DOI: 10.1111/cgf.14947

Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 7

Sharing Model Framework for Zero-Shot Sketch-Based Image Retrieval

Yi-Hsuan Ho[1], Der-Lor Way[2], Zen-Chung Shih[1]

[1]National Yang Ming Chiao Tung University, Taiwan
[2]Taipei National University of the Arts, Taiwan

Abstract

Sketch-based image retrieval (SBIR) is an emerging task in computer vision. Research interests have arisen in solving this problem under the realistic and challenging setting of zero-shot learning. Given a sketch as a query, the search goal is to retrieve the corresponding photographs in a zero-shot scenario. In this paper, we divide the aforementioned challenging work into three tasks and propose a sharing model framework that addresses these problems. First, the weights of the proposed sharing model effectively reduced the modality gap between sketches and photographs. Second, semantic information was used to handle different label spaces during the training and testing stages. The sketch and photograph domains share semantic information. Finally, a memory mechanism is used to reduce the intrinsic variety in sketches, even if they all belong to the same class. Sketches and photographs dominate the embeddings in turn. Because sketches are not limited by language, our ultimate goal is to find a method to replace text searches. We also designed a demonstration program to demonstrate the use of the proposed method in real-world applications. Our results indicate that the proposed method exhibits considerably higher zero-shot SBIR performance than do other state-of-the-art methods on the challenging Sketchy, TU-Berlin, and QuickDraw datasets.

CCS Concepts

• ***Information system*** → *Information retrieval;* • ***Computing methodologies*** → *Machine learning;*

1. Introduction

Sketch-based image retrieval (SBIR) is widely recognized as a crucial vision problem in a wide range of real-world applications. The aim of SBIR is to retrieve natural photographs that are similar to hand-drawn sketches. Although sketches are abstractive, their structural details can provide more information than the text. SBIR can be used to complement conventional text–image cross-modal retrieval methods or classical content-based image retrieval protocols. Moreover, SBIR can serve as an alternative to text-based image retrieval in certain situations. However, there is no guarantee that all possible queries covering all object categories can be collected in a training set. Therefore, zero-shot SBIR (ZS-SBIR) is used as a practical solution for real-world cases. In ZS-SBIR, conventional SBIR and zero-shot learning (ZSL) are combined in a new task.

ZS-SBIR is extremely challenging because a large domain gap, intra-class variability, and limited knowledge regarding unobserved classes must be simultaneously handled in this task. Many studies have investigated the ZS-SBIR [DRD*19, DA19, LXWY19, SLSS18, TXW*21, XDYW20, YRMM18]. The sketch and photograph features were intuitively projected onto a joint embedding space with cross-entropy loss. Because of the heterogeneous characteristics of sketches and photographs, a simple projection cannot effectively reduce the domain gap. Sketches usually depict a specific object using simple strokes, whereas photographs have rich colors and complex backgrounds. A simple projection can result in the loss of many sketch and photographic features. Moreover, the cross-entropy loss cannot reveal the relationship between sketches and photographs of the same class.

A common solution to this problem is to employ a generative model. Additional modules (e.g., generators and discriminators) enable the synthesis of photographic features from sketch features to prevent feature insufficiency. Deep generative models reduce feature loss but ignore the rich discriminative features acquired by the pretrained model and thus might exhibit overfitting. Liu [LXWY19] called this phenomenon catastrophic forgetting, which often occurs during fine-tuning processes because of its irrelevance to new tasks.

To address these shortcomings, we developed a novel sharing model framework. First, certain parameters are shared between the sketch and photograph feature extractors to reduce the domain gap between two modalities. The soft weighting strategy involves learning modality-common and modality-specific features. The layers of a convolutional neural network (CNN) model can effectively learn image characteristics. The batch normalization layer of a CNN model uses an independent mean and variance, so that the model can preserve its unique features in modalities. Second, to transfer knowledge from the training stage to the testing stage, the inconsistency in label spaces in the ZSL is addressed. Semantic embeddings extracted from nonlinear programming models were incorporated into the teacher models to enhance their generality. A teacher model preserves the rich features in the pretrained model by sharing features with a student model. In addition to connecting the two domains, semantic information establishes a link between the seen and unseen categories. Third, to smooth the interclass diversity in the sketches, a memory mechanism was adopted. The sketch samples to be retained in memory are determined according to the average photograph feature, and the average sketch feature in memory controls the embeddings of the sketches; thus, the inherent difference can be smoothed.

The main contributions of this paper are as follows. First, the weights of the proposed sharing model effectively reduced the modality gap between sketches and photographs. Second, semantic information was used to handle different label spaces during the training and testing stages. The text and photograph domains share the same semantic information. Third, a memory mechanism is applied to reduce the intrinsic variety in sketches, even if they all belong to the same class. Fourth, extensive experiments were performed on three large ZS-SBIR datasets: Sketchy, TU-Berlin, and QuickDraw. The experimental results indicate that the proposed approach outperforms the relevant state-of-the-art methods.

2. Related Work

2.1. Sketch-Based Image Retrieval

The aim of SBIR is to bridge the domain gap between sketches and photographs. Studies on SBIR can be divided into two categories: those conduct SBIR using handcrafted features and deep-learning-based methods. SBIR based on handcrafted features involves the use of edge maps extracted from photographs with off-the-shelf descriptors. Some methods for SBIR based on handcrafted features involve extracting the edge map from a natural image and then matching it with the sketch of the image by using a bag-of-words model with specific designed SBIR feature descriptors, such as the histogram of oriented gradients [HC13], histogram of edge local orientations [Saa14], and learned key shapes [SBO15] descriptors. Subsequently, various semantic models [MCCD13, Mil98, PSM14] were used as bridges between the sketches and photographs. However, it is difficult to reduce the domain gap because matching edge maps with unaligned hand-drawn sketches is extremely challenging. This problem is addressed using neural network models that can conduct end-to-end learning on features that can be transferred between the sketch and photograph domains. Some information is inevitably lost because the descriptors are not tailored for SBIR. In SBIR based on deep-learning-based methods, neural networks are adopted to extract deep features. For example, Qi [QSZL16] used a Siamese architecture to improve feature discriminability. Most SBIR approaches based on deep learning methods involve the introduction of ranking losses, such as contrastive loss [KTW*20] and triplet loss [SKP15], in model training.

2.2. Zero-Shot Learning

ZSL in computer vision refers to the recognition of objects that are not observed during the training phase. Thus, ZSL focuses on transferring knowledge from seen classes to unseen classes. Early studies on ZSL [DA19, KTW*20, SLSS18] used attributes in a two-stage approach to infer the label of an image belonging to an unseen class. However, in recent studies on ZSL, direct mapping has been performed from the image feature space to the semantic space. Semantic information is often used in ZSL, such as attribute annotations [APHS15], hierarchical model data [Mil98], and word vectors [MCCD13, PSM14]. ZSL methods can also be divided into two categories: embedding-based and generative approaches. Embedding-based approaches aim to learn multimodal embeddings, and frameworks based on these approaches align visual and semantic spaces [SES*19] or map these spaces to a common intermediate space [AMFS16, LLS*17, ZS16]. For example, a stacked encoder model [KXG17] uses an auto-encoder to map visual features to semantic embeddings. Generative approaches [XDYW20, YRMM18] primarily involve the use of generators to synthesize features of unseen classes in accordance with semantic relations. For instance, Li [LJL*19] trained a conditional Wasserstein generative adversarial network to generate fake features. Moreover, Chen [CD19] integrated these two types of methods into a hybrid model.

2.3. Zero-Shot SBIR

ZS-SBIR involves a combination of ZSL and SBIR, and is a more realistic process than ZSL and SBIR alone. Although producing large sketch datasets is a labor-intensive process because sketches must be drawn manually, the existing data are used appropriately in ZS-SBIR. Shen [SLSS18] pioneered ZS-SBIR and proposed a cross-modal hashing method for this process. Yelamarthi [YRMM18] used an auto-encoder to generate additional details from sketch features. Dutta and Akata [DA19] constructed a paired-cycle-consistent generative model using adversarial training to assist in the alignment of two the modalities. Xu [XDYW20] adopted a progressive projection strategy, in which strong semantic supervision was maintained. They decomposed visual features into domain and semantic ones and then projected all features into a common space. Liu [LXWY19] regarded ZS-SBIR as a catastrophic forgetting problem and designed a teacher–student network for knowledge distillation. They fine-tuned their pre-trained model in an economical manner and leveraged semantic information, such as the inter-class relationship, to achieve knowledge preservation. Tian [TXW*21] used generalizable embeddings rather than semantic embeddings to achieve knowledge distillation. Liu [LXWY19] enabled feature transferability from photos to sketches, and achieved knowledge preservation in realistic cases. Based on knowledge distillation, Wang [WDLT21] introduced a soft-weight strategy and was dedicated to narrowing the domain gap. However, Liu [LXWY19] and Wang [WDLT21] ignored the fact that diverse drawing styles in sketches also degrade performance. Wang [WWY*21] adopted the concept of memory to alleviate this problem, which was concerned with intra-class relationship.

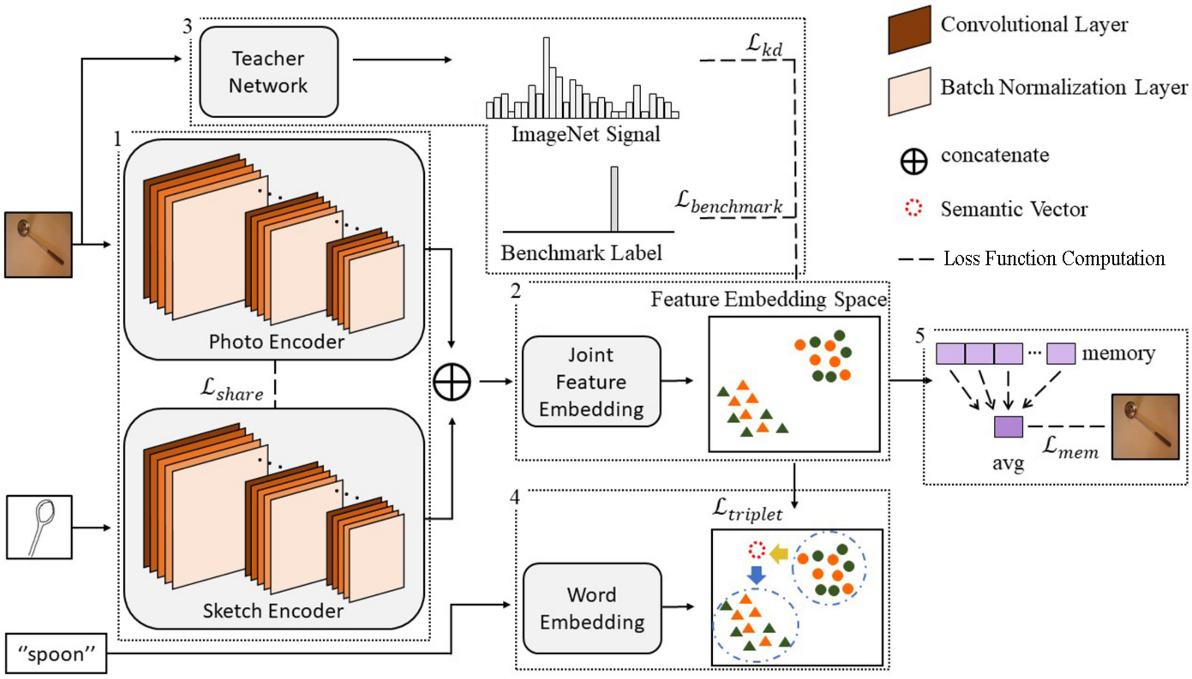


Figure 1: *Architecture of the proposed sharing model framework. (1) Coupled Convolutional Backbones. (2) Joint Embedding Network. (3) Feature Transferability. (4) Semantic Augmentation. (5) Memory Mechanism.*

3. Proposed Approach

This section begins with a description of the ZS-SBIR problem definition and its solution. In particular, intra-class diversity in sketches, which is often neglected, was considered in this study. Figure 1 shows the architecture of the proposed sharing model framework. This framework comprises five modules: two coupled convolutional backbones, a joint embedding network, a feature transferability mechanism, a semantic augmentation mechanism, and a memory mechanism. The concept of sharing forms the basis for our solution to this problem. The details of these modules will be discussed in this section.

First, the problem definition is following: The dataset used in ZS-SBIR comprises training and testing subsets: The training set is composed of photographs ($P^{seen} = \{(p_i, y_i)\,|y_i \in C^{seen}\}_{i=1}^{N1}$), sketches ($S^{seen} = \{(s_i, y_i)\,|y_i \in C^{seen}\}_{i=1}^{N2}$), and semantic embeddings ($W^{seen} = \{w_i^{seen}\}_{i=1}^{N3}$) from the seen categories (C^{seen}), where y_i represents the class label. The testing set contains photographs ($P^{unseen} = \{(p_i, y_i)\,|y_i \in C^{unseen}\}_{i=1}^{M1}$) and sketches ($S^{unseen} = \{(s_i, y_i)\,|y_i \in C^{unseen}\}_{i=1}^{M2}$) from unseen categories (C^{unseen}). $N1, N2, N3, M1$, and $M2$ represent the numbers of corresponding data. Under the zero-shot setting, C^{seen} and C^{unseen} are disjoint, that is, $C^{seen} \cap C^{unseen} = \phi$. Data from C^{seen} are used to construct a retrieval model during the training phase. In the testing phase, given a sketch $s \in S^{unseen}$ with a label $y_s \in C^{unseen}$ as a query, the ultimate goal is to retrieve similar photographs $p \in P^{unseen}$ with labels $y_p \in C^{unseen}$, that is, $y_s = y_p$.

3.1. Coupled Convolutional Backbones

Coupled convolutional backbones comprise a photograph encoder and a sketch encoder, and ResNet50 is adopted as the initial CNN feature extractor. In some previous studies, researchers used two completely independent networks as encoders without sharing parameters [DRD*19, DA19, YRMM18], which resulted in overfitting to a certain extent. In contrast, other researchers have employed a hard weight-sharing strategy [LXWY19, TXW*21, WWY*21], in which a feature extractor is used on all data. This method reduces the domain gap; however, because most of the adopted models are pre-trained on ImageNet, the model parameters exhibit a strong bias toward the photograph dataset. Consequently, inspired by Wang [WDLT21], we adopted a soft weight-sharing strategy as a compromise between the aforementioned strategies. When using the soft weight-sharing strategy, the photograph and sketch encoders can simultaneously learn discriminative representations and reduce the domain gap. The proposed functions $\mathcal{G}_s : R^o \to R^t$ and $\mathcal{G}_p : R^o \to R^t$ (o is the dimension of original data and t is the dimension of intermediate feature vectors) represent actions of sketch encoder and photograph encoder, respectively. These functions can be formulated as follows:

$$V^s = \mathcal{G}_s\left(S;\theta_s\right), \quad V^p = \mathcal{G}_p\left(P;\theta_p\right)$$

where V^s and V^p are intermediate representations of the sketch and photograph encoders, respectively. The sketch encoder with the parameter θ_s and the photograph encoder with the parameter θ_p optimize the sketch and photograph modalities, respectively. We used convolutional layers, pooling layers, and batch normalization layers to implement the soft weight-sharing strategy. Convolutional layers are designed to learn visual information. Pooling layers then extract the patterns of images (e.g., edges). The batch normalization layers solve internal covariate shift by calculating the mean and variance of batch wise data, which ensures the uniqueness of the input. Consequently, during the fine-tuning process, the two encoders shared all the parameters, except for the parameters in the batch normalization layers. The sharing loss is defined as follows:

$$\mathcal{L}_{share} = \sum_{l} 1\left[l \notin BN\right] \cdot \left\|\theta_s^l - \theta_p^l\right\|_2^2$$

where BN is the abbreviation for batch normalization and $1\left[l \notin BN\right]$ is an indicator function. The terms θ_s^l and θ_p^l are the parameters of the sketch and photograph encoders in layer l, respectively. For $\left\|\theta_s^l - \theta_p^l\right\|_2^2$, the Frobenius norm is used to compute the matrix norm. If l does not belong to the BN layer, then the matrix norm is accumulated. Furthermore, hard and soft sharing strategies can be analyzed from the gradient aspect. The loss is formulated as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_s\left(\mathcal{G}_s\left(s, y_s\right); \theta_s\right) + \mathcal{L}_p\left(\mathcal{G}_p\left(p, y_p\right); \theta_p\right) + \lambda_{share} \cdot \mathcal{L}_{share} \\ &= \mathcal{L}_s\left(\mathcal{G}_s\left(s, y_s\right); \theta_s\right) + \mathcal{L}_p\left(\mathcal{G}_p\left(p, y_p\right); \theta_p\right) + \lambda_{share}\|\theta_s - \theta_p\|_2^2\end{aligned}$$

where λ_{share} is the coefficient of sharing loss and y_s and y_p are the labels of the sketches and photographs, respectively. For the hard sharing strategy, all models share the same parameters, including those of the batch normalization layers. Therefore, $\theta = \theta_s = \theta_p$, and the indicator function in $\mathcal{L}_{share}$ was deleted. The loss gradient is computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}_s\left(\mathcal{G}_s\left(s, y_s\right); \theta\right) + \partial \mathcal{L}_p\left(\mathcal{G}_p\left(p, y_p\right); \theta\right)}{\partial \theta}$$

No other restrictions exist on the two modalities; thus, the entire training process may be imbalanced and has the tendency to optimize the photograph modality because of the pretrained model. For the soft-sharing strategy, the gradients of θ_s and θ_p are computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \theta_s} = \frac{\partial \mathcal{L}_s\left(\mathcal{G}_s\left(s, y_s\right); \theta_s\right) + \lambda_{share} \cdot \mathcal{L}_{share}}{\partial \theta_s}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_p} = \frac{\partial \mathcal{L}_p\left(\mathcal{G}_p\left(p, y_p\right); \theta_p\right) + \lambda_{share} \cdot \mathcal{L}_{share}}{\partial \theta_p}$$

The antecedent and consequent of the numerator represent the tradeoff between discriminability and reducing the domain gap. Thus, encoders obey similar feature extraction rules and simultaneously reserve their own characteristics. The two aforementioned encoders not only learn modality-common features but also retain their modality-specific features.

3.2. Joint Embedding Network

After intermediate representations of sketches (V^s) and photographs (V^p) are acquired, retrieval features can be obtained. Figure 1 illustrates the function of the joint embedding network. The representations V^s and V^p passed through the same single fully connected layer to learn the retrieval features. This operation also reduces the domain gap to certain extent. The term $f : R^t \rightarrow R^d$ (d is the dimension of retrieval features) represents the function of the joint embedding network. Moreover, the generation of retrieval features can be formulated as follows:

$$E^s = f\left(V^s; \theta\right), \quad E^p = f\left(V^p; \theta\right)$$

where θ is the parameter of the joint embedding network f. The standardization and reduction of the dimension of retrieval features make the retrieval task more practical because the dimension of the features influences the speed of the retrieval task. After the aforementioned operation , all the features are projected into a common space, namely the embedding feature space. In Figure 1, the same color in the embedding feature space represents data from the same modality, and the same shape represents data from the same category.

3.3. Feature Transferability

To increase feature discriminability, cross-entropy loss (defined as the benchmark loss in Figure 1) was incorporated into the training process. This loss can be expressed as follows:

$$\mathcal{L}_{benchmark} = -E\left[\log P\left(y_i | X_i^b\right)\right] = \frac{1}{N}\sum_i\left(-\log\frac{\exp\left(X_i^b\right)}{\sum_{k \in C^{seen}} \exp\left(X_k^b\right)}\right)$$

where the superscript b represents the SBIR benchmark dataset, and y represents the one-hot label. Term X can be E^s or E^p. By using benchmark loss, feature knowledge from the ImageNet classification task can be fine-tuned and implemented in a new task on the SBIR dataset. The inconsistency in the label space between seen and unseen classes must be addressed. Therefore, the concept of knowledge distillation was adopted. Knowledge distillation [YXQY18] involves transferring deep knowledge from a teacher network to a student network. The teacher model teaches the student model to remember abundant visual features and make good ImageNet label predictions. We selected an ImageNet-pretrained network, ResNet50, as the teacher model. The knowledge distillation loss is related only to the photograph modality. This loss is similar to the benchmark loss and can be formulated as follows:

$$\mathcal{L}_{kd} = -E\left[\log P\left(y_i | X_i^t\right)\right] = \frac{1}{N}\sum_i \sum_{m \in C^t} -q_{i,m}^t \log\frac{\exp\left(X_i^t\right)}{\sum_{k \in C^t} \exp\left(X_k^t\right)}$$

where the superscript t represents the ImageNet dataset, q_i^t is the supervised probability of sample X belonging to each category in C^t, and q_i^t is the pseudo ImageNet label generated by the teacher model. During the training phase, the parameters of the teacher model are fixed such that stable implicit semantic information is included in the pseudo-labels. The feature transferability loss is defined as follows:

$$\mathcal{L}_{trans} = \mathcal{L}_{benchmark} + \mathcal{L}_{kd}$$

3.4. Semantic Augmentation

Semantic supervision is crucial in ZSL, because semantic information is similar to a dictionary that provides a global view. Therefore, we adopted a widely used text model, namely Word2Vec [MCCD13], to obtain semantic information. Word2Vec was trained on a part of the Google News dataset (roughly 100 billion words) to obtain text representations. Words are embedded into vectors and projected into a high-dimensional vector space, and triplet loss is used to constrain the relationships between the features under the supervision of the text model. Triplet loss is a type of ranking loss that shortens the distance between samples if they are similar and increases the distance between samples if they are dissimilar.

As displayed in Figure 2, the sketch and photograph features and semantic vectors are projected into a common space. The anchor is the result of interpolation between the sketch or photograph features and semantic vectors. The interpolation formula is as follows:

$$Z_i = h\left(W_i^{seen}, \theta_w\right), \quad I_i^A = \beta \cdot Z_i + (1-\beta) \cdot X_i, \quad 0 \leq \beta \leq 1$$

where h is a fully connected layer with the parameter θ_w that projects semantic vectors W_i^{seen} from the original dimension to R^d. X_i can be E^s or E^p. I_i^A represents the interpolation result, and the superscript A denotes the anchor. The value of β is random.

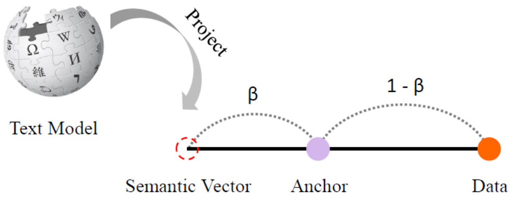


Figure 2: *Interpolation between semantic vectors and sketch or photograph features.*

Some studies [DA19, YRMM18] have used generative models to handle semantic information. Compared with interpolation models, generative models require additional epochs for training, and they may exhibit overfitting in certain model branches. Interpolation provides sufficient diversity for training, and complex treatment is not required because all the adopted modules are already well trained for their original tasks. To ensure maximum improvement , the hardest positive and negative samples were selected for anchoring to calculate the triplet loss. The hardest positive sample represents the sketch or photograph sample that belongs to the same class as the anchor but is farthest from the anchor. The hardest negative sample represents the sketch or photograph sample that belongs to a different class and is closest to the anchor. The aforementioned process is expressed as follows:

$$I_i^{pos} = \underset{j, W_j^{seen}=W_i^{seen}}{\arg\max} \; D\left(I_i^A, X_j\right), \quad I_i^{neg} = \underset{j, W_j^{seen}\neq W_i^{seen}}{\arg\min} \; D\left(I_i^A, X_j\right)$$

$$\mathcal{L}_{triplet} = \sum_{j=1}^{batch} \varphi\left(D\left(I_j^A, I_j^{pos}\right) - D\left(I_j^A, I_j^{neg}\right) + margin\right)$$

where the superscript *pos* represents a positive sample, the superscript *neg* represents a negative sample, and D is the Euclidean distance function. A distance function is used to select the hardest positive and negative samples. The margin value depends on the dataset used and is employed to ensure that after triplet loss adjusts the interclass distance, different categories can be strongly distinguished. The term φ represents the soft-plus activation function. Figure 3 illustrates the process . In summary, the addition of semantic information to a model can be considered to be a form of data augmentation. The generality of the model is enhanced by using a text model.

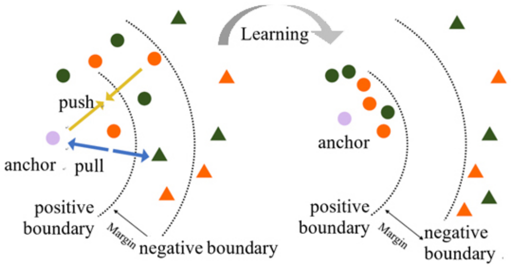


Figure 3: *(Left) Initial projection in the common space and (right) our ideal result after adopting triplet loss.*

3.5. Memory Mechanism

Hand-drawn sketches can be quite different to and less abundant than photographs. Therefore, feature embeddings are often considered outliers. Inspired by Wang [WWY*21], we use a memory mechanism to solve this problem. Prototype learning [YZYL18] is the basis of the adopted memory mechanism. This type of learning was originally designed for few-shot learning. In prototype learning, the average of features becomes the prototype of each category to avoid extreme cases. In the adopted memory mechanism, a memory bank that can store k samples per class is prepared. When the memory storage is less than k, a sketch feature of the class can be added at will to the storage. However, when the memory storage is full, a decision must be made regarding whether to remove a sample from the memory and add a new one or simply retain the existing samples. Take "spoon" category as an example. The y label with hat is the ground truth; otherwise, is the prediction. $U^p = \left\{E^p | y^{\hat{p}} = y^{\hat{s}}\right\}$ represents a set of photographs that belongs to the "spoon" class as a new "spoon" sketch query. Although our predictions might be incorrect during training, we only considered the correct predictions here. E_{avg}^p represents the mean feature of U^p. The sketch in the memory that is farthest from E_{avg}^p is replaced by the new sketch. Otherwise, the memory maintains the original state if the new sketch is the farthest one. The criterion for calculating distance is cosine similarity. The aforementioned rules are followed to update the memory $M^s = \left\{E^s | y^s = y^{\hat{s}}\right\}$. Moreover, the magnitude of M^s is equal to k (i.e. $|M^s| = k$). We considered only the top k sketch features near to E_{avg}^p. Figure 4 illustrates the adopted update mechanism. Finally, the mean feature of the sketches (E_{avg}^s) in memory M^s is calculated, and all photograph features should not be too far from E_{avg}^s. Their summation is defined as the memory loss. Figure 5 illustrates the aforementioned calculation process. The relevant formulas for this process are as follows:

$$E_{avg}^p = \frac{1}{|U^p|} \sum_{E^p \in U^p} E^p, \; E_{avg}^s = \frac{1}{|M^s|} \sum_{E^s \in M^s} E^s$$

$$\mathcal{L}_{memory} = \sum_{i}^{batch} cosine\ similarity(E_i^p, E_{avg}^s)$$

The photograph features dominate the sketch embedding results because they determine the members of the memory bank . Therefore, the intrinsic sketch diversity can be reduced. Irrespective of the type of sketches, the sketch embeddings are not considered outliers.

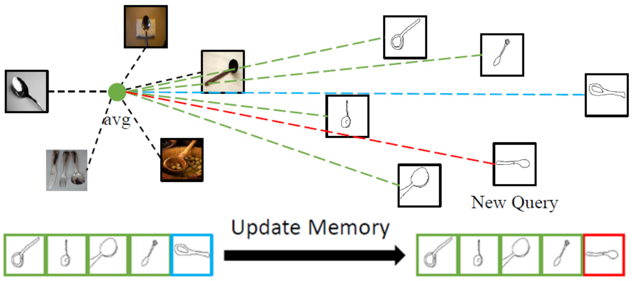


Figure 4: *Memory update when a new sketch query is generated. The green solid circle represents E^p_{avg}. The dotted lines between E^p_{avg} and the sketches indicate their cosine similarities.*

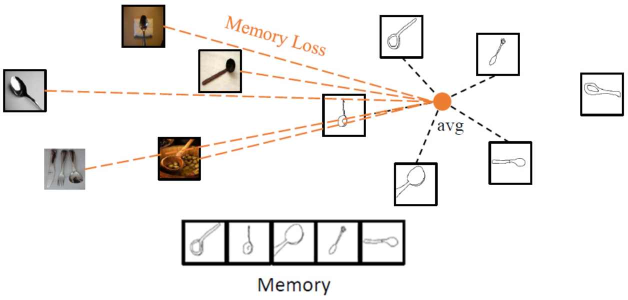


Figure 5: *Calculation of the memory loss for a data batch. For example, for all photographs labeled "spoon," the loss is computed with E^s_{avg}, which is denoted by the orange solid circle.*

3.6. Objective and Optimization

Because our proposed modality loss is constrained in terms of sharing, feature transferability, triplet loss, and memory mechanism, the total modality loss can be expressed as follows:

$$\mathcal{L} = \lambda_{share} \cdot \mathcal{L}_{share} + \lambda_{trans} \cdot \mathcal{L}_{trans} + \lambda_{triplet} \cdot \mathcal{L}_{triplet} + \lambda_{memory} \cdot \mathcal{L}_{memory}$$

where $\lambda_{share}, \lambda_{trans}, \lambda_{triplet}, \lambda_{memory}$ are coefficients for balancing the overall model performance . Our objective is to minimize the loss function.

4. Experiments

4.1. Datasets

Three large-scale sketch datasets are widely used for the ZS-SBIR: Sketchy, TU-Berlin, and QuickDraw.

Sketchy was initially created for fine-grained retrieval [SBHH16], and each sketch was drawn using a corresponding photograph as a reference. This dataset initially contains 75,471 sketches and 12,500 photographs from 125 categories. Later, Liu [LSS*17] collected an additional 60,502 photographs from ImageNet so that Sketchy could be adapted for SBIR. Sketchy now contains 73,002 photographs and 75,471 sketches from 125 categories. Shen [SLSS18] randomly selected 25 classes from this category as the test set. However, the classes selected in the ZSL should not overlap with the ImageNet categories. Consequently, Yelamarthi [YRMM18] selected 21 classes that did not overlap with ImageNet categories. These classes were used in the experiments.

TU − Berlin consists of 20,000 freehand sketches belonging to 250 categories [EHA12]. The dataset was originally used for sketch classification and recognition. The main disadvantage of TU-Berlin is the ambiguity in the naming of its labels, such as “seagull” and “flying-bird.” Zhang [ZLZ*16] collected an additional 204,489 photographs to extend the dataset. We selected the 30 classes used by Shen [SLSS18] as a test set and used the remaining 220 classes as a training set.

QuickDraw is a dataset obtained by Google [JRK*16] that comprises 50 million amateur sketches belonging to 345 categories. To make this dataset suitable for SBIR, Dey [DRD*19] selected 110 of 345 categories. QuickDraw contains 330,000 sketches and 204,000 photographs. The sketches in QuickDraw are more abstract than those in TU-Berlin and Sketchy. Thus, QuickDraw is more realistic than TU-Berlin and Sketchy. In accordance with the study by Dey [DRD*19], we selected 30 classes as the test set and the remaining classes as the training set.

4.2. Implementation Details

All the experiments in this study were conducted on Pytorch using two GTX 1080 Ti graphics processing units. The proposed framework was trained using the Adam optimizer with β_1 = 0.9 and β_2 = 0.999. The initial learning rate was 0.0001, weight decay parameter was 0.0005, and batch size was 36. The maximum number of epochs was set to 25, and the model with the highest sketch classification accuracy was selected. The input size of the photographs and sketches was 224 × 224 pixels. Except for the margin, the other parameters were fixed for the three datasets (λ_{share}= 1000, λ_{trans}=1, $\lambda_{triplet}$= 1, and λ_{memory}= 1). The margin value was 1.0 for Sketchy, 0.1 for TU-Berlin, and 1.0 for QuickDraw. Because labels in TU-Berlin are ambiguous, a strict or large margin is unsuitable. Moreover, we used the ImageNet-pretrained ResNet50 architecture as our teacher model, sketch encoder, and photograph encoder. The parameters of the teacher model were fixed during the training. The adopted pre-trained model can be replaced if a more powerful feature extractor is developed in the future.

4.3. Comparison with Peer Methods

To evaluate our model, we compared its performance with that of relevant state-of-the-art methods proposed in previous studies on SBIR, ZSL, and ZS-SBIR. A performance comparison was conducted for the 64- and 512-dimensional features. The 64-dimensional features are of two forms: real-valued features and binary codes. Table 1 presents the performance of the compared methods. In Table 1, mAP and Prec@N are commonly used metrics for evaluating the performance of retrieval and object detection

Table 1. Comparison of the performance of our proposed method and existing SBIR, ZSL, and ZS-SBIR methods.

Method	Dimension	Sketchy Split 1		Sketchy Split 2		TU-Berlin		QuickDraw	
		mAP@all	Prec@100	mAP@all	Prec@200	mAP@all	Prec@100	mAP@all	Prec@100
Siamese CNN [QSZL16]	64	0.132	0.175	-	-	0.109	0.141	-	-
GN-Triplet [SBHH16]	1024	0.204	0.296	-	-	0.175	0.253	-	-
DSH [LSS*17]	64b	0.164	0.210	0.059	0.153	0.122	0.198	-	-
DeViSE[FCS*13]	300	0.067	0.077	-	-	0.059	0.071	-	-
SAE [KXG17]	100	0.210	0.302	.0136	0.238	0.161	0.210	-	-
ZSH [YCL*16]	64b	0.165	0.217	-	-	0.139	0.174	-	-
ZSIH [SLSS18]	64b	0.254	0.340	-	-	0.220	0.291	-	-
CAAE [YRMM18]	4096	0.196	0.284	0.156	0.260	-	-	-	-
CVAE [YRMM18]	4096	-	-	0.225	0.333	-	-	0.003	-
Doodle [DRD*19]	256	-	-	0.460	0.370	0.109	-	0.075	-
SEM-PCYC [DA19]	64	0.349	0.463	-	-	0.297	0.426	-	-
	64b	0.344	0.399	-	-	0.293	0.392	-	-
SAKE [LXWY19]	64b	0.364	0.487	0.356	0.477	0.359	0.481	-	-
	512	0.547	0.692	0.497	0.598	0.475	0.599	-	-
OCEAN[ZXS*20]	64	0.462	0.590	-	-	0.333	0.467	-	-
PDFD [XDYW20]	64b	**0.638**	**0.755**	-	-	0.386	**0.542**	-	-
	512	**0.661**	**0.781**	-	-	0.483	0.600	-	-
TCN [WDLT21]	64b	0.488	0.644	0.401	0.514	0.381	0.506	0.110	0.150
	512	0.616	0.763	0.516	**0.608**	0.495	0.616	0.140	0.231
RDKD [TXW*21]	64b	0.423	0.536	0.371	0.485	0.361	0.491	-	-
	512	0.613	0.723	0.502	0.598	0.486	0.612	0.143	**0.230**
3JOIN [ZLW*22]	64b	0.462	0.595	-	-	0.361	0.487	-	-
	512	0.620	0.724	-	-	0.496	0.613	-	-
Ours	64b	0.503	0.659	**0.415**	**0.524**	**0.400**	0.509	**0.113**	**0.151**
	512	0.642	0.770	**0.519**	0.607	**0.507**	**0.620**	**0.147**	0.227

Notes:
1. SBIR: row 1-3; ZSL: row 4-6; ZS-SBIR: other rows.
2. "b" denotes the binary codes, and "-" represents results that are not reported in the original papers.
3. The best result in each experiment setting is shown in underline bold and higher value is better.

tasks. mAP stands for Mean Average Precision and it involves taking the average precision for each class and then averaging them across all classes. Prec@N represents the accuracy within the top N predictions. ZS-SBIR outperformed SBIR. Unlike ZSL, only the domain gap is considered when neglecting the label space differences in SBIR. The knowledge learned from the training stage was not successfully transferred to the testing stage. Thus, the model exhibits overfitting and cannot be generalized to new categories. The performance of the ZSL was considerably lower than that of ZS-SBIR. In ZSL, only the consistency label space is considered during training and testing; however, the domain gap between the two modalities is not addressed, which results in the two modalities lying in different manifolds. The proposed ZS-SBIR method outperformed the existing ZSL, SBIR, and ZS-SBIR methods in the majority of the conducted experiments.

- *For Sketchy*, our method outperformed the other methods in all cases except for PDFD [XDYW20] under split 1 and TCN [WDLT21] under split 2. Sketchy has an implicit one-to-one mapping characteristic because it was originally created for fine-grained retrieval. Therefore, previous studies [DA19, XDYW20] trained a more fitting model as the pretrained model. If our sketch and photograph encoders same as PDFD are adopted [XDYW20], the model performance would improve considerably. The results of the performance comparison are listed in Table 2.

Table 2. Performance of the proposed method and PDFD methods using the same pretrained encoders.

Method	Dimension	Sketchy Split 1	
		mAP@all	Prec@100
PDFD [XDYW20]	64b	0.638	0.755
	512	0.661	0.781
Ours	64b	**0.666**	**0.819**
	512	**0.724**	**0.860**

ps : The best result is shown in bold and higher value is better.

- *For TU − Berlin*, our proposed method exhibited the best mAP@all results in all situations, especially for 512-dimensional features. Regardless of whether we win or lose, our overall performance remains topnotch.

- *For QuickDraw*, the proposed method exhibited a good performance in terms of mAP@all ; however, its performance was marginally lower than that of RDKD in terms of Prec@100. Although the precision of the proposed method for the first 100 retrievals was not very high, it still retrieved more correct photographs than the other methods. The QuickDraw dataset contains many amateur sketches that humans find difficult to distinguish. Therefore, the adopted methods exhibited poorer performance on this dataset than on the other two datasets.

- *On average*, the proposed method exhibits state-of-the-art performance for all datasets. In particular, the proposed method outperformed state-of-the-art methods on Sketchy by a considerable margin. The experimental results indicate the superiority of the proposed approach over the other methods.

4.4. Ablation Studies

To analyze the effect of every component of our model, we ablated the loss term of each component. For convenience, the experiments were conducted only on Sketchy under split 1, and the relevant results are presented in Table 3. Model 1 contains all the loss terms and represents the proposed method. One loss term was removed in sequence to create four additional models, models 2–5, to test the usefulness of each loss function. The inferences obtained from Table 3 are described in the following text. The results obtained with model 2 indicated that $\mathcal{L}_{share}$ was the most important loss component. In the absence of $\mathcal{L}_{share}$, the sketch embeddings lacked supervision from the photograph domain, which resulted in the neglect of the domain gap and a reduction in discriminability. Weight sharing has a significant impact on capturing the crucial features. Because we only utilize a hand-drawn sketch to search during the testing stage, the ability to capture features is essential. The results obtained with model 3 indicate that $\mathcal{L}_{triplet}$ improves model performance through the addition of semantic information and metric learning. Traditional triplet metric learning aids retrieval of features while maintaining sample similarity. The aforementioned results also suggest that the addition of semantic information to the anchor brought it closer to the class center. Model 4 ablated $\mathcal{L}_{trans}$, and $\mathcal{L}_{trans}$ consisted of $\mathcal{L}_{kd}$ and $\mathcal{L}_{benchmark}$, which were used to boost the model discriminability. In particular, $\mathcal{L}_{kd}$ preserved the feature extraction knowledge from the teacher model so that we could better deal with the zero-shot scenario under the supervision of ImageNet labels, whereas $\mathcal{L}_{benchmark}$ concentrated on learning the benchmark dataset features. ZS-SBIR benefited from $\mathcal{L}_{triplet}$ and $\mathcal{L}_{trans}$ despite being less powerful than $\mathcal{L}_{share}$. Finally, although $\mathcal{L}_{memory}$ was not as informative as the other loss terms, the results indicated that it could solve the sketch diversity problem.

Table 3. Ablation results for split1of Sketchy

	$\mathcal{L}_{share}$	$\mathcal{L}_{triplet}$	$\mathcal{L}_{trans}$	$\mathcal{L}_{memory}$	Sketchy	
					mAP@all	Prec@100
Model 1	v	v	v	v	0.642	0.770
Model 2	x	v	v	v	0.347	0.451
Model 3	v	x	v	v	0.608	0.745
Model 4	v	v	x	v	0.600	0.738
Model 5	v	v	v	x	0.637	0.764

Note: “v” indicates that the corresponding loss term was used during training, whereas “x” indicates that the corresponding loss term was not used during training.

4.5. Demo Application

An application program was developed to demonstrate the proposed method using a simple user interface, as shown in Figure 6. Our search application consists of one drawing area, two buttons (clear and submit), and a search result area. Users can draw sketches freely. After finishing the drawing, the user can submit using the right button and search for the corresponding photos according to the user's sketch. Users can clean the sketch using the left “Clear” button. Finally, the corresponding top ten photos were obtained according to the input sketch.

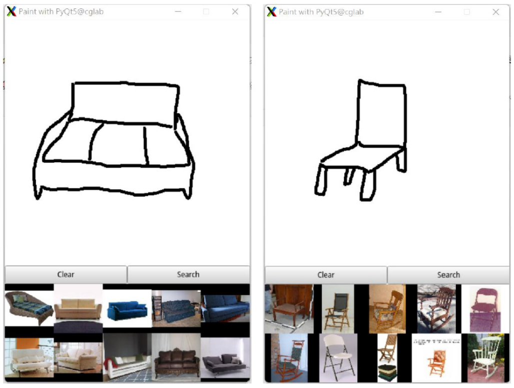
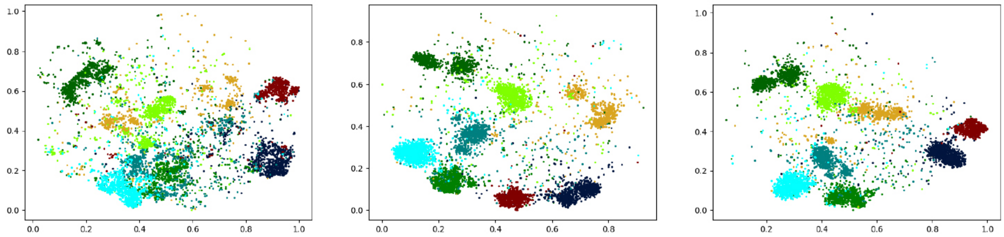


Figure 6: *Our demo searching application. Users drew a sofa sketch and a chair sketch as queries, respectively.*



(a) Model without $\mathcal{L}_{share}$ (b) Model without $\mathcal{L}_{trans}$ (c) Our proposed method

Figure 7: *The figures from left to right illustrate the training process of the proposed model. We selected categories with 64-dimensional features to demonstrate our results, and one color represents one class.*

4.6. Experiment Analysis

Visualization of Embeddings: In accordance with the ablation results presented in Table 3, we selected three settings to prove the effectiveness of our model. The worst result was obtained when using the model without $\mathcal{L}_{share}$, as shown in Figure 7(a). A large domain gap exists in the two-dimensional embedding space because the retrieval features are almost randomly distributed. Figure 7(b) depicts the result obtained without $\mathcal{L}_{trans}$, which is better than the result displayed in Figure 7(a). Dots of the same color are closer in Figure 7(b) than in Figure 7(a), where each color represents a different category. The final result is showed in Figure 7(c), which represents an improvement over the embedding results displayed in Figure 7(b). The clusters converged most strongly.

Visualization of the Retrieval Results: We visualized some retrieval results from the three benchmark datasets for qualitative analysis, as shown in Figures 8, 9 and 10. The input is a hand-drawn sketch in the first column that retrieves the top eight photograph results for each query. The color boxes are marked by hand additionally for ease of understanding. Green border and Red border stands for correct and incorrect retrievals. All the photographs were obtained with 512 retrieval features. In most cases, our model found photographs belonging to the same category as the sketch query. However, as displayed in the fourth row of Figure 9, incorrect results were obtained in some cases because the shape of the bottle opener resembles that of a frying pan. Similar errors are depicted in Figure 10, and incorrect retrievals could occur when the shape and structure patterns of the sketch query and photograph candidates are strikingly similar. Consequently, we inferred that shape relations rather than semantic information dominated retrieval outcomes.

4.7. Efficiency Discussion

Because of its low storage cost and fast query speed, hashing has been widely used in large-scale image retrieval tasks. Because binary hashing simplifies the distance vectors, it often results in poor performance. Therefore, most ZS-SBIR papers focus on search accuracy, and little attention has been given to retrieval speed using binary hashing [WDLT21, ZLW*22]. The main purpose of this study is to focus on search accuracy. Consequently, our program was developed without hashing, to demonstrate the application of sketch-based image retrieval. In our program, if there are N images in the database, the search time will increase as the number of images increases because each image must be compared with the sketch query for similarity. As a result, the search process is obviously time consuming.

In the future, the performance of the hashing method, which is encoded as a binary code from real-valued features to accelerate retrieval speed, such as multiple code hashing for efficient image retrieval, will be investigated. Additionally, parallel programming is another simple way to improve efficiency and adapt to real-world applications.

5. Conclusions and Future Work

In this study, a novel method was developed to solve ZS-SBIR problems. Extensive experiments were conducted on three large-scale ZS-SBIR datasets, and these experiments indicated that the proposed approach is superior to the existing state-of-the-art methods. This study focuses on the trade-off between shared and private features. In the proposed method, shared encoders and a joint embedding network are first used to learn the effective features and reduce the domain gap. Second, using a teacher model and benchmark loss, knowledge is successfully transferred to the target task while preserving the feature extraction ability. Third, the obstacle in ZSL is overcome by adding semantic information to the model and then matching it through triplet metric learning. Finally, a memory mechanism was used to reduce the diversity of hand-drawn sketches.

ZS-SBIR is a complex problem derived from the real world, and we thoroughly examined ZS-SBIR and suggested appropriate solutions for each task. Our experiments showed that the most effective technique would be to propose appropriate solutions for each sub-challenge. Our advantage lies in comprehensively addressing aspects that were previously overlooked and in achieving performance improvements. In addition, according to our observations, the most important factors affecting the performance of the proposed model are the shape and structure of the unknown samples. Zero-shot classification may be a possible solution. Therefore, improving the feature extraction ability is suggested for future research. Moreover, to make ZS-SBIR more practical for real-world applications, the search speed of this method should be increased. With the foundation established in our paper, we believe that striving in these directions will lead to improved performance.

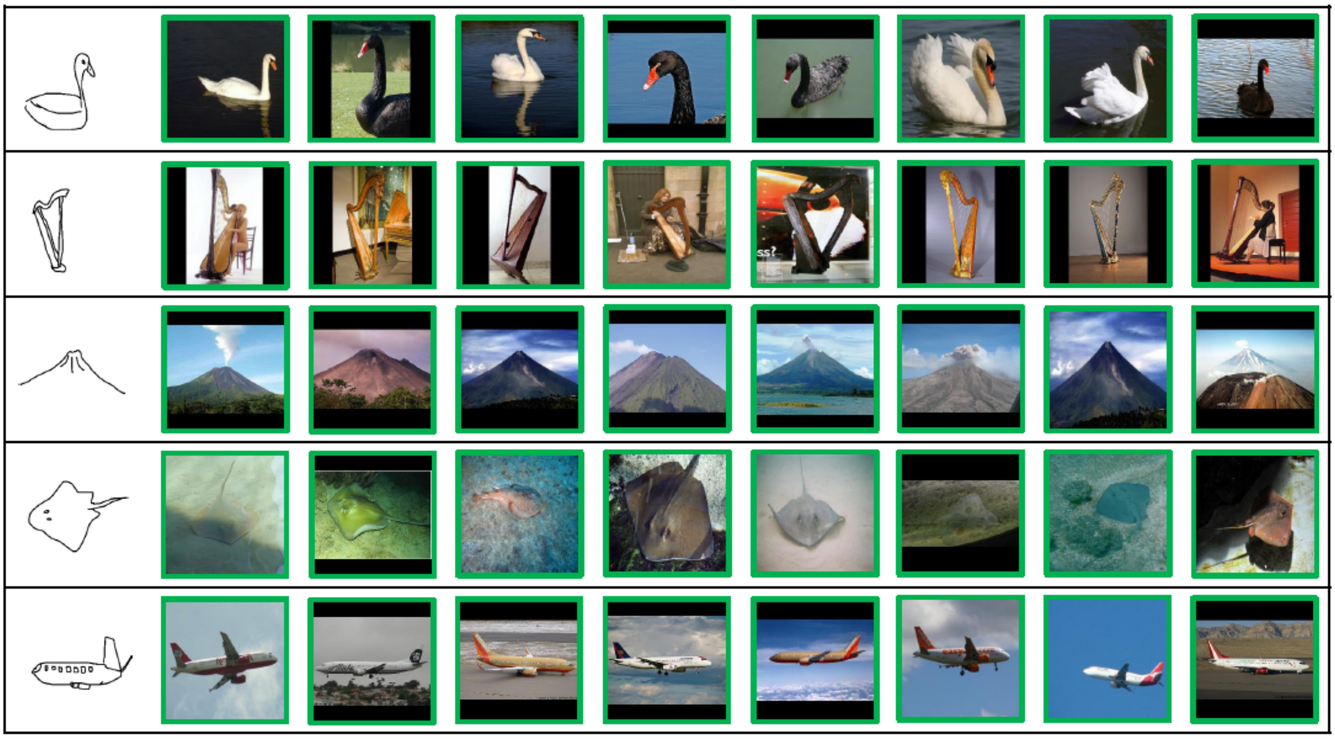


Figure 8: *The figures from left to right illustrate the training process of the proposed model. We selected categories with 64-dimensional features to demonstrate our results, and one color represents one class.*

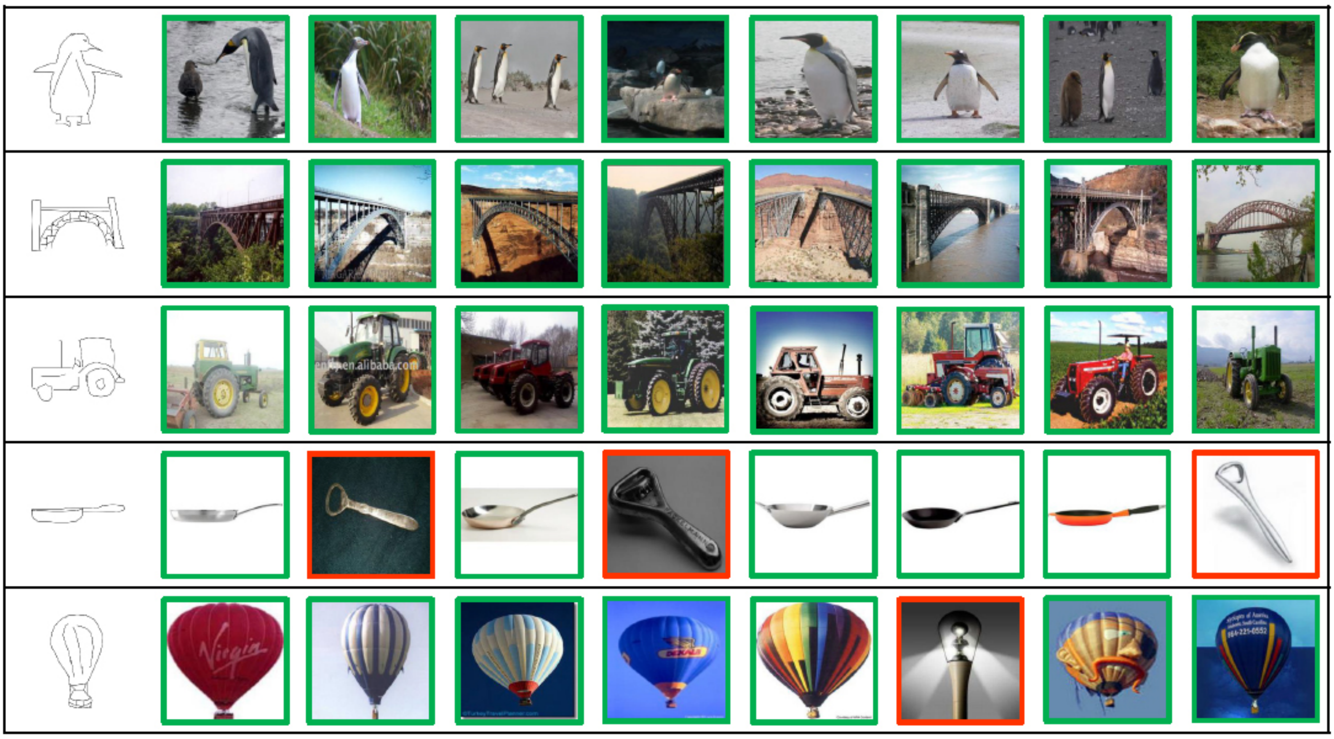


Figure 9: *The figures from left to right illustrate the training process of the proposed model. We selected categories with 64-dimensional features to demonstrate our results, and one color represents one class.*

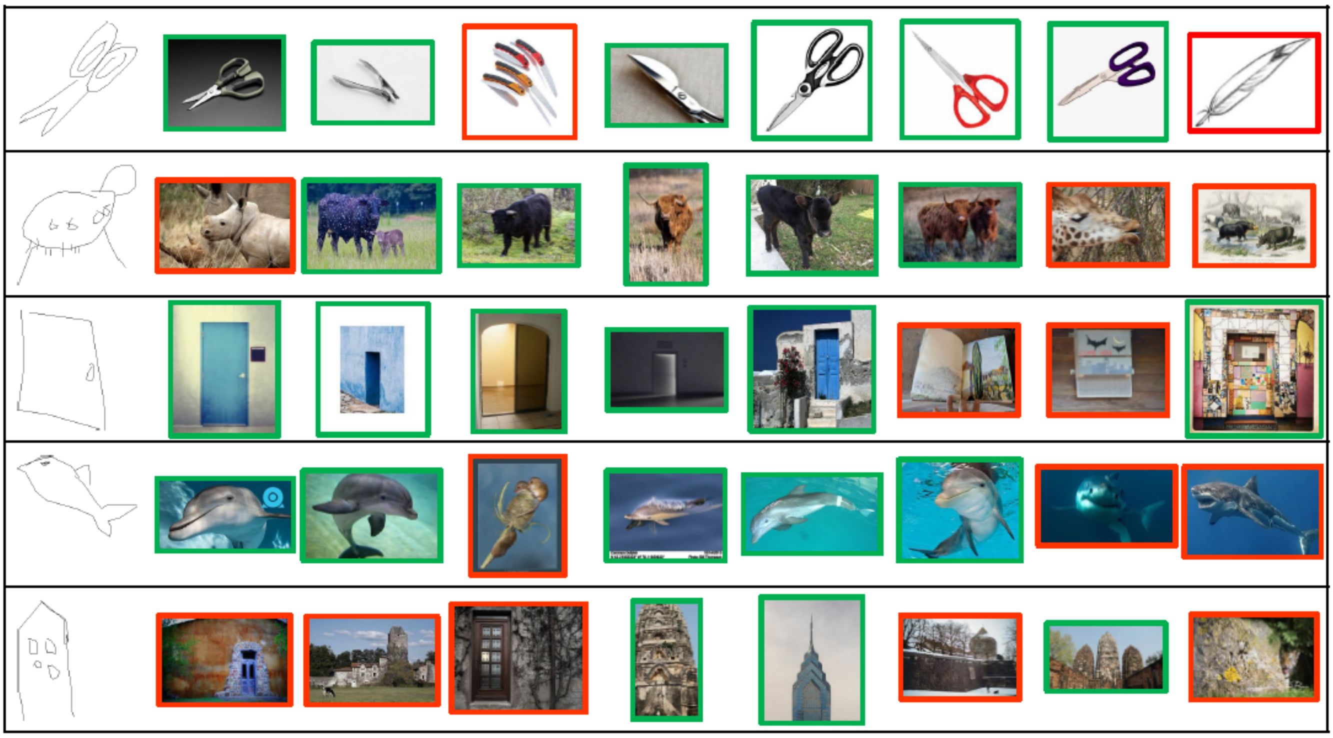


Figure 10: *The figures from left to right illustrate the training process of the proposed model. We selected categories with 64-dimensional features to demonstrate our results, and one color represents one class.*

References

[AMFS16] Akata Z., Malinowski M., Fritz M., Schiele B.: Multi-cue zero-shot learning with strong supervision. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016, pp. 59–68.

[APHS15] Akata Z., Perronnin F., Harchaoui Z., Schmid C.: Label embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015, 38, (7), pp.1425-1438. abs/1503.08677.

[CD19] Chen B., Deng W.: Hybrid attention based decoupled metric learning for zero-shot image retrieval. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, pp. 2750-2759.

[DA19] Dutta A., Akata Z.: Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, pp. 5089-5098 abs/1903.03372.

[DRD*19] Dey S., Riba P., Dutta A., Lladós J., Song Y.: Doodle to search: Practical zero-shot sketch-based image retrieval. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(*CVPR), 2019, pp. 2179-2188, abs/1904.03451.

[EHA12] Eitz M., Hays J., Alexa M.: How do humans sketch objects? *Proc. SIGGRAPH*, 2012, 31, (4), pp.1-10.

[FCS*13] Frome A., Corrado G. S., Shlens J., Bengio S., Dean J., Ranzato M., Mikolov T.: Devise: A deep visual-semantic embedding model. *Conference on Neural Information Processing Systems(NeurIPS)*, 2013, pp. 2121–2129.

[HC13] Hu R., Collomosse J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 2013, vol. 117, no. 7, pp. 790–806.

[JRK*16] Jongejan J., Rowley H., Kawashima T., Kim J., Foxgieg N.: Quick, draw! a.i. experiment. *https://quickdraw.withgoogle.com*, 2016.

[KTW*20] Khosla P., Teterwak P., Wang C., Sarna A., Tian Y., Isola P., Maschinot A., Liu C., Krishnan D.: Supervised contrastive learning. *Conference on Neural Information Processing Systems(NeurIPS)*, 2020, 33, pp.18661-18673.

[KXG17] Kodirov E., Xiang T., Gong S.: Semantic auto-encoder for zero-shot learning. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017, pp. 3174–3183.

[LJL*19] Li J., Jing M., Lu K., Ding Z., Zhu L., Huang Z.: Leveraging the invariant side of generative zero-shot learning. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, pp. 7402-7411, abs/1904.04092.

[LLS*17] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, J. Han: From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017, pp. 1627–1636.

[LSS*17] L. Liu, F. Shen, Y. Shen, X. Liu, L. Shao: Deep sketch hashing: Fast freehand Sketch-based image retrieval. *Proc. of* *the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017, pp. 2862-2871, abs/1703.05605.

[LXWY19] LIU Q., XIE L., WANG H., YUILLE A. L.: Semantic aware knowledge preservation for zero-shot Sketch-based image retrieval. *Proc. of the International Conference on Computer Vision(ICCV)*, 2019, vol. 9487, pp. 3662–3671.

[MCCD13] T. MIKOLOV, K. CHEN, G. CORRADO, J. DEAN: Efficient estimation of word representations in vector space. *International Conference on Learning Representations(ICLR)*, 2013, arXiv:1301.3781.

[Mil98] MILLER G.: Wordnet: An electronic lexical database. *MIT press*, 1998.

[PSM14] PENNINGTON J., SOCHER R., MANNING C.: Glove: Global vectors for word representation. *Conference on Empirical Methods in Natural Language Processing(EMNLP)*, 2014, pp.1532–1543.

[QSZL16] QI Y., SONG Y. Z., ZHANG H., LIU J.: Sketch-based image retrieval via siamese convolutional neural network. *Proc. of the International Conference on Image Processing (ICIP)*, 2016, pp. 2460–2464.

[Saa14] SAAVEDRA J. M.: Sketch based image retrieval using a soft computation of the histogram of edge local orientations (shelo). *Proc. of the International Conference on Image Processing(ICIP)*, 2014, pp. 2998–3002.

[SBHH16] SANGKLOY P., BURNELL N., HAM C., HAYS J.: The sketchy database: Learning to retrieve badly drawn bunnies. *Proc. SIGGRAPH*, 2016, 35, (4), pp. 1-12.

[SBO15] SAAVEDRA J. M., BARRIOS J. M., ORAND S.: Sketch based image retrieval using learned key shapes (lks). *Proc. of the British Machine Vision Conference(BMVC)*, 2015, vol. 1, pp. 1–11.

[SES*19] SCHONFELD E., EBRAHIMI S., SINHA S., DARRELL T., AKATA Z.: Generalized zero and few shot learning via aligned variational auto-encoders. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, p. 8247–8255.

[SKP15] SCHROFF F., KALENICHENKO D., PHILBIN J.: Facenet: A unified embedding for face recognition and clustering. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015, pp. 815-823.

[SLSS18] SHEN Y., LIU L., SHEN F., SHAO L.: Zero-shot sketch image hashing. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 3598-3607, abs/ 1803.02284.

[TXW*21] TIAN J., XU X., WANG Z., SHEN F., LIU X.: Relationship Preserving Knowledge Distillation for Zero-shot Sketch Based Image Retrieval. *Proc. of the 29th ACM International Conference on Multimedia(ACMMM)*, 2021, pp. 5473-5481.

[WDLT21] WANG H., DENG C., LIU T., TAO D.: Transferable coupled network for zero-shot Sketch-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 44, (12), pp.9181-9194.

[WWY*21] WANG Z., WANG H., YAN J., WU A., DENG C.: Domain smoothing network for zero-shot Sketch-based image retrieval. *International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, 2021, arXiv:2106.11841.

[XDYW20] XU X., DENG C., YANG M., WANG H.: Progressive domain independent feature decomposition network for zero-shot sketch based image retrieval. *Proc. of the Twenty Ninth International Joint Conference on Artificial Intelligence(IJCAI)*, 2020, arXiv:2003.09869.

[YCL*16] YANG Y., CHEN W., LUO Y., SHEN F., SHAO J., SHEN H. T.: Zero-shot hashing via transferring supervised knowledge. *Proc. of the 24th ACM International Conference on Multimedia (ACMMM)*, 2016, pp. 1286-1295, abs/1606.05032.

[YRMM18] YELAMARTHI S. K., REDDY S. K., MISHRA A., MITTAL A.: A zero-shot framework for sketch based image retrieval. *The European Conference on Computer Vision (ECCV)*, 2018, pp. 300-317.

[YXQY18] YANG C., XIE L., QIAO S., YUILLE A. L.: Knowledge distillation in generations: More tolerant teachers educate better students. *ArXiv*, 2018, arXiv:1805.05551.

[YZYL18] H. YANG, X. ZHANG, F. YIN, C. LIU: Robust classification with convolutional prototype learning. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 3474-3482, abs/1805.03438.

[ZLW*22] Y.-W. ZHAN, X. LUO, Y. WANG, Z.-D. CHEN, X.-S. XU: Three-stream joint network for zero-shot sketch-based image retrieval. *Arxiv*, 2022, arXiv:2204.05666.

[ZLZ*16] H. ZHANG, S. LIU, C. ZHANG, W. REN, R. WANG, X. CAO: Sketchnet: Sketch classification with web images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1105–1113.

[ZS16] ZHANG Z., SALIGRAMA V.: Zero-shot learning via joint latent similarity embedding. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 6034–6042.

[ZXS*20] ZHU J., XU X., SHEN F., LEE R. K.-W., WANG Z., SHEN H. T.: Ocean: A dual learning approach for generalized zero-shot sketch-based image retrieval. *IEEE International Conference on Multimedia and Expo(ICME)*, 2020, pp. 1–6.