Domain-Smoothing Network for Zero-Shot Sketch-Based Image Retrieval

Zhipeng Wang , **Hao Wang** , **Jiexi Yan** , **Aming Wu*** and **Cheng Deng***

Xidian University

{zpwang1996, haowang.xidian}@gmail.com, jxyan@stu.xidian.edu.cn, amwu@xidian.edu.cn, chdeng.xd@gmail.com

Abstract

Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) is a novel cross-modal retrieval task, where abstract sketches are used as queries to retrieve natural images under zero-shot scenario. Most existing methods regard ZS-SBIR as a traditional classification problem and employ a cross-entropy or triplet-based loss to achieve retrieval, which neglect the problems of the domain gap between sketches and natural images and the large intra-class diversity in sketches. Toward this end, we propose a novel Domain-Smoothing Network (DSN) for ZS-SBIR. Specifically, a cross-modal contrastive method is proposed to learn generalized representations to smooth the domain gap by mining relations with additional augmented samples. Furthermore, a category-specific memory bank with sketch features is explored to reduce intra-class diversity in the sketch domain. Extensive experiments demonstrate that our approach notably outperforms the state-of-the-art methods in both Sketchy and TU-Berlin datasets.

1 Introduction

Sketch-based image retrieval (SBIR) [Sangkloy *et al.*, 2016; Bhunia *et al.*, 2020] is an interesting and practical task aiming to retrieve similar images based on the queries of the hand-drawn sketches. With the help of large-scale labeled datasets, SBIR can achieve astonishing performance. However, collecting and annotating large-scale datasets is time-consuming and labor-intensive. To this end, zero-shot SBIR (ZS-SBIR) [Shen *et al.*, 2018] is proposed, which aims to retrieve categories or samples unseen at the training stage. However, solving ZS-SBIR is significantly challenging since it simultaneously deals with the inherent domain gap and limited knowledge about the unseen categories [Deng *et al.*, 2020].

Previous ZS-SBIR methods [Xu *et al.*, 2020] attempt to learn a common representation of sketches and images with cross-entropy loss or triplet-based loss, which neglect the impact of the inherent domain gap. Specifically, cross-entropy loss equally treats sketches and images for classification and
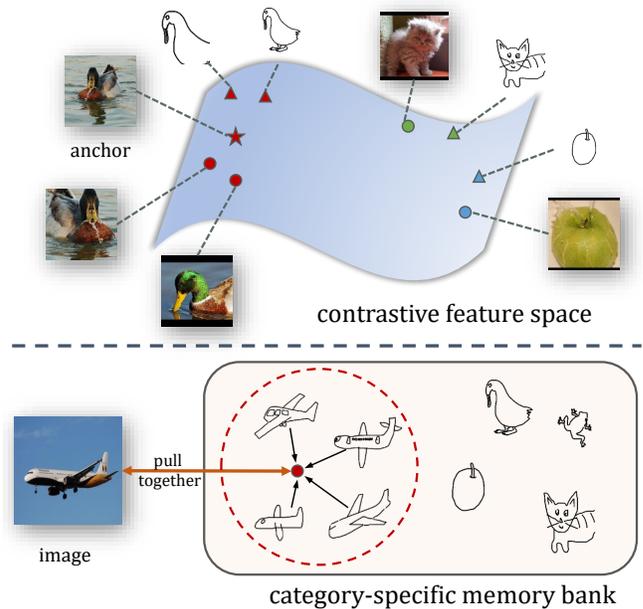
---

*Contact Author

Figure 1: The illustration of our DSN method. (Top) To smooth the domain gap between sketches and images, we propose a cross-modal contrastive method to mine the relations between samples with sufficiently positive and negative samples. (Bottom) Besides, a category-specific memory bank is built to address the large intra-class diversity in the sketch domain, where prototypes of each sketch category are learned to reduce the variability in appearance.

neglects the relations between sketches and images of the same category. Besides, triplet-based loss is usually utilized to measure the relations between samples. However, due to the limited positive and negative samples, triplet-based loss may not mine the relations sufficiently. Furthermore, sketches within the same category might be extremely diverse in appearance due to the drawing styles of different people, which is ignored by most methods and degrades retrieval performance.

To alleviate the domain gap and the intra-class diversity in sketch domain, we propose a Domain-Smoothing Network (DSN) for ZS-SBIR (seen Fig. 1), mainly including a cross-modal contrastive method and a category-specific memory bank. Specifically, a pre-trained CNN backbone is first ap-

plied to extract visual features of the images and sketches. To smooth the domain gap between the sketches and images, a cross-modal contrastive method is then conducted after a projection head for visual features. Meanwhile, a category-specific memory bank is proposed, where sketch category prototypes are learned to reduce the intra-class diversity in the sketch domain. Besides, to enhance the discrimination of obtained representations, our model is trained with the knowledge of the teacher model and label information inspired by previous work on teacher signal [Liu *et al.*, 2019]. To validate the superiority of our DSN method, extensive experiments are conducted and the results demonstrate our method remarkably outperforms the state-of-the-art methods on two popular datasets.

The main contributions of this work are summarized as:

- We propose the DSN model for ZS-SBIR task, where a cross-modal contrastive method is utilized to mine the relations between sketches and images with sufficient positive and negative samples to significantly smooth the domain gap.

- To address the large intra-class diversity in the sketch domain, we elaborate a category-specific memory bank to learn prototypes for each sketch category to reduce the variability in appearance.

- Extensive experiments conducted on two popular datasets, *i.e.*, Sketchy and TU-Berlin, demonstrate that our approach outperforms the state-of-the-art methods by a large margin.

2 Related Work

Since ZS-SBIR is an extremely challenging task integrated with SBIR and zero-shot learning (ZSL), we will review both of them in this section.

2.1 Sketch-Based Image Retrieval

The main target of SBIR task is to alleviate the domain gap between image and sketch by aligning their features in a common semantic space. Early works utilize hand-crafted features to represent sketches and match them with the edge maps of images by various Bag-of-Words model. However, it is difficult to align the edge map with hand-crafted features directly, which makes the problem of domain gap unresolved. With the advent of deep neural networks, researchers tend to employ CNN to extract deep features from images and sketches to learn transferrable representation. Among them, Siamese architecture based methods [Qi *et al.*, 2016] play an important role by handling sketches and images separately with ranking loss. Compared with these methods, our DSN model proposes a cross-modal contrastive method to address the domain gap effectively. By pulling together samples with the same label and pushing apart those from different categories in different modalities and views, the cross-modal contrastive method can mine the relations between sketches and images sufficiently to smooth the inherent domain gap.

2.2 Zero-Shot Learning

Zero-shot learning (ZSL) aims to recognize objects whose instances are unseen during the training period [Xian *et*

al., 2018a; Wei *et al.*, 2019; Wei *et al.*, 2020a]. Existing ZSL methods can be classified into two categories: embedding-based [Wei *et al.*, 2020b] and generative-based [Li *et al.*, 2021] approaches. For the embedding-based method, [Kodirov *et al.*, 2017] learned an auto-encoder to map visual features to semantic embeddings, where the nearest neighbor search is conducted. For the generative-based method, [Xian *et al.*, 2018b] proposed a feature generating network to synthesize unseen visual features to mitigate the data imbalance problem in ZSL. Besides, various semantic representation methods have been proposed, including attributes annotations [Akata *et al.*, 2015] and word vectors [Mikolov *et al.*, 2013] to bridge the seen and unseen categories. Due to the space limitation, more details of ZSL can refer to [Xian *et al.*, 2018a].

2.3 Zero-Shot Sketch-Based Image Retrieval

ZS-SBIR is a novel task that conducts SBIR under zero-shot setting and is more practical in realistic applications. [Shen *et al.*, 2018] first combined ZSL with SBIR, and proposed a cross-modal hashing method for ZS-SBIR. [Yelamarthi *et al.*, 2018] built an auto-encoder to generate image features by sketches for alignment, while [Dutta and Akata, 2019] proposed a paired cycle-consistent generative model with adversarial training to align the sketch and image in the same semantic space. [Dutta and Biswas, 2019] presented a feature disentanglement method, where the images generated by sketches and corresponding styles can be conducted in retrieval as queries. [Liu *et al.*, 2019] treated ZS-SBIR as a catastrophic forgetting problem and employed a CNN model pre-trained on ImageNet as the teacher model to preserve the knowledge. However, all these methods ignore the large intra-class diversity in the sketch domain. In this work, we propose a category-specific memory bank to learn prototypes for each sketch category to reduce the intra-class diversity in the sketch domain.

3 Methodology

3.1 Problem Definition

In ZS-SBIR, the dataset can be divided into two parts: the training set and the testing set. The training set, denoted as $\mathcal{D}^{seen} = \{\mathcal{I}^{seen}, \mathcal{S}^{seen}\}$, is utilized for training the retrieval model, where I^{seen} and S^{seen} represent images and sketches from seen categories, respectively. Similarly, the testing set is utilized for validating the performance of the retrieval model, which can be denoted as $\mathcal{D}^{unseen} = \{\mathcal{I}^{unseen}, \mathcal{S}^{unseen}\}$. Mathematically, let $\mathcal{I}^{seen} = \{(I_j, y_j)|y_j \in \mathcal{C}^{seen}\}_{j=1}^{N_1}$ and $\mathcal{S}^{seen} = \{(S_i, y_i)|y_i \in \mathcal{C}^{seen}\}_{i=1}^{N_2}$, where y represents the category label. $\mathcal{C}^{seen}$ denotes the seen category set. N_1 and N_2 denote the number of images and sketches in training set, respectively.

During the training period, we sample batches of images and sketches from the training set, where corresponding labels are utilized to guide the model to learn discriminative features. And at testing stage, a sketch query $S' \in \mathcal{S}^{unseen}$ with its label $y \in \mathcal{C}^{unseen}$, where $\mathcal{C}^{unseen}$ denotes the unseen category set, is given to match the similar image $I' \in \mathcal{I}^{unseen}$. If the label y' of the retrieved image is the same as
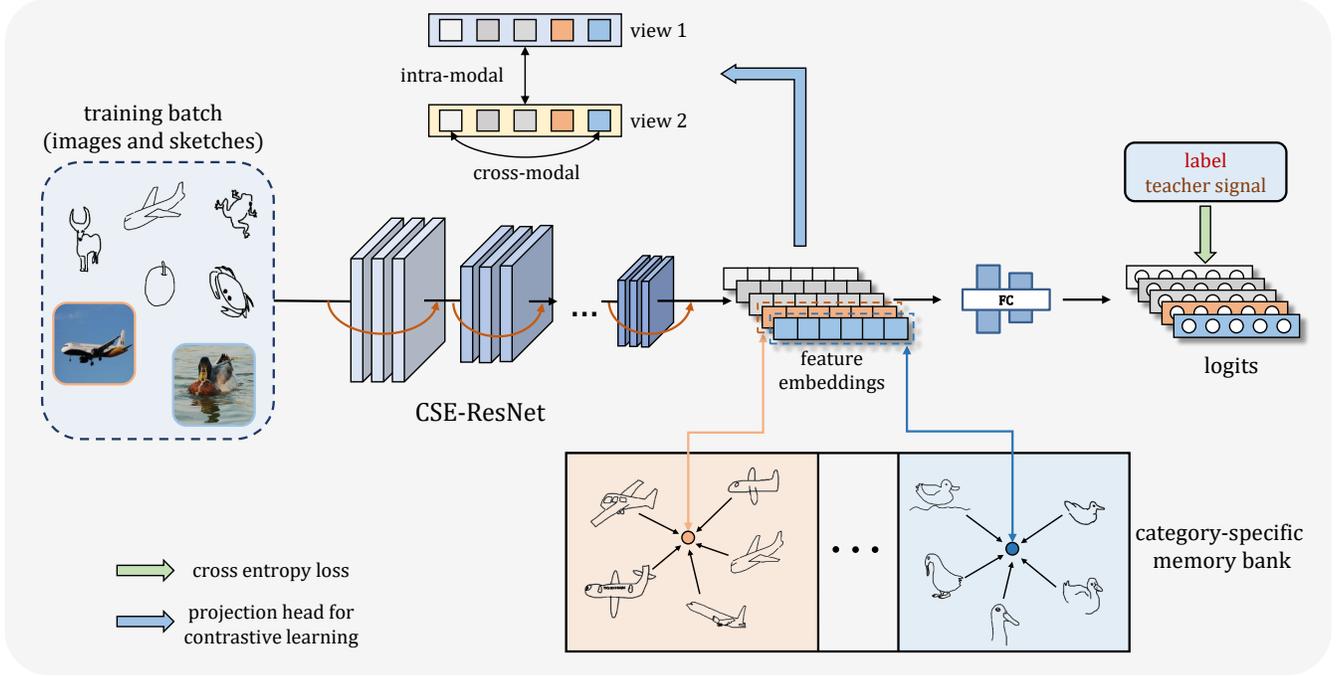
Figure 2: An overview of our model. We use a CSE-ResNet-50 as the backbone to learn a common representation for input sketches and images. After obtaining feature embeddings, three branches of modules are applied. The first one is to employ cross-modal contrastive method in different views after a projection head with fully-connected layers. The second one is to enforce the image feature to be closer to the prototype of sketch features in the category-specific memory bank. The last one is to leverage the teacher signal and label information to train the whole model.

that of the given sketch, *i.e.*, $y = y'$, we name this as a successful retrieval. In zero-shot setting, there is no testing category appears in training period, *i.e.*, $\mathcal{C}^{seen} \bigcap \mathcal{C}^{unseen} = \varnothing$.

3.2 Domain-Smoothing Network

The architecture of our DSN model is illustrated in Figure 2. Specifically, we employ a CSE-ResNet-50 backbone (CSE) [Lu *et al.*, 2018] to extract feature embeddings for images and sketches. To smooth the domain gap between sketch and image, we propose a cross-modal contrastive method to align these two domains. Besides, a category-specific memory bank is built to address the large intra-class diversity in sketch domain. The details of these modules will be discussed in this section.

Cross-modal Contrastive Method

The domain gap is an inherent challenge in SBIR task. Previous methods usually use a unified framework to learn a common representation of sketches and images by cross-entropy loss or triplet-based loss. However, cross-entropy loss may neglect the relations between sketches and images as mentioned above. Besides, triplet-based loss is usually utilized to measure the relations between samples. However, due to the limited positive and negative samples, triplet-based loss may not mine the relations sufficiently.

To effectively smooth the domain gap, we propose a cross-modal contrastive method. In detail, for each image I and sketch S from $\mathcal{D}^{seen}$, two random augmentations are

generated for contrastive learning by augmentation method [Khosla *et al.*, 2020], denoted as $\{\hat{I}, \tilde{I}, \hat{S}, \tilde{S}\}$. Then CSE is applied to extract feature embeddings for each image, sketch and these augmentations as:

$$[f_{img}, \hat{f}_{img}, \tilde{f}_{img}] = [\text{CSE}(I), \text{CSE}(\hat{I}), \text{CSE}(\tilde{I})],$$
$$[f_{ske}, \hat{f}_{ske}, \tilde{f}_{ske}] = [\text{CSE}(S), \text{CSE}(\hat{S}), \text{CSE}(\tilde{S})]. \quad (1)$$

After obtaining feature embeddings f, a projection head with two fully-connected layers, named as P, is utilized for mapping f to vector $v = P(f)$. In this way, we can obtain the vectors for two augmentations of sketch and image, which can be written as $\{\hat{v}_{img}, \tilde{v}_{img}, \hat{v}_{ske}, \tilde{v}_{ske}\} \subseteq V$. To illustrate clearly, we denote $\hat{v}_{img}$ as the anchor. It should be noted that even though $\hat{v}_{img}$ and $\tilde{v}_{img}$ are in different views, they share the same label. So $\tilde{v}_{img}$ is an **intra-modal** positive sample for $\hat{v}_{img}$. Similarly, $\hat{v}_{ske}$ and $\tilde{v}_{ske}$ are **cross-modal** samples for $\hat{v}_{img}$. Their roles of positive or negative depends on whether their labels are the same as $\hat{v}_{img}$. Then contrastive method is conducted to pull together the anchor with the positive samples and push apart the anchor from the negative samples in a cross-modal scenario.

Our cross-modal contrastive method (CMCM) loss can be summarized as:

$$\mathcal{L}_{cmcm} = \sum_{i=1}^{|V|} \left[\frac{-1}{2N_i - 1} \sum_{j=1}^{2|V|} \mathbf{1}_{y_i = y_j} \log \frac{\exp(v_i \cdot v_j / \tau)}{\sum_{k=1}^{2|V|} \mathbf{1}_{i \neq k} \exp(v_i \cdot v_k / \tau)} \right], \quad (2)$$

where for each vector $v \in V$, there is a corresponding label y. $|V|$ represents the cardinality of the vector set V, and $N_i =$

$|\{y_j|y_j = y_i, j \neq i, v_i \in V, v_j \in V\}|$. $\mathbf{1}_{\text{condition}} \in \{0, 1\}$ is an indicator function evaluating to 1 iff condition is true. τ is the temperature parameter for contrastive learning.

Category-specific Memory Bank

To reduce the large intra-class diversity in the sketch domain, we propose a category-specific memory bank to learn a prototype representation for sketches from the same category, where the image is enforced to be closer to the prototype of top-k similar sketches at the feature level.

Specifically, for each training batch, we can obtain a batch of feature embeddings $\{\mathcal{F}_{img}, \mathcal{F}_{ske}\}$, where $\mathcal{F}_{img} = \{f_{img}|y_{img} \in \mathcal{C}_{seen}\}$ and $\mathcal{F}_{ske} = \{f_{ske}|y_{ske} \in \mathcal{C}_{seen}\}$. In order to update the memory bank and leverage the history knowledge at the same time, we maintain a stack filled with k sketch features for each category adaptively. First, for f_{ske} and its label y_{ske}, we search for the corresponding image features $\mathcal{F}'_{img}$ with the same label in a batch, where $\mathcal{F}'_{img} = \{f_{img}|y_{img} = y_{ske}\}$. The mean value of corresponding image features f'_{img} can be obtained as:

$$f'_{img} = \frac{1}{|\mathcal{F}'_{img}|} \sum_{f_{img} \in \mathcal{F}'_{img}} f_{img}. \quad (3)$$

Next, we calculate the cosine similarity between f'_{img} and $\mathcal{F}'_{ske} \cup \{f_{ske}\}$ to select new top-k similar sketch features and store them in the memory bank, where $\mathcal{F}'_{ske}$ is the previous top-k sketch features.

After updating the memory bank, the prototype of new top-k sketch features can be obtained as:

$$p_{ske} = \frac{1}{|\mathcal{M}_{ske}|} \sum_{f_{ske} \in \mathcal{M}_{ske}} f_{ske}, \quad (4)$$

where $\mathcal{M}_{ske}$ denotes the new top-k sketch features. Then the memory loss can be formulated as:

$$\mathcal{L}_{ml} = -\frac{f_{img}^{\top} \cdot p_{ske}}{|f_{img}| \cdot |p_{ske}|}, \quad (5)$$

where $|\cdot|$ indicates the norm of the feature vector.

Discriminative Learning

To learn discriminative representations, we use cross-entropy loss to align the learned feature with its label:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N} \log \frac{\exp(\alpha_{y_i}^{\top} f_i + \beta_{y_i})}{\sum_{j \in \mathcal{C}^{seen}} \exp(\alpha_j^{\top} f_i + \beta_j)}, \quad (6)$$

where α and β is the weight and bias of the classifier, and N denotes the number of samples in a training batch.

To bridge the seen and unseen categories for zero-shot setting, auxiliary semantic knowledge should be introduced to transfer the knowledge learned from the seen category. Inspired by [Liu *et al.*, 2019], we employ a teacher model pre-trained on ImageNet to help our DSN model learn transferrable representations.

In detail, we utilize the prediction of the teacher model $p \in \mathbb{R}^{|\mathcal{C}^T|}$ as the label information instead of one-hot vector used in cross-entropy loss, where $\mathcal{C}^T$ denotes the label space of the teacher model and $\mathcal{C}^T \cap \mathcal{C}^{unseen} = \varnothing$ in our zero-shot setting. Finally, the auxiliary semantic knowledge (ASK) loss can be formulated as:

$$\mathcal{L}_{ask} = -\sum_{i=1}^{N} \sum_{k \in \mathcal{C}_T} p'_{i,k} \log \frac{\exp(\gamma_k^{\top} f_i + \delta_k)}{\sum_{j \in \mathcal{C}^T} \exp(\gamma_j^{\top} f_i + \delta_j)}, \quad (7)$$

where p' denotes the teacher signal p with semantic constraints proposed in [Liu *et al.*, 2019] and γ, δ are weight and bias of the another classifier, respectively.

The total discriminative learning loss can be formulated as:

$$\mathcal{L}_{dl} = \mathcal{L}_{cls} + \mathcal{L}_{ask}. \quad (8)$$

3.3 Objective and Optimization

In summary, the full objective function of our DSN model is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cmcm} + \lambda_2 \mathcal{L}_{ml} + \lambda_3 \mathcal{L}_{dl}, \quad (9)$$

where λ_1, λ_2 and λ_3 are the hyperparameters to balance the performance of our model.

4 Experiments

4.1 Datasets

To verify the effectiveness of our method, We evaluate our DSN method on two popular SBIR datasets, *i.e.*, Sketchy [Sangkloy *et al.*, 2016] and TU-Berlin [Eitz *et al.*, 2012].

Sketchy. It consists of 75,471 sketches and 12,500 images from 125 categories. [Liu *et al.*, 2017] extended the image gallery with additional 60,502 photos so there are 73,002 samples in extended Sketchy dataset totally.

TU-Berlin. It consists of 20,000 sketches uniformly distributed over 250 categories. [Liu *et al.*, 2017] extended this dataset with 204,489 natural images.

For comparison, we follow the split of [Shen *et al.*, 2018] to randomly choose 25 categories and the rest 100 categories for testing and training in Sketchy. For TU-Berlin, 30 categories and the rest 220 categories are chosen for testing and training. Also, we follow [Shen *et al.*, 2018] to select at least 400 images for each category in the testing set.

4.2 Implementation Details

We implement our method with PyTorch on two TITAN RTX GPUs. In our model, we utilize a CSE-ResNet-50 pre-trained on ImageNet [Lu *et al.*, 2018] to initialize the backbone. Inspired by [Liu *et al.*, 2019], we use the same pre-trained backbone to provide teacher signal as auxiliary semantic knowledge. In the training stage, we train our model with Adam optimizer. The initial learning rate is set to 0.0001 and exponentially decays to $1e^{-7}$ in training. The batch size in our original setting is 96 and the training epochs is 10. Usually, we set hyperparameters $\lambda_1 = 0.1$, $\lambda_2 = 1$ and $\lambda_3 = 1$ unless stated. For cross-modal contrastive method, we set the temperature τ as 0.07 and the dimension of contrastive vectors is 128 in all the experiments. For category-specific memory bank, we set $k = 10$.

Task	Methods	Dimension	Sketchy		TU-Berlin	
			mAP@all	Prec@100	mAP@all	Prec@100
SBIR	Siamese CNN [Qi et al., 2016]	64	0.132	0.175	0.109	0.141
	SaN [Yu et al., 2017]	512	0.115	0.125	0.089	0.108
	GN Triplet [Sangkloy et al., 2016]	1024	0.204	0.296	0.175	0.253
	DSH$_b$ [Liu et al., 2017]	64	0.171	0.231	0.129	0.189
	GDH$_b$ [Zhang et al., 2018]	64	0.187	0.259	0.135	0.212
ZSL	SAE [Kodirov et al., 2017]	300	0.216	0.293	0.167	0.221
	FRWGAN [Felix et al., 2018]	512	0.127	0.169	0.110	0.157
	ZSH$_b$ [Yang et al., 2016]	64	0.159	0.214	0.141	0.177
ZS-SBIR	ZSIH$_b$ [Shen et al., 2018]	64	0.258	0.342	0.223	0.294
	CVAE [Yelamarthi et al., 2018]	4096	0.196	0.284	0.005	0.001
	SEM-PCYC [Dutta and Akata, 2019]	64	0.349	0.463	0.297	0.426
	SEM-PCYC$_b$ [Dutta and Akata, 2019]	64	0.344	0.399	0.293	0.392
	SAKE$_b$ [Liu et al., 2019]	64	0.364	0.487	0.359	0.481
	CSDB [Dutta and Biswas, 2019]	64	0.376	0.484	0.254	0.355
	DSN (ours)	64	**0.484**	**0.610**	**0.442**	**0.538**
	DSN$_b$ (ours)	64	**0.436**	**0.553**	**0.385**	**0.497**
	SAKE [Liu et al., 2019]	512	0.547	0.692	0.475	**0.599**
	DSN (ours)	512	**0.583**	**0.704**	**0.481**	0.586
	DSN$_b$ (ours)	512	**0.581**	**0.700**	**0.484**	**0.591**

Table 1: The retrieval performance comparison with existing SBIR, ZSL and ZS-SBIR methods. (The best performance results are in **bold** and the subscript b denotes binary hash results).

4.3 Comparing with Existing Methods

We compare our DSN model with various methods related to SBIR, ZSL and ZS-SBIR. Especially for ZS-SBIR, we employ all the experiments under the same setting as others for fairness. In addition, to generate binary hash codes for comparison, we use iterative quantization (ITQ) [Gong et al., 2012]. Then the hamming distance is computed to conduct retrieval while cosine distance is utilized for real value feature retrieval. The results of our model for ZS-SBIR are shown in Table 1.

As we can see, our DSN method outperforms other existing methods especially in Sketchy dataset with 64-d features. Specifically, in 64-d real value feature retrieval, our DSN outperforms the state-of-the-art (SOTA) method [Dutta and Biswas, 2019; Dutta and Akata, 2019] by 29% in Sketchy and 49% in TU-Berlin dataset, respectively. In 512-d real value feature retrieval, our DSN outperforms the SOTA [Liu et al., 2019] by 7% in Sketchy and 1% in TU-Berlin. For binary hash retrieval, we conduct ITQ [Gong et al., 2012] following the same setting with [Liu et al., 2019] and outperform SOTA by 16% and 7% for Sketchy and TU-Berlin respectively in 64-d setting. Also, we provide 512-d binary hash code retrieval results in the Table 1, which is similar to 512-d real value feature results. It demonstrates that our DSN method can not only generate accurate real value features but also perform stably even in binary hash setting for ZS-SBIR.

All of these results show that our DSN method can alleviate the domain gap between sketch and domain effectively, and reduce the large intra-class diversity in sketch domain which is ignored by most existing methods.
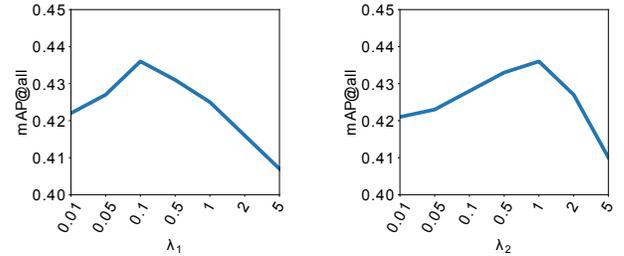

Figure 3: The 64-d binary hash retrieval results with different λ_1 for cross-modal contrastive method loss and λ_2 for memory loss in Sketchy dataset.

4.4 Experiment Analysis

Quantitative Analysis

Inspired by [Liu et al., 2019], we employ a pre-trained teacher model to provide teacher signal as auxiliary semantic knowledge. So in our experiments, we set λ_3 to 1 and explore different λ_1 and λ_2 for our proposed method.

In Figure 3, we show mAP@all results with different hyperparameters. As we can see, setting $\lambda_1 = 0.1$ and $\lambda_2 = 1$ can achieve the best performance with 64-d binary hash codes in Sketchy. It should be noted that we adopt this setting in all the experiments unless stated otherwise.

Qualitative Analysis

Visulization of Similarity. To demonstrate the superiority of our method, we also visualize the similarity between sketch and image in retrieval feature level (seen Fig. 4). Here,
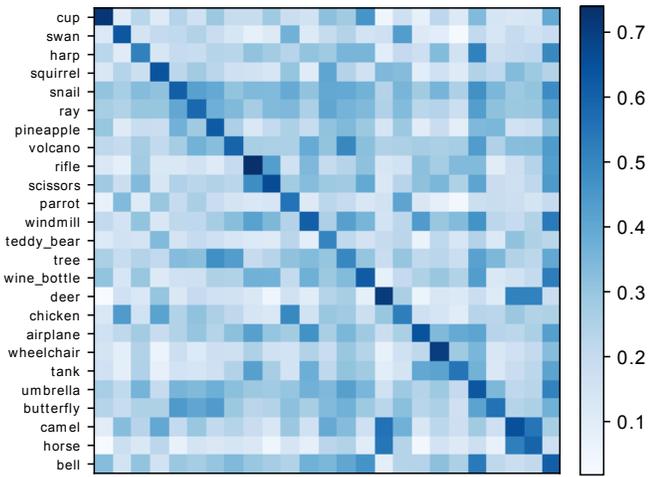
Figure 4: The similarity matrix between the sketches and images with 64-d retrieval features in the testing set of Sketchy dataset. The horizontal and vertical axes represent image and sketch categories respectively. Best view in color.
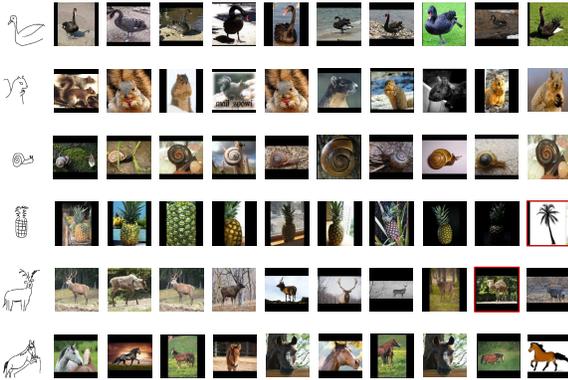

Figure 5: The top 10 zero-shot sketch-based image retrieval results using our DSN model with 512-d real value feature in Sketchy dataset. The mis-retrieved images are highlighted in red border.

we calculate the cosine similarity using the mean sketch feature and mean image feature for each category in the testing set of Sketchy. The darkness of matrix color represents the similarity. As shown in Figure 4, our DSN method smooths the domain gap to learn discriminative features, which can boost the performance of retrieval. However, negative cases still exist. For example, the camel and horse are similar in appearance, which leads to a higher similarity value in the matrix compared with other pairs of categories.

Examples of Retrieval. We show top 10 ZS-SBIR results using our DSN model with 512-d real value features in Figure 5. The mis-retrieved results are in the red border. As we can see from these examples, our DSN model performs well in most cases except for some visual similarly images. For example, in the retrieval results of pineapple sketch (the 4_{th} row in Figure 5), a tree image is mis-retrieved, which is similar to the pineapple query in appearance.

Models	Sketchy	TU-Berlin
Baseline	0.361	0.355
Baseline + $\mathcal{L}_{cmcm}$	0.420	0.373
Baseline + $\mathcal{L}_{ml}$	0.412	0.372
Baseline + $\mathcal{L}_{cmcm}$ + $\mathcal{L}_{ml}$	**0.436**	**0.385**

Table 2: The mAP@all results with 64-d binary hash codes of our ablation study. (The best results are in **bold**).

4.5 Ablation Study

We conduct four ablation studies to validate the performance of our DSN model: 1) A baseline that learns a common representation for sketches and images with a pre-trained model as teacher model inspired by [Liu *et al.*, 2019]; 2) Adding our cross-modal contrastive method $\mathcal{L}_{cmcm}$ to the baseline that smooths the domain gap by mining relations between samples; 3) Adding our category-specific memory bank $\mathcal{L}_{ml}$ to the baseline that reduces the large intra-class diversity; 4) Full DSN model. The experiments are conducted with 64-d binary hash codes for retrieval in Sketchy dataset.

As shown in Table 2, $\mathcal{L}_{cmcm}$ can improve baseline performance by 16% and 5% in Sketchy and TU-berlin respectively, while $\mathcal{L}_{ml}$ can improve baseline performance by 14% and 5%. However, combining $\mathcal{L}_{cmcm}$ with $\mathcal{L}_{ml}$ improves the baseline only by 21% and 8% for each dataset. It should be noted that the improvement with $\mathcal{L}_{cmcm} + \mathcal{L}_{ml}$ isn't a simple addition operation with $\mathcal{L}_{cmcm}$ and $\mathcal{L}_{ml}$. In our knowledge, the reason may be that our cross-modal contrastive method has already reduced the intra-class diversity in sketch domain to some extent, which pulls the sketches with a same label together in a training batch. These two novel methods have intersection in reducing intra-class diversity, which results in the phenomenon mentioned above. Besides, the experiment results demonstrating $\text{Imp}(\mathcal{L}_{cmcm}) > \text{Imp}(\mathcal{L}_{ml})$, where Imp denotes the improvement from baseline, confirms our conjecture from another perspective.

5 Conclusion

In this paper, we present a novel yet effective Domain-Smoothing Network for ZS-SBIR. First, a cross-modal contrastive method is proposed to smooth the domain gap by mining relations between sketches and images with sufficient positive and negative samples. Second, we build a category-specific memory bank to address the large intra-class diversity in the sketch domain, where prototypes of each sketch category are learned to reduce the variability in appearance. We conduct extensive experiments on two popular datasets, *i.e.*, Sketchy and TU-Berlin, which demonstrate our DSN method outperforms existing methods by a large margin and validate its superiority.

Acknowledgments

References

[Akata *et al.*, 2015] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7):1425–1438, 2015.

[Bhunia *et al.*, 2020] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *CVPR*, pages 9779–9788, 2020.

[Deng *et al.*, 2020] Cheng Deng, Xinxun Xu, Hao Wang, Muli Yang, and Dacheng Tao. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. *IEEE Transactions on Image Processing*, 29:8892–8902, 2020.

[Dutta and Akata, 2019] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, pages 5089–5098, 2019.

[Dutta and Biswas, 2019] Titir Dutta and Soma Biswas. Style-guided zero-shot sketch-based image retrieval. In *BMVC*, page 209, 2019.

[Eitz *et al.*, 2012] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44:1–44:10, 2012.

[Felix *et al.*, 2018] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018.

[Gong *et al.*, 2012] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2916–2929, 2012.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[Kodirov *et al.*, 2017] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017.

[Li *et al.*, 2021] Xiangyu Li, Zhe Xu, Kun Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. 2021.

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

[Liu *et al.*, 2019] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, pages 3662–3671, 2019.

[Lu *et al.*, 2018] Peng Lu, Gao Huang, Yanwei Fu, Guodong Guo, and Hangyu Lin. Learning large euclidean margin for sketch-based image retrieval. *arXiv preprint arXiv:1812.04275*, 2018.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Qi *et al.*, 2016] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, pages 2460–2464. IEEE, 2016.

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

[Shen *et al.*, 2018] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, pages 3598–3607, 2018.

[Wei *et al.*, 2019] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3741–3749, 2019.

[Wei *et al.*, 2020a] Kun Wei, Cheng Deng, and Xu Yang. Lifelong zero-shot learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 551–557, 2020.

[Wei *et al.*, 2020b] Kun Wei, Cheng Deng, Xu Yang, and Maosen Li. Incremental embedding learning via zero-shot translation. *arXiv preprint arXiv:2012.15497*, 2020.

[Xian *et al.*, 2018a] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2251–2265, 2018.

[Xian *et al.*, 2018b] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018.

[Xu *et al.*, 2020] Xinxun Xu, Cheng Deng, Muli Yang, and Hao Wang. Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:2003.09869*, 2020.

[Yang *et al.*, 2016] Yang Yang, Yadan Luo, Weilun Chen, Fumin Shen, Jie Shao, and Heng Tao Shen. Zero-shot hashing via transferring supervised knowledge. In *ACM-MM*, pages 1286–1295, 2016.

[Yelamarthi *et al.*, 2018] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, pages 316–333. Springer, 2018.

[Yu *et al.*, 2017] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *IJCV*, 122(3):411–425, 2017.

[Zhang *et al.*, 2018] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, pages 297–314, 2018.