al. 2021) extended the advantages of previous works, such as momentum encoder (He et al. 2020)and multi-crop training (Caron et al. 2020), to ViTs (Dosovitskiy et al. 2021). Through unsupervised self-distillation, Dino established an excellent baseline for subsequent work. Moreover, self-supervised methods based on masked image modeling aroused widespread attention, including BEiT (Bao et al. 2022), which predicts discrete tokens, MAE (He et al. 2022), which relies on masked reconstruction, and the simpler SimMIM (Xie et al. 2022) framework. This paper focuses on less redundant sketch data, which is more suitable for employing contrastive methods.

Zero-Shot Sketch-Based Image Retrieval

ZS-SBIR requires the model to acquire robust generalization knowledge from the seen training set and apply it to an unseen testing set. The work of (Yelamarthi et al. 2018) introduced this task and proposes a generative model based on autoencoders that utilize matched sketch-image pairs for training. (Dutta and Akata 2019) employed adversarial training to reduce the sketch-photo domain gap. Besides, other works (Dey et al. 2019; Zhang et al. 2020) explored the semantic information contained within word embeddings to enhance the transferability of the model. Subsequent works introduced knowledge distillation, utilizing pre-trained teacher networks on large-scale datasets to optimize student networks (Liu et al. 2019; Wang et al. 2021; Tian et al. 2021). This also included prototype-based selective knowledge distillation (Wang et al. 2022) and the fusion of image and sketch feature representation space (Tian et al. 2022). Furthermore, recent works focused on various facets, including a test-time training paradigm (Sain et al. 2022), prompt learning for ZS-SBIR (Sain et al. 2023), and patch correspondences in explainable style (Lin et al. 2023). Nevertheless, these methods rely on labeled information from images and sketches, which is inherently scarce. To address this issue, we aim to investigate a novel setting where both images and sketches are unlabeled.

Unsupervised Cross-Domain Image Retrieval

Unsupervised cross-domain image retrieval involves learning the semantic concepts of images and achieving data alignment between different domains without label supervision. (Kim et al. 2021a) introduced a cross-domain self-supervised pre-training method, which learns discriminative and domain-invariant features using unlabeled multi-domain data. (Yue et al. 2021) utilized prototypes to perform unified, unsupervised, and adaptive semantic structure learning, discriminative feature learning, and cross-domain alignment. These works focus on minimizing the domain discrepancy through self-supervised learning on both intra-domain and inter-domain data, aiming to attain models with improved domain adaptability similar to (Yan et al. 2023b; Chen et al. 2023). In order to enhance its applicability in retrieval tasks, (Hu and Lee 2022) proposes a novel distance-of-distance loss to facilitate improved domain alignment. Besides, (Wang et al. 2023) enforces the predictions by different domain-specific classifiers to consistently obtain domain-invariant representations from unlabeled data. For the sketch, data-free sketch-based image retrieval (Chaudhuri et al. 2023) achieved sketch-based retrieval without requiring training set images. However, it heavily relies on the quality of the pre-trained classifier, which is often associated with labeled data. The problem we aim to tackle is an entirely unsupervised problem without any data labels.

Method

Problem Setting

Unsupervised Zero-shot Sketch Retrieval comprises two distinct settings: the training set and the testing set, each encompassing varying categories, all within the context of the

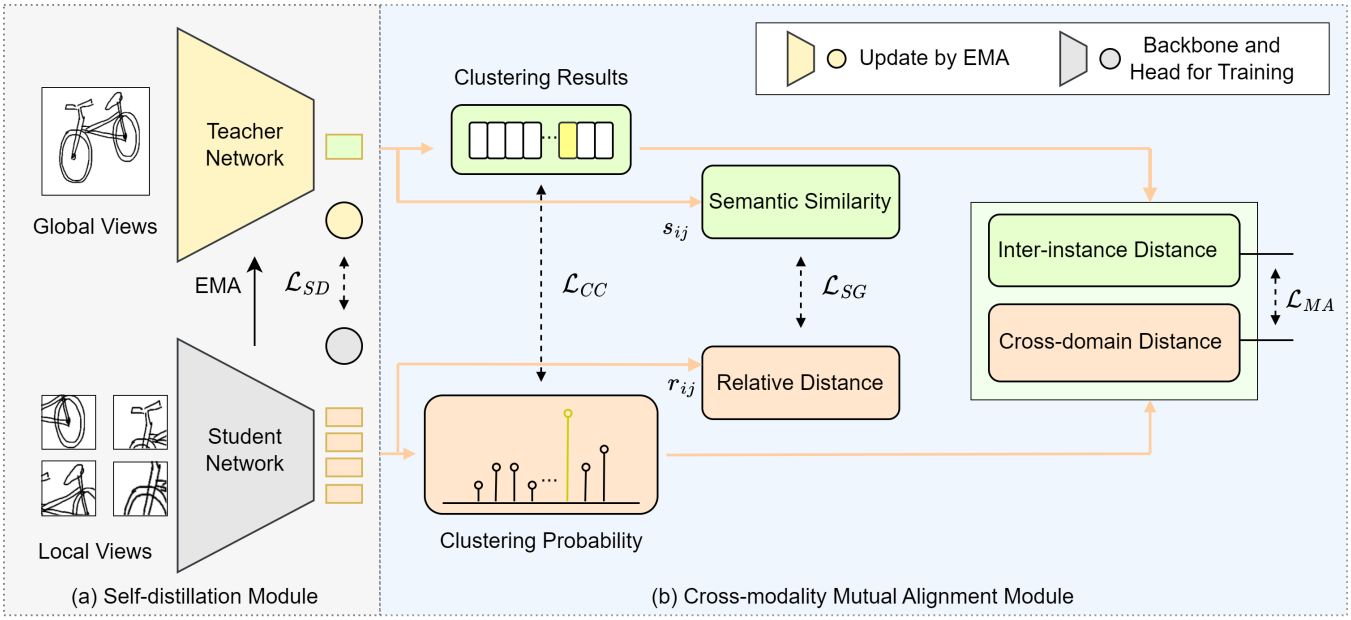


Figure 2: The illustration of basic architectures of our proposed AMA. (a) Self-distillation module finds the most informative feature representation. (b) Cross-modality mutual alignment module captures the modality-related correlations and generalizable knowledge.

absence of category labels. In the training phase, we define the training set for seen classes as $\mathcal{D}_{train} = \{\mathcal{I}^S, \mathcal{S}^S\}$, where$\mathcal{I}^S = \{I_i\}_{i=1}^{N_I}$ and $\mathcal{S}^S = \{S_j\}_{j=1}^{N_S}$ represent unlabeled actual images and unlabeled sketches belonging to the set of seen classes denoted as $\mathcal{C}^S$, respectively. Correspondingly, the test set is designated as $\mathcal{D}_{test} = \{\mathcal{I}^U, \mathcal{S}^U\}$, specifically pertaining to the unseen classes categorized under $\mathcal{C}^U$. In this particular zero-shot scenario, the essential requirement is the mutual exclusivity of $\mathcal{C}^S$ and $\mathcal{C}^U$. Consequently, the challenge at hand revolves around effectively extracting features that are both invariant to domain shifts and generalizable across categories.

Self-Distillation Module

Inspired by Dino (Caron et al. 2021), we have recognized the potential of self-distillation within the realm of unsupervised learning. Particularly noteworthy is the pivotal role played by the multi-views augmentation strategy, significantly enhancing the performance of downstream tasks. In order to harness this advantage more effectively within the context of sketch-based tasks, we introduce a novel self-distillation paradigm tailored to the Unsupervised Zero-shot Sketch-Based Image Retrieval (UZS-SBIR) domain.

To enhance the representations of distinct modalities, we undertake separate self-distillation processes for the different image modalities, specifically sketches and actual images. For each instance x, which can be either a sketch or an image, we employ a multi-crop strategy to generate a collection of multi-views denoted as $\{\mathrm{x}_k\}_{k=1}^{V}$. Within this set of views, those designated with indices $k = 1, 2, \cdots, V_1$ correspond to global views, while the remaining indices pertain to local views. Given the inherent sparsity of regions in sketches in comparison to natural images, we allocate a higher proportion of local views for sketches. Embracing the concept of "local-to-global" correspondences, our self-distillation framework incorporates a configuration wherein the teacher network solely processes the global views $\{\mathrm{x}_k\}_{k=1}^{V_1}$, whereas the student network operates on inputs from all views.

Identified as g_{θ_t} and g_{θ_s}, the teacher and student networks encompass a backbone f_T (f_S) and a head layer h_T (h_S). Given an input instance x, our overarching objective is to facilitate the training of the student network to generate predictions that align with the outputs produced by the teacher network. The specific formulation of this objective is expressed as follows:

$$R\left(\mathrm{x}\right) = \sum_{k=1}^{V_1} \sum_{\substack{k'=1 \\ k' \neq k}}^{V} \mathrm{CE}\left(\psi\left(g_{\theta_t}(\mathrm{x}_k)/\tau\right), \psi\left(g_{\theta_s}(\mathrm{x}_{k'})/\tau\right)\right) \tag{1}$$

where ψ is the softmax function, $\mathrm{CE}(\cdot,\cdot)$ denotes cross entropy operation, and τ is a temperature parameter.

Hence, the total self-distillation loss for both sketches and images can be formulated as follows:

$$\mathcal{L}_{\mathrm{SD}} = \frac{\beta_1}{N_I} \sum_{i=1}^{N_I} R\left(I_i\right) + \frac{\beta_2}{N_S} \sum_{j=1}^{N_S} R\left(S_j\right) \tag{2}$$

Unlike knowledge distillation, self-distillation applies exponential moving average (EMA) updates on the parameters of the teacher network with respect to the student network. This update can be expressed mathematically as:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s \tag{3}$$

Cross-Modality Mutual Alignment Module

While the Self-distillation Module effectively handles the challenge of unlabeled sketches and images, it does not thoroughly delve into cross-modality correlations. It struggles to overcome class shifts, leading to suboptimal generalization. To address this, we extend the paradigm to capitalize on the synergy between a stable teacher network and a highly adaptable student network. This extension aims to facilitate an asymmetric yet mutually advantageous alignment between sketches and images.

Teacher-Guided Contrastive Clustering An intuitive approach to achieve class alignment involves utilizing pseudo-labels derived from clustering. Nonetheless, before embarking on direct alignment, it is crucial to ascertain the reliability of clustering for both sketches and images. To further leverage the asymmetric nature of parameter updates inherent in the self-distillation module, we introduce a novel concept termed "teacher-guided contrastive clustering learning". This technique aims to enhance the discriminability of clustering within the feature spaces of sketches and images.

We employ a teacher network f_T to extract features for image clustering, a strategy consistent with the utilization of momentum encoders as discussed in (He et al. 2020). Independently applying k-means to the extracted features of sketches and images leads to the creation of two distinct sets of K clusters: $\left\{M_c^S\right\}_{c=1}^K$ for sketches and $\left\{M_c^I\right\}_{c=1}^K$ for images. Concurrently, each instance is assigned a pseudo-label based on the outcomes of the clustering procedure. The objective of the contrastive clustering learning approach can be succinctly formulated as follows:

$$\mathcal{L}_{\mathrm{CC}} = -\sum_{\mathbf{x}\in\mathcal{D}} I(\mathbf{x}) \sum_{\mathbf{p}\in\mathcal{X}} \log \frac{\exp(f_S(\mathbf{x})^\top f_S(\mathbf{p})/\tau)}{\sum_{\mathbf{a}\in\mathcal{D}'} \exp(f_S(\mathbf{x})^\top f_S(\mathbf{a})/\tau)} \tag{4}$$

where $I(\mathbf{x}) \in \{0,1\}$ denotes teacher-guided filter, $\mathcal{X}$ denotes $\{\mathbf{x}_k\}$, $\mathcal{D}$ and $\mathcal{D}'$ denote all training samples and the domain to which the instances belong (sketches or images) respectively.

For an instance $\mathbf{x}$, we calculate its clustering probability based on its distance to the cluster centroids:

$$p^c(\mathrm{x}, M) = \frac{\exp(f_T(\mathbf{x})^\top M_c/\tau)}{\sum_{k=1}^{K} \exp(f_T(\mathbf{x})^\top M_k/\tau)} \tag{5}$$

Therefore, $I(\mathbf{x})$ can be succinctly expressed as:

$$I(\mathbf{x}) = \mathbf{1}_{[}p^i \geq \mu] \tag{6}$$

where i is the pseudo label for $\mathbf{x}$ and μ denotes a filter parameter.

Semantic Similarity Guidance The efficacy of contrastive learning for clustering hinges on the dependability of the feature space in which the clustered features are embedded. In the absence of labels, we must undertake more self-exploration of instances in the feature space. In contrast to the student network, the teacher network, updated through EMA, possesses a more stable grasp of semantic information. This stability enables it to offer pragmatic guidance for steering the self-exploration process in the right direction.

We capitalize on the teacher network's ability to gauge the semantic similarity between instances, utilizing it as soft labels to navigate the student network's contrastive learning process. The quantification of the similarity between samples A and B is defined as follows:

$$s_{ij} = \exp\left(\frac{\|f_S(\mathbf{x}_i) - f_S(\mathbf{x}_j)\|_2^2}{\sigma}\right) \tag{7}$$

where σ is the Gaussian kernel bandwidth. We employ the relaxed metric loss (Kim et al. 2021b) to facilitate the guidance of semantic similarity. The relative distance between samples A and B is defined as:

$$r_{ij} = \frac{\phi^(f_S(\mathbf{x}_i) - f_S(\mathbf{x}_j))}{\frac{1}{N}\sum_{k=1}^{N} \phi(f_S(\mathbf{x}_i) - f_S(\mathbf{x}_k))} \tag{8}$$

where ϕ denotes L2 norm distance.The similarity guidance loss is given by:

$$\mathcal{L}_{\mathrm{SG}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} s_{ij}\left(r_{ij}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(1 - s_{ij}\right)\left[\delta - r_{ij}\right]_+^2 \tag{9}$$

where n is the number of samples in the batch, and δ is the margin.

Mutual Alignment While effective discrimination between categories is achieved within the sketch and image domains, there remains room for enhancing the consistency of their distributions. The clustering outcomes within their respective domains partly reflect these distributions. However, variations in intra-class distributions continue to hinder the establishment of semantic consistency between sketches and images.

Our approach to mutual alignment delves deeper into exploring inter-instance relationships for cross-domain alignment. In an ideal scenario free from domain shifts, relationships among arbitrary instances should exhibit domain-invariant traits between the sketch and image domains. Consequently, we advocate for aligning both intra-class and inter-class distributions, aiming to minimize cross-domain disparities in in-domain distances among instances. To measure inter-instance distance, we utilize cosine similarity:

$$d(\mathbf{x}_i, \mathbf{x}_j, M) = sim(\psi(f_S(\mathbf{x}_i)^\top M/\tau), \psi(f_S(\mathbf{x}_j)^\top M/\tau))) \tag{10}$$

The inter-instance distance for sketches S_i and S_j can be represented as $d(S_i, S_j, M^S)$, and the cross-domain distance can be represented as $d(S_i, S_j, M^I)$.Likewise, the same applies to images I_i and I_j.The objective of domain-invariant learning can be formulated as:

$$\mathcal{L}_{\mathrm{MA}} = \sum_{\mathcal{D}'\in\{\mathcal{I}^S, \mathcal{S}^S\}} \sum_{\mathbf{x}\in\mathcal{D}'} \phi(d(\mathbf{x}_i, \mathbf{x}_j, M^S), d(\mathbf{x}_i, \mathbf{x}_j, M^I)) \tag{11}$$

Methods	Sketchy Ext1		Sketchy Ext2		TU-Berlin Ext		Quickdraw	
	mAP	Prec@100	mAP@200	Prec@200	mAP	Prec@100	mAP	Prec@200
MoCo v2(CVPR'2020)	0.161	0.203	0.127	0.212	0.157	0.247	0.053	0.059
DINO(ICCV'2021)	0.197	0.252	0.215	0.335	0.177	0.273	0.057	0.064
CDS(ICCV'2021)	0.194	0.243	0.128	0.206	0.150	0.239	0.053	0.053
PCS(CVPR'2021)	0.184	0.257	0.145	0.227	0.159	0.254	0.056	0.059
FRL(ECCV'2022)	0.198	0.252	0.187	0.274	0.182	0.276	0.054	0.063
SD-Base	0.457	0.587	0.393	0.509	0.353	0.517	0.101	0.173
AMA(Ours)	**0.548**	**0.684**	**0.491**	**0.585**	**0.429**	**0.592**	**0.112**	**0.192**

Table 1: Experimental results on the three datasets for UZS-SBIR. SD-Base denotes the self-distillation paradigm.

Models	Sketchy Ext1	Sketchy Ext2	TU-Berlin Ext
SD-Base	0.457	0.377	0.353
AMA w/o $\mathcal{L}_{SD}$	0.165	0.191	0.146
AMA w/o $\mathcal{L}_{CC}$	0.498	0.411	0.372
AMA w/o $\mathcal{L}_{SG}$	0.506	0.458	0.394
AMA w/o $\mathcal{L}_{MA}$	0.493	0.434	0.346
Full AMA	**0.548**	**0.507**	**0.429**

Table 2: Ablation results (mAP) for each loss term on Sketchy Ext1, Sketchy Ext1, and TU-Berlin Ext.

Overall Objective

Finally, the overall objective of our AMA for UZS-SBIR is given as follows:

$$\mathcal{L} = \mathcal{L}_{SD} + \beta_3 \mathcal{L}_{CC} + \beta_4 \mathcal{L}_{SG} + \beta_5 \mathcal{L}_{MA} \quad (12)$$

where β_3, β_4, and β_5 are hyperparameters that balance the objective function's different parts.

Experiments

Datasets We validate our method on three commonly used benchmark datasets in ZS-SBIR: Sketchy Ext (Liu et al. 2017), TuBerlin Ext (Zhang et al. 2016), and Quickdraw (Dey et al. 2019). **Sketchy Ext** is an extended version of Sketchy (Sangkloy et al. 2016), which contains 75,471 sketches belonging to 125 categories and 100 images per category. It expands on Sketchy by incorporating 60,502 additional images from ImageNet (Russakovsky et al. 2015), resulting in 73,002 images. We used two types of splits (Shen et al. 2018; Yelamarthi et al. 2018), denoted as Ext1 and Ext2. In Sketchy Ext1, we randomly selected 25 classes for testing and used the remaining 100 classes for training. In Sketchy Ext2, we used 21 unseen classes not present in ImageNet for testing, with the remaining 104 classes used for training. **Tu-Berlin Ext** consists of 250 categories, with 80 sketches per category and 204,489 images expanded based on Tu-Berlin (Eitz, Hays, and Alexa 2012). Follow (Liu et al. 2019), our testing includes 30 randomly selected categories, while the remaining 220 classes are used for testing. **Quickdraw** is the largest SBIR dataset. It consists of 330,000 sketches and 204,000 images belonging to 110 categories. We utilized the default split for our training and testing on Quicdraw, where 80 classes were used for training, and 30 classes were used for the test.

Implementation Details Our method is implemented using the popular PyTorch toolbox, and we adopt VIT-B/16 (Dosovitskiy et al. 2021) as our feature extractor. To ensure the entire training process is fully unsupervised, we initialize the backbone and head layer parameters using the DINO (Caron et al. 2021) model pre-trained on the unlabeled ImageNet dataset. The initial learning rate is set to 0.0002. We train the model for 50 epochs using the SGD optimizer with a momentum of 0.9 and a batch size 32. The learning rate is gradually decayed to 0 using a cosine annealing schedule. The temperature τ is always 0.1. The filter parameter μ is 0.2. The number of multi-crops V and global views V_1 were set to 10 and 2. The global and local views were resized to resolutions of 224^2 and 96^2, respectively, before being fed into the model. During the training process, β_1 and β_2 are consistently set to 1.0 and 2.0. Particular attention should be paid to the adjustments of the hyperparameters β_3, β_4, and β_5. They are initialized as 0. β_3 is positively correlated with the reliability of clustering, linearly increasing to 0.2 between 20 and 40 epochs.β_4 and β_5 set explicitly to 0.02 and 0.1 after 20 and 40 epochs, respectively.

Evaluation Metrics Consistent with recent ZS-SBIR literature (Lin et al. 2023; Sain et al. 2023), we utilized mean average precision (mAP), top-200 mean average precision (mAP@200), top-100 precision (prec@100), 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.

Baselines We use the following works as the baselines to evaluate our proposed method. **MoCo v2** (Chen et al. 2020b) achieves self-supervised learning through instance contrastive learning and momentum encoders, enabling effective discrimination between instances. **DINO** (Caron et al. 2021) is a fully self-distilled method that does not rely on the contrastive relationships between instances. It employs the vision transformers as the backbone for multi-crop training. **CDS** (Kim et al. 2021a) provides cross-domain self-supervised pretraining, which includes within-domain in-

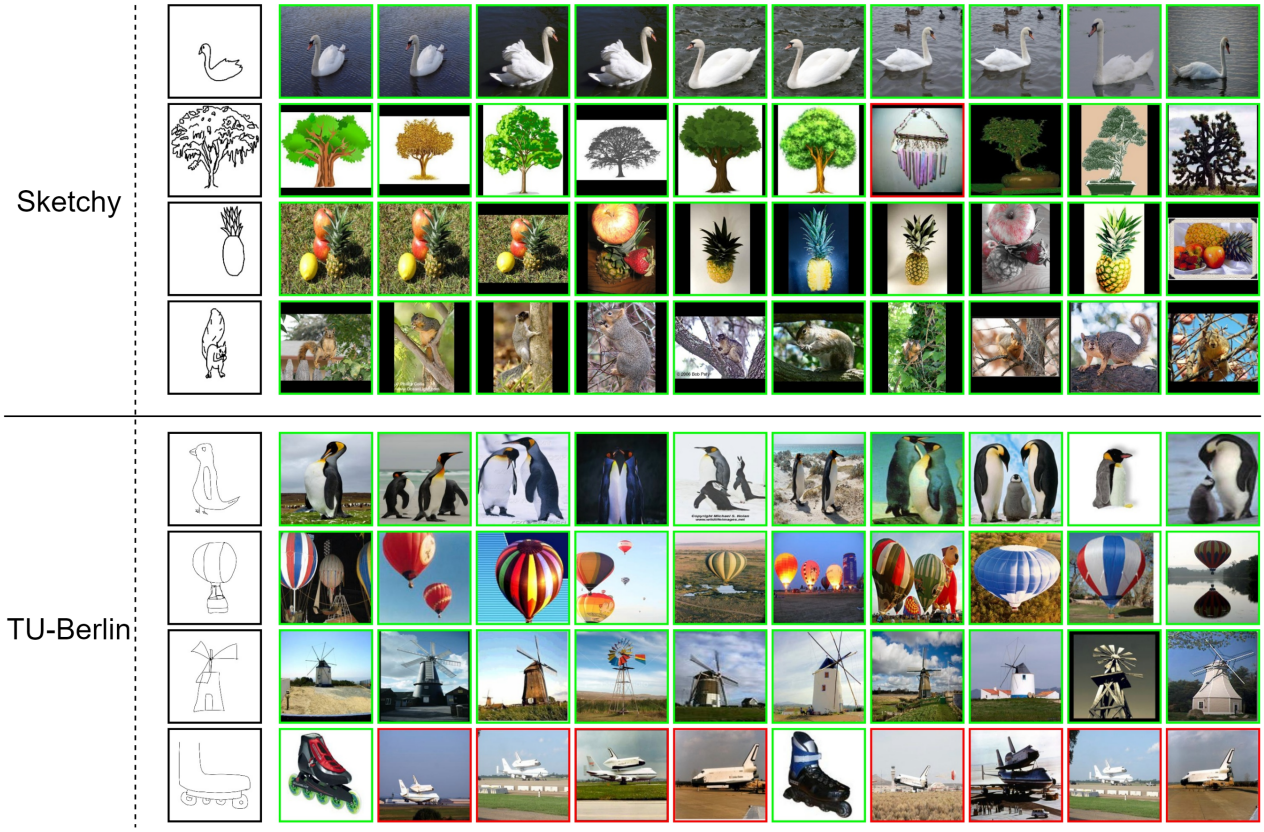


Figure 3: The Top-10 retrieval examples of our AMA for UZS-SBIR. Green and red bounding boxes indicate correct retrieval results and false positives, respectively.

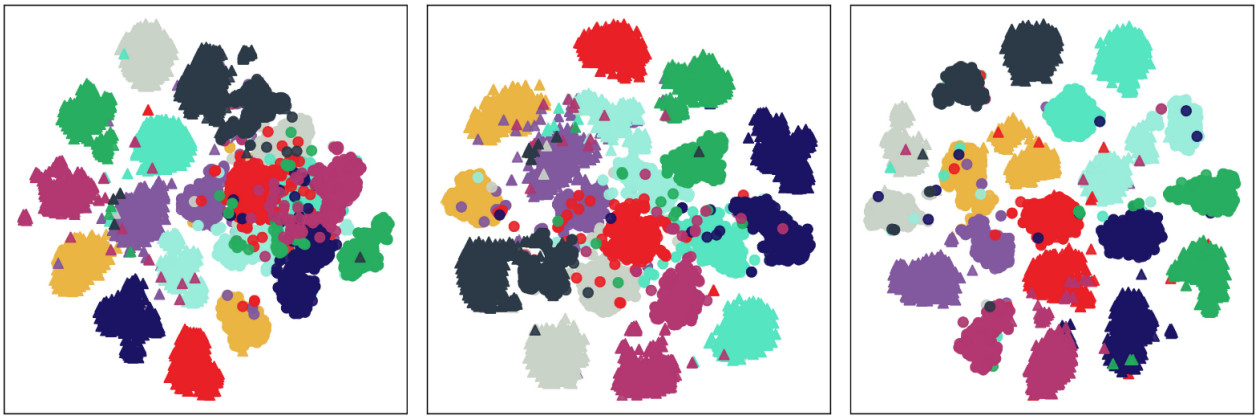
stance discrimination and cross-domain alignment through entropy minimization. **PCS** (Yue et al. 2021) incorporates prototypical contrastive learning into unsupervised instance discrimination for cross-domain matching. **FRL** (Yue et al. 2021) is an unsupervised cross-domain retrieval method that employs cluster-wise contrastive learning for intra-domain feature learning and measures the cross-domain distance to achieve category alignment.

Quantitative Analysis

We report our experimental results on datasets Sketchy Ext1, Sketchy Ext2, Tu-Berlin Ext, and Quickdraw in Table 1. It can be observed that our method achieves superior performance compared to other methods evaluated on the UZS-SBIR task. In traditional supervised ZS-SBIR, Sketchy Ext1 and TU-Berlin Ext datasets are typically considered essential datasets, where the unseen categories are randomly selected. Specifically, compared to both cross-domain retrieval and self-supervised pretraining methods on these two datasets, AMA achieves over 20% improvements in mAP scores and over 30% improvements in Prec@100 scores. Moreover, AMA still exhibits an 8%$\sim$10% improvement compared to our self-distillation paradigm on these metrics. Sketchy Ext2 and Quickdraw are relatively challenging datasets since they ensure unseen classes do not overlap with the categories in ImageNet. On Sketchy Ext2, AMA achieves a twofold increase in MAP and precision metrics over other unsupervised methods. All the methods experimented on Quickdraw face difficulties in achieving good performance, primarily due to the sheer volume of data and the high level of abstraction in the sketches. AMA achieves a score improvement of 5.5% and 12.8% on metrics mAP and Prec@200, respectively. All these comparisons can demonstrate that our AMA effectively leverages category semantics and domain generalization information from unlabeled sketches and images, enabling cross-domain alignment and category transfer between sketches and images.

Qualitative Analysis

Figure 3 illustrates a sample retrieval example of AMA. It demonstrates the top 10 retrieved candidate images for each sketch, with the majority of them belonging to the same category as the query sketch. Furthermore, we can observe that neither complex backgrounds nor the presence of multiple objects significantly affect the retrieval performance. However, suppose the shape and fine-grained characteristics of the target closely resemble those of the query image. In that case, there is a high likelihood of being retrieved as a false positive, as exemplified by the red-boxed candidates in the second and last rows of the figure. The reasons for erroneous



(a) DINO (b) SD-Base (c) AMA (Ours)

Figure 4: The t-SNE visualization of results of the learned feature embeddings on Sketchy Ext1, with colored circles (◯) representing images and colored upper triangles (△) representing sketches, using distinct colors to indicate different categories.

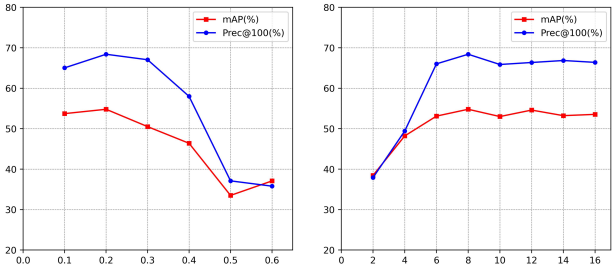
retrieval are likely due to the sparsity of the sketch itself and the scarcity of crucial fine-grained information.

Ablation Study

Performance Contributions of Different Components We explored the contribution of each respective loss term in Eq.12 by ablating it during the training process. The ablation experiment results of datasets Sketchy Ext1, Sketchy Ext2, and TU-Berlin Ext are presented in Table 2, where "w/o" denotes the ablation behavior. We can see: 1) Without $\mathcal{L}_{SD}$, the performance drops dramatically, indicating the importance of considering the self-distillation paradigm as the foundation. 2) The results of AMA w/o $\mathcal{L}_{CC}$ and AMA w/o $\mathcal{L}_{SG}$ show that contrastive clustering and similarity guidance exhibit remarkable synergistic effects with the asymmetric parameter update scheme of self-distillation. 3) The comprehensive integration of all loss functions enables the entire model to achieve optimal results, leveraging each loss's advantages.

Analysis on Multi-Views Augmentation As shown in figure 5, we analyze the impact of global and local view numbers on the model's performance. The result indicates that considering a greater number of local views than the global view count can lead to improved outcomes. Furthermore, as this quantity increases, the performance initially improves and stabilizes.

Visualization Comparison Figure 4 presents the visualization of the embeddings of our approach on the Sketchy Ext1 dataset using the t-SNE (Van der Maaten and Hinton 2008) algorithm. We can observe that for single-domain data, both sketches and images of the same category are well-clustered. However, since our method does not rely on labels, the sketches and authentic images of the same category can only maintain a relatively close distance instead of achieving a complete cross-domain fusion. Fortunately, this relative distance still enables them to be separated as a distinct cluster from other sketch-image clusters. This result indicates that our method positively facilitates the alignment between the sketch domain and the image domain.



(a) The proportion of global views with all views (b) The number of local views

Figure 5: The performance on Sketchy Ext1 with different values of global and local views numbers. The number of all views is set to 10 in (a) and the number of global view is set to 2 in (b).

Conclusion

In this paper, we delve into the Unsupervised Zero-Shot Sketch-Based Image Retrieval problem, introducing an asymmetric mutual alignment approach for cross-domain alignment and category transfer between sketches and images. Firstly, we summarize the Self-distillation Paradigm, which enhances the feature expressiveness of images and sketches through multi-views learning. Secondly, we design teacher-guided contrastive clustering and Semantic Similarity Guidance to perform self-exploration on category and instance levels. Finally, we define consistency constraints for achieving Mutual Alignment between sketches and images. Our experimental results on four dataset partitions demonstrate the effectiveness of our approach. We will investigate better domain alignment strategies for sketches and images in the future.

Acknowledgments

Our work was supported by Joint Fund of Ministry of Education of China (8091B022149), Key Research and Development Program of Shaanxi (2021ZDLGY01-03), National Natural Science Foundation of China (62132016, 62171343, 62071361, and 62302372), and Fundamental Research Funds for the Central Universities (ZDRC2102).

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. Beit: Bert pre-training of image transformers. In *ICLR*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 33: 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Chaudhuri, A.; Bhunia, A. K.; Song, Y.-Z.; and Dutta, A. 2023. Data-Free Sketch-Based Image Retrieval. In *CVPR*, 12084–12093.

Chen, S.; Gong, C.; Li, J.; Yang, J.; Niu, G.; and Sugiyama, M. 2022. Learning Contrastive Embedding in Low-Dimensional Space. *NeurIPS*, 35: 6345–6357.

Chen, S.; Gong, C.; Li, X.; Yang, J.; Niu, G.; and Sugiyama, M. 2023. Boundary-restricted metric learning. *MACH LEARN*, 1–40.

Chen, S.; Gong, C.; Yang, J.; Tai, Y.; Hui, L.; and Li, J. 2019. Data-adaptive metric learning with scale alignment. In *AAAI*, 3347–3354.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

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

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

Dutta, A.; and Akata, Z. 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 5089–5098.

Eitz, M.; Hays, J.; and Alexa, M. 2012. How do humans sketch objects? *ACM T GRAPHIC*, 31(4): 1–10.

Eitz, M.; Hildebrand, K.; Boubekeur, T.; and Alexa, M. 2010. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *COMPUT GRAPH-UK*, 34(5): 482–498.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *CVPR*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

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

Kim, D.; Saito, K.; Oh, T.-H.; Plummer, B. A.; Sclaroff, S.; and Saenko, K. 2021a. Cds: Cross-domain self-supervised pre-training. In *ICCV*, 9123–9132.

Kim, S.; Kim, D.; Cho, M.; and Kwak, S. 2021b. Embedding transfer with label relaxation for improved metric learning. In *CVPR*, 3967–3976.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3): 453–465.

Lin, F.; Li, M.; Li, D.; Hospedales, T.; Song, Y.-Z.; and Qi, Y. 2023. Zero-Shot Everything Sketch-Based Image Retrieval, and in Explainable Style. In *CVPR*, 23349–23358.

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

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

Qi, Y.; Song, Y.-Z.; Zhang, H.; and Liu, J. 2016. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, 2460–2464. IEEE.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115: 211–252.

Sain, A.; Bhunia, A. K.; Chowdhury, P. N.; Koley, S.; Xiang, T.; and Song, Y.-Z. 2023. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *CVPR*, 2765–2775.

Sain, A.; Bhunia, A. K.; Potlapalli, V.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2022. Sketch3t: Test-time training for zero-shot sbir. In *CVPR*, 7462–7471.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM T GRAPHIC*, 35(4): 1–12.

Shen, Y.; Liu, L.; Shen, F.; and Shao, L. 2018. Zero-shot sketch-image hashing. In *CVPR*, 3598–3607.

Song, J.; Yu, Q.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 5551–5560.

Tian, J.; Xu, X.; Shen, F.; Yang, Y.; and Shen, H. T. 2022. Tvt: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In *AAAI*, 2370–2378.

Tian, J.; Xu, X.; Wang, Z.; Shen, F.; and Liu, X. 2021. Relationship-preserving knowledge distillation for zero-shot sketch based image retrieval. In *ACM MM*, 5473–5481.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *J MACH LEARN RES*, 9(11).

Wang, H.; Deng, C.; Liu, T.; and Tao, D. 2021. Transferable coupled network for zero-shot sketch-based image retrieval. *IEEE TPAMI*, 44(12): 9181–9194.

Wang, K.; Wang, Y.; Xu, X.; Liu, X.; Ou, W.; and Lu, H. 2022. Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval. In *ACM MM*, 601–609.

Wang, X.; Peng, D.; Yan, M.; and Hu, P. 2023. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. In *AAAI*.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, 41(9): 2251–2265.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *CVPR*, 9653–9663.

Yan, J.; Luo, L.; Deng, C.; and Huang, H. 2023a. Adaptive hierarchical similarity metric learning with noisy labels. *IEEE TIP*, 32: 1245–1256.

Yan, J.; Yang, E.; Deng, C.; and Huang, H. 2022. MetricFormer: A Unified Perspective of Correlation Exploring in Similarity Learning. *NeurIPS*, 35: 33414–33427.

Yan, J.; Yin, Z.; Yang, E.; Yang, Y.; and Huang, H. 2023b. Learning with Diversity: Self-Expanded Equalization for Better Generalized Deep Metric Learning. In *ICCV*, 19365–19374.

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

Yu, Q.; Liu, F.; Song, Y.-Z.; Xiang, T.; Hospedales, T. M.; and Loy, C.-C. 2016. Sketch me that shoe. In *CVPR*, 799–807.

Yue, X.; Zheng, Z.; Zhang, S.; Gao, Y.; Darrell, T.; Keutzer, K.; and Vincentelli, A. S. 2021. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *CVPR*, 13834–13844.

Zhang, H.; Liu, S.; Zhang, C.; Ren, W.; Wang, R.; and Cao, X. 2016. Sketchnet: Sketch classification with web images. In *CVPR*, 1105–1113.

Zhang, Z.; Zhang, Y.; Feng, R.; Zhang, T.; and Fan, W. 2020. Zero-shot sketch-based image retrieval via graph convolution network. In *AAAI*, 12943–12950.