Semi-transductive Learning for Generalized Zero-Shot Sketch-Based Image Retrieval

Ce Ge[*], Jingyu Wang[*], Qi Qi, Haifeng Sun[†], Tong Xu, Jianxin Liao[†]

State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing 100876, China
{nwlgc, wangjingyu, qiqi8266, hfsun}@bupt.edu.cn, xutong@ebupt.com, jxlbupt@gmail.com

Abstract

Sketch-based image retrieval (SBIR) is an attractive research area where freehand sketches are used as queries to retrieve relevant images. Existing solutions have advanced the task to the challenging zero-shot setting (ZS-SBIR), where the trained models are tested on new classes without seen data. However, they are prone to overfitting under a realistic scenario when the test data includes both seen and unseen classes. In this paper, we study generalized ZS-SBIR (GZS-SBIR) and propose a novel semi-transductive learning paradigm. Transductive learning is performed on the image modality to explore the potential data distribution within unseen classes, and zero-shot learning is performed on the sketch modality sharing the learned knowledge through a semi-heterogeneous architecture. A hybrid metric learning strategy is proposed to establish semantics-aware ranking property and calibrate the joint embedding space. Extensive experiments are conducted on two large-scale benchmarks and four evaluation metrics. The results show that our method is superior over the state-of-the-art competitors in the challenging GZS-SBIR task.

Introduction

With the prevalence of touchscreen devices, hand-drawn input has been widely adopted and is infiltrating into every aspect of everyday life. Sketching therefore, as a simple, intuitive, and informative means of communication, is showing great potential in human-computer interaction. Sketch-based image retrieval (SBIR) is a representative multimedia retrieval application that aims to retrieve relevant images with freehand sketches. It has attracted wide attention in recent years and brought to various fruition in computer vision (Li et al. 2014; Saavedra and Barrios 2015; Liu et al. 2017; Song et al. 2017; Guo et al. 2017; Li, Zhou, and Yang 2019; Pang et al. 2019), computer graphics (Eitz et al. 2010, 2011; Shrivastava et al. 2011; Eitz, Hays, and Alexa 2012; Xu et al. 2013; Chen, Wang, and Liu 2020), and human-computer interaction (Huang, Canny, and Nichols 2019; Huang and Canny 2019).

The main difficulty in the SBIR community is the scarcity of hand-drawn sketches. Unlike photos that are easily accessible on the Internet, the acquisition of sketches relies heavily on manual drawing. With the explosive growth of multimedia content in daily life, the existing sketch datasets are always insufficient in scale and variety. Deployed retrieval systems trained on a limited number of classes (only a few hundred) most likely receive query sketches from new classes. The zero-shot SBIR (ZS-SBIR) (Shen et al. 2018) has emerged with the vision of generalizing the trained retrieval model to unseen classes, thereby bypassing the need to collect sketches for every novel category to work with.

ZS-SBIR shares the same challenges as conventional SBIR. As a cross-modal learning task, the main difficulty lies in bridging the domain gap between sketch and image. Freehand sketches are abstract and sparse line drawings, which require a feature extraction flow different to colorful images. Additionally, "zero-shot" means that no examples of test classes are available in the training phase. The retrieval model should have the ability to learn ranking property from seen classes and transfer the knowledge to unseen classes. Zero-shot learning itself is a non-trivial task and the introduction of sketch modality makes ZS-SBIR harder due to the high inter- and intra-class variances.

The usual solution for ZS-SBIR is to shape a cross-modal embedding space through deep metric learning. Both the query sketch and the candidate images are mapped to the high-dimensional feature space and the retrieval is performed by finding the nearest neighbors. However, since no data of unseen classes is available, the learned feature mapping likely overfits, and the embedding space biases toward the clusters of seen classes (Chao et al. 2016). Since ZS-SBIR is only evaluated on unseen classes, most retrieval models are deceived by the idealized setting and exhibit overestimated performance. Recently, the more realistic and challenging generalized ZS-SBIR (GZS-SBIR) is attracting more attention. Sketches and images from both the seen and unseen classes are mixed to form a new generalized test set. GZS-SBIR imposes higher requirements on preventing overfitting and balancing the semantic distribution.

It has been explored that generalized zero-shot learning can not be implemented without knowing any information about unseen classes (Liu et al. 2018). Side information, like attribute vectors or word embeddings in the semantic space, is needed to bridge the seen and unseen classes. Generally,

[*] These authors contributed equally.

[†] Corresponding authors.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

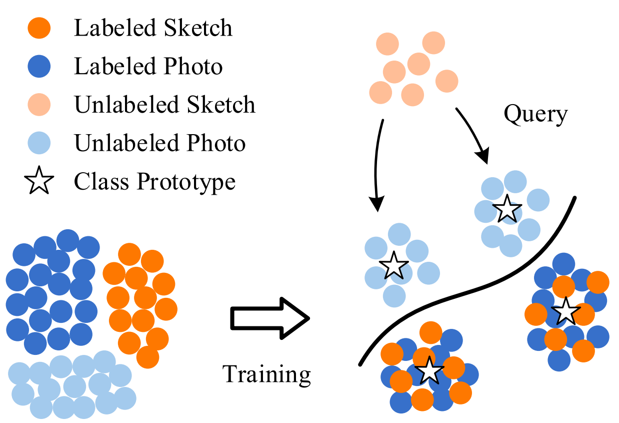


Figure 1: Schematics of semi-transductive GZS-SBIR.

a visual-semantic joint embedding space will be constructed to align visual features with corresponding class prototypes. Then the ranking property learned from seen classes is expected to be transferred to unseen classes through semantic relevance. However, the drawback is that this only connection between visual subspace and semantic subspace is weak. The semantic space is originally built via the vector similarity of class prototypes, while the data distribution in the visual space has a far larger variance between classes. The weak information conveyed by class semantics is insufficient to eliminate the mapping bias from unseen classes.

The key point of generalized zero-shot learning is to avoid overfitting to seen classes and increase the generalization toward unseen classes. We have realized that there is still great potential to improve GZS-SBIR if the model can perceive the latent data distribution of unseen classes. In this paper, we introduce a novel semi-transductive learning paradigm. The core idea is illustrated in Fig. 1. Based on the fact that images are easier to collect than sketches, we suggest conducting transductive learning in the image modality by utilizing unlabeled images of unseen classes. By exploring the potential adaptation with unlabeled images, the model is able to better generalize to unseen classes and reduce overfitting to seen classes. Since SBIR is a cross-modal learning problem, we transfer the learned image distribution of unseen classes to the sketch modality by sharing high-level learning paths through a semi-heterogeneous network structure. While learning a semantics-guided ranking property, a hybrid metric learning strategy is designed to calibrate the biases between seen and unseen classes. In the test phase, sketches and images from the the unseen classes will reuse the adapted feature encoding. They are mapped around the corresponding class prototypes and inherit ranking property. Toward a practical GZS-SBIR, the proposed semi-transductive learning paradigm is expected to benefit from the expansion and enrichment of the retrieval database without additional annotation effort.

The main contributions are summarized as follows:

- The novel semi-transductive learning paradigm is introduced to effectively improve generalized ZS-SBIR.
- A hybrid metric learning strategy is designed to calibrate semantic biases and learn semantics-guided ranking property.
- Extensive experiments are conducted on four evaluation metrics to show the effectiveness of our method.

Related Work

Sketch-Based Image Retrieval

Due to the inherent cross-modal nature of sketch-based image retrieval (SBIR), the main challenge lies in bridging the domain gap between freehand sketches and natural images. The breakthrough of deep learning has driven the trend of learning cross-domain high-level representations from raw pixels directly (Sangkloy et al. 2016; Yu et al. 2016). The current SBIR study has been identified as two separate fields based on the matching precision required: fine-grained (i.e., instance-level) SBIR (Song et al. 2017; Huo et al. 2018; Xu et al. 2018; Pang et al. 2019, 2020) and coarse-grained (i.e., category-level) SBIR (Sangkloy et al. 2016; Qi et al. 2016; Liu et al. 2017; Jiang, Xia, and Lu 2017; Li, Zhou, and Yang 2019). Fine-grained SBIR aims to capture fine-grained intra-class similarities, while coarse-grained SBIR performs a wider search among multiple categories. Zero-shot SBIR is a continuation of coarse-grained SBIR.

Zero-Shot Learning

Zero-shot learning (ZSL) in computer vision refers to identifying objects whose examples are not seen during training. Early works relied on manually annotated class attributes to infer unseen classes (Lampert, Nickisch, and Harmeling 2014; Al-Halah, Tapaswi, and Stiefelhagen 2016). With the help of convolutional neural networks (CNNs), recent studies focused on learning cross-modal embedding of images and class prototypes. These works can be roughly categorized into three types: (i) mapping visual features to the semantic space (Frome et al. 2013; Socher et al. 2013; Norouzi et al. 2014; Akata et al. 2016; Xian et al. 2016), (ii) mapping class semantics to the visual space (Changpinyo, Chao, and Sha 2017; Zhang, Xiang, and Gong 2017), and (iii) learning another common feature space (Zhang and Saligrama 2015; Fu et al. 2015; Zhang and Saligrama 2016). Our method falls into the first scheme of aligning visual space with semantic space, while additionally calibrating distribution and imposing ranking constraints in the embedding space.

Zero-Shot Sketch-Based Image Retrieval

Zero-shot sketch-based image retrieval (ZS-SBIR) is a combination of category-level SBIR and zero-shot learning. Shen et al. first investigated this problem and designed a deep three-branch network to learn cross-modal encoding and shared binary representations. Yelamarthi et al. proposed two autoencoder-based conditional generative models CAAE and CVAE and provided a new evaluation benchmark. Conditional generative models were further combined with inverse autoregressive flow (IAF) (Verma et al. 2019) and graph convolution network (GCN) (Zhang et al. 2020). In order to reduce the intra-class variance, adversarial training was adopted (Dutta and Akata 2019; Zhu et al. 2020;

Pandey et al. 2020; Dutta and Biswas 2020) to map the visual data to the semantic space by means of cycle-consistent generative models. Toward practical ZS-SBIR, Dey et al. combined triplet-based ranking metric, semantic reconstruction, and domain disentanglement to learn domain-agnostic cross-modal embedding. Our solution shares some similar concepts like semi-heterogeneity and semantic alignment (Liu et al. 2019; Chaudhuri et al. 2020).

The above methods are primarily designated to address *non-generalized* ZS-SBIR. There are very few studies that have specifically investigated GZS-SBIR (Dutta and Akata 2019; Dutta and Biswas 2020; Pandey et al. 2020). Recently, Zhu et al. proposed a dual learning cOmmon Conditional Encoder Adversarial Network (OCEAN) to address the semantic gap issue in GZS-SBIR. However, the knowledge transfer still only relied on semantic auxiliary information, thereby the encoding bias remaining unresolved. Our core point is that in addition to class prototypes, accessible abundant unlabeled images can be fully utilized to improve model generalization.

Problem Formulation

Generalized ZS-SBIR

Let $D = \{ X, Y \}$ be a database of sketches and images belonging to class set C. For regular SBIR, the goal is to retrieve images $y_i \in Y$ that belong to the same category as the query sketch $x_i \in X$, i.e., $\eta(x_i) = \eta(y_i)$, where $\eta : D \mapsto C$ is a labeling function. However, in practical applications, the database D cannot cover all common object categories. Therefore, the key is to make the trained retrieval models generalize to unseen data.

ZS-SBIR builds database D with two disjoint subsets $D^s = \{ X^s, Y^s \}$ and $D^u = \{ X^u, Y^u \}$ in accordance with *seen* classes C^s and *unseen* classes C^u, where $C^s \cap C^u = \emptyset$. Models are trained on seen data D^s but are expected to perform well on unseen data D^u. However, when data from seen and unseen classes are mixed for test, retrieval models designed for ZS-SBIR tend to overly bias toward seen classes. In order to evaluate generalized retrieval performance, GZS-SBIR extends the test set by including some seen-class data as $D^{te} = \{ X^u \cup X^{s\prime}, Y^u \cup Y^{s\prime} \}$, where $X^{s\prime} \subset X^s$ and $Y^{s\prime} \subset Y^s$.

In generalized zero-shot learning tasks, some form of auxiliary information is always needed to transfer the learned knowledge from seen classes to unseen classes. Typically, the embedding vectors of class words learned with natural language models are adopted as $E = \{ \vec{e}_i \mid \vec{e}_i \in \mathbb{R}^d, 1 \leq i \leq |C| \}$, where d is the pretrained embedding dimension. Although the semantic information of unseen classes is involved, the unseen image data is still unavailable in the training phase, which is considered to satisfy the essential hypothesis of zero-shot learning (Liu et al. 2018).

Semi-Transductive GZS-SBIR

SBIR is essentially a cross-modal learning task, in which the freehand sketches are scarce, whereas the natural images are abundant and easy to collect. In most real-word scenarios, the image database to be retrieved is bounded; thus a closed-world assumption is generally reasonable. We propose semi-transductive learning (STL) by including idle natural images to explore the potential data distribution of unseen classes. Specifically, unlabeled images of unseen classes participate in model training in a semi-supervised manner. The training set is extended as $D^{tr} = \{ X^s, Y^s \cup Y^u \}$, where the labels of Y^u are unknown. In this paradigm, the image modality forms transductive semi-supervised learning, while the sketch modality still performs zero-shot learning; thus we name it semi-transductive GZS-SBIR.

Methodology

Semi-Heterogeneous Encoding

We resolve sketch-based image retrieval by first extracting cross-modal feature encoding and then learning representation ranking. Given a query sketch $x_i \in X$, the retrieval model is expected to produce a ranking list of all candidate images $y \in Y$ satisfying the following criterion:

$$\mathrm{sim}(\phi(x_i), \psi(y_i)) > \mathrm{sim}(\phi(x_i), \psi(y_j)), \qquad (1)$$

where $\eta(x_i) = \eta(y_i)$ and $\eta(x_i) \neq \eta(y_j)$. The function $\mathrm{sim}(\cdot, \cdot)$ measures the vector similarity of any pair of sketch and image representation. The cosine similarity is employed in this work:

$$\mathrm{sim}(\vec{v}_1, \vec{v}_2) = \cos\langle \vec{v}_1, \vec{v}_2 \rangle = \frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_1\| \|\vec{v}_2\|} \qquad (2)$$

The mapping functions $\phi : X \mapsto \mathbb{R}^d$ and $\psi : Y \mapsto \mathbb{R}^d$ are the feature encoders that need to be learned for each modality. On one hand, since SBIR involves two instinct modalities (i.e, line drawings vs. colorful images), two modality-specific feature encoding paths are needed. On the other hand, the extracted visual features should be comparable in the joint embedding space, which means that the final representations should have certain modality invariance.

As shown in Fig. 2a, we designed two independent CNN feature encoders to extract modality-specific underlying visual features. They are built with the same backbone network but own separately trainable weights. As conventionally, two soft-attention modules (Dey et al. 2019) are plugged onto each CNN feature maps $\boldsymbol{F}$ to localize important regions and focus on representative features. The attention module learns a differentiable soft-mask $\boldsymbol{M}$ that indicates the importance of each neuron and re-weight the features maps through softmax as: $\boldsymbol{F} + \boldsymbol{F} \times \sigma_{\text{softmax}}(\boldsymbol{M})$.

The output feature volume of each modality-specific CNN is flattened and fed into a shared multi-layer perceptron (MLP). The shared MLP summarizes middle-level features and extracts high-level visual representations that are independent of modalities. This semi-heterogeneous design ensures the comparability of sketches and images in the common ranking space. Moreover, the shared part eliminates modality and thus enables knowledge sharing between sketches and images. This is the basis for our semi-transductive idea to work. As will be explained later and shown Fig. 2c, the learned information about unseen classes will be transmitted form unseen images to unseen query

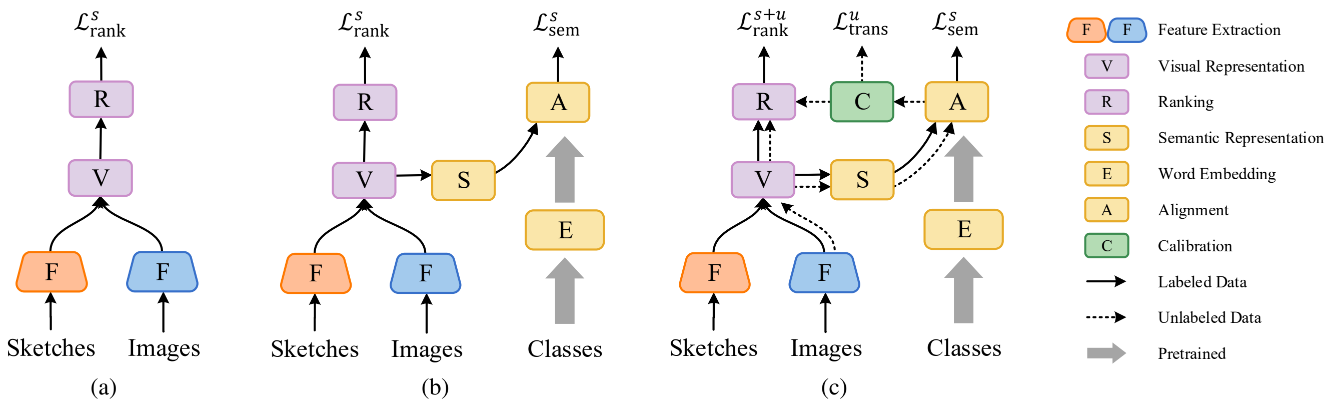


Figure 2: Model evolution. (a) Basic semi-heterogeneous network with ranking property learning. (b) Semantics extraction and alignment with class prototypes. (c) Our final model with distribution calibration and enhanced semi-transductive ranking.

sketches through the shared visual representation.

Ranking Property

We employ the margin-based triplet loss (Schroff, Kalenichenko, and Philbin 2015) to learn Eq. 1. First, a triplet $t = (x_i, y_i, y_j)$ is sampled from the seen data $\{ X^s, Y^s \}$, where $x_i \in X^s$ is the *anchor* sketch and $y_i, y_j \in Y^s$ are the *positive* and *negative* images, respectively, which satisfies $\eta(x_i) = \eta(y_i)$ and $\eta(x_i) \neq \eta(y_j)$. Their corresponding visual representations are extracted as $\vec{v}_a = \phi(x_i)$, $\vec{v}_p = \psi(y_i)$, and $\vec{v}_n = \psi(y_j)$. The goal is minimize the following cost:

$$\mathcal{L}_{\text{rank}}(t) = \max(0, \Delta + \text{dist}(\vec{v}_a, \vec{v}_p) - \text{dist}(\vec{v}_a, \vec{v}_n)), \quad (3)$$

where Δ is a hyper-parameter of margin. The margin-based metric learning strengthens the ranking property by forcing the negative image to be farther away than the positive image by some distance. The distance function $\text{dist}(\cdot,\cdot)$ is defined based on Eq. 1 as:

$$\text{dist}(\vec{v}_1, \vec{v}_2) = 1 - \cos(\vec{v}_1, \vec{v}_2). \quad (4)$$

After training, retrieval is performed by computing the distance (or equivalently, cosine similarity) between the query sketch and all candidate images to obtain a ranking list.

Discriminative Semantic Alignment

The problem with vanilla triplet ranking is that the model is supervised by seen classes C^s but never exposed to any data of unseen classes C^u (illustrated in Fig. 2a). The model has no knowledge about the relation between C^s and C^u. Thus the retrieval on new classes would be rather undetermined.

In order to transfer the ranking property from seen to unseen classes, we establish a linkage by aligning the visual feature space with the semantic space as shown in Fig. 2b. The relevance of class prototypes can provide guidance for the adaptation of unseen classes. The embedding vectors $E = \{ \vec{e}_i \mid \vec{e}_i \in \mathbb{R}^{300}, 1 \leq i \leq |C| \}$ learned by word2vec (Mikolov et al. 2013) are used as the initial auxiliary information. The class prototypes $H = \{ \vec{h}_i \mid \vec{h}_i \in \mathbb{R}^{d'}, 1 \leq i \leq |C| \}$ are obtained by transforming E with normalization to hold class-specific angular information as:

$$\vec{h}_i = \frac{\boldsymbol{W}^\top \vec{e}_i}{\|\boldsymbol{W}^\top \vec{e}_i\|_2} \cdot \lambda, \quad (5)$$

where $\boldsymbol{W}$ is the learnable weights and $\|\cdot\|_2$ denotes l_2-norm. The main purpose of scaling factor λ is to stabilize model training; its exact value is not sensitive and we set it to 10 throughout all experiments.

The semantic representation of sketches and images are extracted from their visual representation through a shared fully-connected layer $\zeta : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$. Then we calculate the vector similarity between the semantic representation (of sketches and images) and the class prototypes to obtain discriminative scores. Note that ζ also includes $l2$-normalization, so that cosine similarity $S(\cdot)$ can be easily implemented by dot product:

$$\mathcal{S}_c(x) = \zeta(\phi(x)) \cdot \vec{h}_c, \qquad x \in X^s, c \in C, \quad (6)$$

$$\mathcal{S}_c(y) = \zeta(\psi(y)) \cdot \vec{h}_c, \qquad y \in Y^s, c \in C. \quad (7)$$

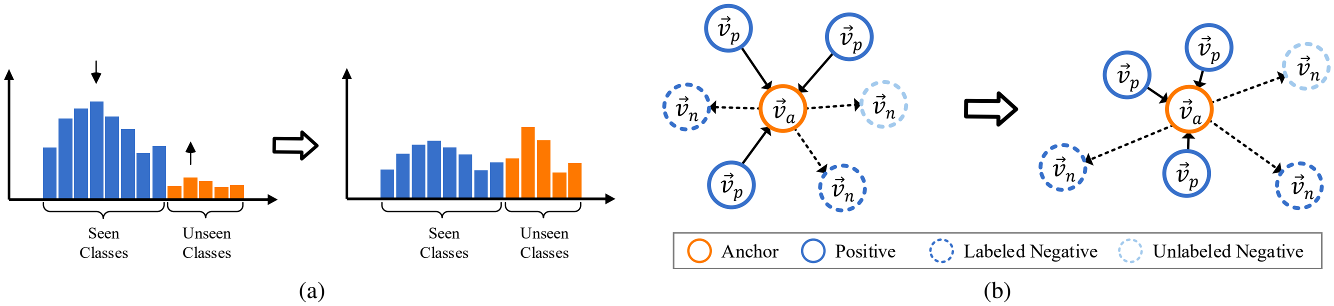
Several studies have reveled that CNNs make absolute confident scores on the supervised classes. Despite high discriminative accuracy, overfitting to seen classes generally hurts model generalization to unseen classes. In order to mitigate the overconfidence, we employ the temperature technique (Hinton, Vinyals, and Dean 2015) to reform the softmax prediction as:

$$\mathcal{P}_c(x) = \frac{\exp(\mathcal{S}_c(x) \cdot \gamma)}{\sum_{c' \in C} \exp(\mathcal{S}_{c'}(x) \cdot \gamma)}, \quad x \in X^s, c \in C, \quad (8)$$

where $\gamma \in \mathbb{N}$ is the temperature parameter. As $\gamma > 1$, the distribution becomes more skewed; as $\gamma < 1$, the distribution is softened (when $\gamma = 0$, all classes have equal probability). We empirically found that a value between 15 and 20 works well. Finally, the discriminative semantic alignment is trained on seen data by cross-entropy loss:

$$\mathcal{L}_{\text{sem}}(x) = \sum_{c \in C^s} \eta(x) \log \mathcal{P}_c(x), \qquad x \in X^s \quad (9)$$

$\mathcal{P}_c(y)$ and $\mathcal{L}_{\text{sem}}(y)$ for $y \in Y^s$ can be similarly defined.



(a) (b)

Figure 3: (a) Distribution calibration between seen and unseen classes. (b) Semi-transductive triplet ranking.

Semi-Transductive Adaptation

A semantics-aware sketch-based image retrieval model is built under the joint guidance of $\mathcal{L}_{\text{rank}}$ and $\mathcal{L}_{\text{sem}}$ (Fig. 2b). The learning paths $\phi(\cdot)$, $\psi(\cdot)$, and $\zeta(\cdot)$ will be reused when testing on new data. By sharing the semantic relevance established by the class prototypes, unseen data is likely mapped to more suitable positions in the visual embedding space. However, the feature encoding is still biased due to the high inter-class variances.

We calibrate the biased distribution by utilizing unlabeled images Y^u from unseen classes C^u. First, the discriminative scores of an unlabeled image $y^u \in Y^u$ can be predicted according to Eq. 7. Due to overfitting, the top prediction generally falls in seen classes, i.e.,

$$\operatorname*{argmax}_{c\in C} \mathcal{P}_c(y) \in C^s, \qquad y \in Y^u. \tag{10}$$

We formalize the problem as a binary classification task: *whether the highest prediction belongs to seen or unseen classes?* Consequently, a variant of hinge loss is used to calibrate the distribution of unlabeled images $y \in Y^u$:

$$\mathcal{L}_{\text{trans}}(y) = \max(0,\ \delta + \max_{c\in C^s} \mathcal{P}_c(y) - \max_{c'\in C^u} \mathcal{P}_{c'}(y)). \tag{11}$$

The model weights are updated in the direction of decreasing the predictions on seen classes and increasing the confidences on unseen classes. Conversely, the loss is claimed by zero. Finally, the highest prediction on unseen classes should be at least δ higher than that on seen classes. The process is conceptually shown in Fig. 3a.

Based on the calibrated distribution, we can assign each unlabeled image $y \in Y^u$ a pseudo-label as:

$$\eta(y) = \operatorname{argmax}_{c\in C^u} \mathcal{P}_c(y) \tag{12}$$

The unlabeled images with their predicted labels are incorporated into ranking property learning to reflect the calibration, i.e., the triplet set $\mathcal{T}^{s+u}$ is now sampled from $\{X^s, Y^s \cup Y^u\}$ rather than $\{X^s, Y^s\}$. Please refer to Fig. 2c and Fig. 3b for intuitive concepts.

Learning Objectives

The final objective is to minimize the weighted losses:

$$\min_{\boldsymbol{\Theta}} \left(\mathcal{L}^{s+u}_{\text{rank}} + \omega_1 \mathcal{L}^s_{\text{sem}} + \omega_2 \mathcal{L}^u_{\text{trans}}\right), \tag{13}$$

where ω_1 and ω_2 are weighting factors, and

$$\mathcal{L}^{s+u}_{\text{rank}} = \frac{1}{|\mathcal{T}^{s+u}|} \sum_{t\in\mathcal{T}^{s+u}} \mathcal{L}_{\text{rank}}(t), \tag{14}$$

$$\mathcal{L}^s_{\text{sem}} = \frac{1}{|X^s \cup Y^s|} \left(\sum_{x\in X^s} \mathcal{L}_{\text{sem}}(x) + \sum_{y\in Y^s} \mathcal{L}_{\text{sem}}(y) \right), \tag{15}$$

$$\mathcal{L}^u_{\text{trans}} = \frac{1}{|Y^u|} \sum_{y\in Y^u} \mathcal{L}_{\text{trans}}(y). \tag{16}$$

Experiments

Settings

Datasets. We employ two widely used SBIR datasets: Sketchy (Sangkloy et al. 2016) and TU-Berlin (Eitz, Hays, and Alexa 2012). Sketchy consists of 75,471 freehand sketches and 12,500 photos belonging to 125 object categories. Liu et al. enriched the dataset by collecting additional 60,502 natural images from ImageNet (Deng et al. 2009). The TU-Berlin dataset includes 250 categories, each with 80 hand-drawn sketches. Zhang et al. extended it with 191,067 real images crawled from ImageNet and search engines, thus enabling SBIR experiments. We split each dataset into seen and unseen classes following conventions (Dey et al. 2019). During evaluation, 20% seen classes and all unseen classes are mixed to form the generalized test set (Zhu et al. 2020).

Evaluation Metrics. Several different evaluation metrics are used in the literature. Some studies reported mean Average Precision (mAP) on the whole dataset and Precision on the top-100 recalled images (P@100) (Shen et al. 2018; Dutta and Akata 2019; Dey et al. 2019; Liu et al. 2019; Chaudhuri et al. 2020; Dutta and Biswas 2020; Zhu et al. 2020). While others considered the top-200 recalls and reported mAP@200 and P@200 (Yelamarthi et al. 2018; Verma et al. 2019; Dey et al. 2019; Liu et al. 2019; Pandey et al. 2020; Dutta and Biswas 2020; Zhu et al. 2020). We endeavored to gather all relevant competitors and conduct a comprehensive assessment on four common metrics.

Implementation Details. The feature encoders $\phi(\cdot)$ and $\psi(\cdot)$ make use of the convolutional layers of VGG-16 net

Method	Sketchy		TU-Berlin	
	mAP(%)	P@100(%)	mAP(%)	P@100(%)
ZSIH (Shen et al. 2018)	21.9	29.6	14.2	21.8
SEM-PCYC (Dutta and Akata 2019)	30.7	36.4	19.2	29.8
CSDB (Dutta and Biswas 2020)	33.1	38.1	14.9	22.6
OCEAN (Zhu et al. 2020)	44.5	54.8	31.2	34.1
STL (Ours)	**53.0**±0.3	**58.1**±0.3	**40.2**±0.4	**49.8**±0.5

Table 1: GZS-SBIR Performance in terms of mAP and P@100.

Method	Sketchy		TU-Berlin	
	mAP@200(%)	P@200(%)	mAP@200(%)	P@200(%)
CAAE (Yelamarthi et al. 2018)	12.4	18.6	9.1	16.2
CVAE (Yelamarthi et al. 2018)	13.4	20.2	9.9	17.7
SAN (Pandey et al. 2020)	22.7	30.4	12.4	20.3
SEM-PCYC (Dutta and Akata 2019)	35.5	32.7	30.1	26.7
OCEAN (Zhu et al. 2020)	54.7	44.3	36.9	31.9
STL (Ours)	**63.4**±0.2	**53.8**±0.3	**52.9**±0.3	**46.7**±0.5

Table 2: GZS-SBIR Performance in terms of mAP@200 and P@200.

	Sketchy	TU-Berlin
$\mathcal{L}^{s}_{\text{rank}}$ (Heterogeneous)	8.7	5.3
$\mathcal{L}^{s}_{\text{rank}}$ (Siamese)	11.2	6.7
$\mathcal{L}^{s}_{\text{rank}}$	16.3	9.6
$\mathcal{L}^{s}_{\text{rank}} + \mathcal{L}^{s}_{\text{sem}}$	27.4	13.8
$\mathcal{L}^{s+u}_{\text{rank}} + \mathcal{L}^{s}_{\text{sem}} + \mathcal{L}^{u}_{\text{trans}}$	53.0	40.2

Table 3: Ablation study for GZS-SBIR in terms of mAP.

(Simonyan and Zisserman 2014) that is pre-trained on ImageNet (Deng et al. 2009). The attention mechanism upon feature maps is implemented using 1×1 convolutions. The shared MLP for visual representation consist of the last three fully connected layers in VGG-16, among which the first two layers are followed by ReLU non-linearity and dropout regularization with probability 0.5. The feature dimension of the embedding space is set to 1024D. Note that although our model introduces more computation during training, only the feature encoder and nearest neighbor search are used in the test phase. The test-time computational cost is in the same order of magnitude as the recent competitive methods.

Hyperparameters. The weighting factors for each dataset are determined by grid search with $\omega_1 \in [0.01, 1]$ and $\omega_2 \in [0.001, 10]$. For Sketchy, $\omega_1 = 0.5$, $\omega_2 = 0.1$, and for TU-Berlin $\omega_1 = 0.5$, $\omega_2 = 0.5$. The margin hyperparameters in $\mathcal{L}_{\text{rank}}$ (Eq. 3) and $\mathcal{L}_{\text{trans}}$ (Eq. 11) are empirically set to $\Delta = 0.1$ and $\delta = 0.01$, respectively. The whole model is implemented on top of PyTorch (Paszke et al. 2019) and is trained end-to-end by stochastic gradient descent with learning rate 1e-3 and a mini-batch size 20. The early stopping strategy is adopted to combat overfitting.

Quantitative Results

The quantitative results for generalized zero-shot sketch-based image retrieval are presented in Tab. 1 and Tab. 2. We conducted five independent experiments and reported the mean and standard deviation to be statistically significant. Our model is denoted as STL (acronym for Semi-Transductive Learning) in the last row.

It can be seen that our solution is superior to all other state-of-the-art alternatives by a large margin. On the Sketchy dataset, our approach has achieved an improvement of more than 8% on mAP, mAP@200, and P@200 compared with the *OCEAN* method. On the TU-Berlin dataset, our model also outperforms *OCEAN* by 9% on mAP (40.2% vs. 31.2%). More significantly, our method gains about 15%–16% improvements on P@100, mAP@200, and P@200 compared with *OCEAN* method. Comparing Tab. 1 and Tab. 2, the performance of all methods on the TU-Berlin dataset is obviously lower than that on the Sketchy dataset. We attribute it to the inherent hierarchy ambiguity in TU-Berlin, which will be further explained in Sec. Visualization.

In terms of evaluation metrics, P@100 and P@200 only measure the retrieval precision under specific recalls, which are usually unstable indicators. The slightly larger standard deviation reported in the tables may confirm this. The mAP (i.e., mAP@all) and mAP@200 metrics deserve special attention as they indicate the average retrieval performance.

Ablation Study

Our final model is learned under the joint supervision of three losses. We gradually remove modules to investigate their contributions. The results are shown in Tab. 3. Obviously, $\mathcal{L}^{u}_{\text{trans}}$ makes the greatest contribution. In fact, once it is removed, the semi-transductive learning paradigm degen-

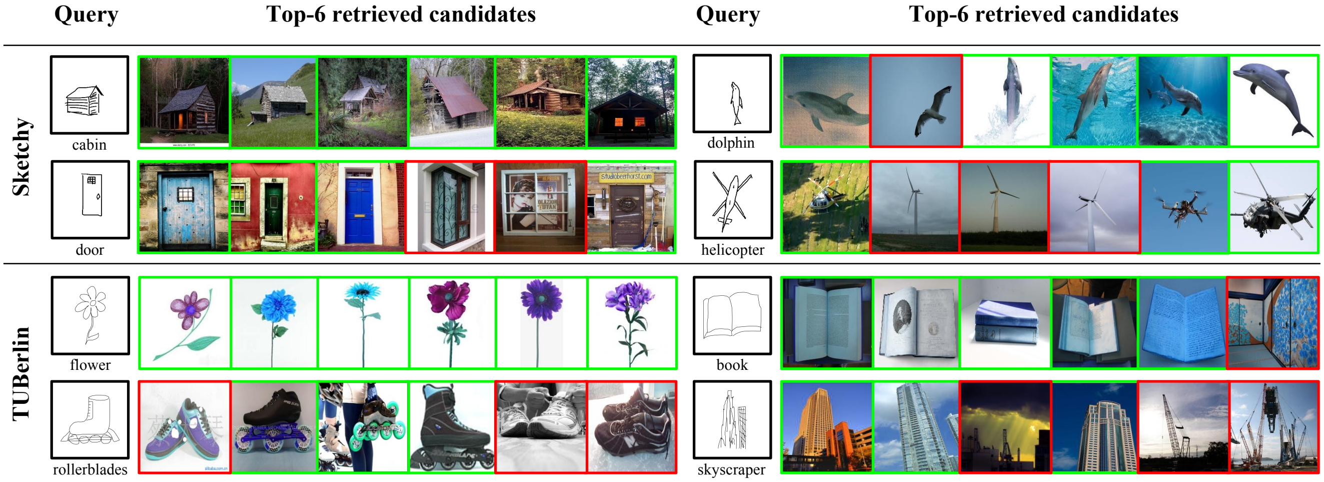


Figure 4: Visualization of the top-6 retrieval results on Sketchy and TU-Berlin datasets for generalized zero-shot sketch-based image retrieval. The green and red bounding boxes indicate true and false retrieved images, respectively.

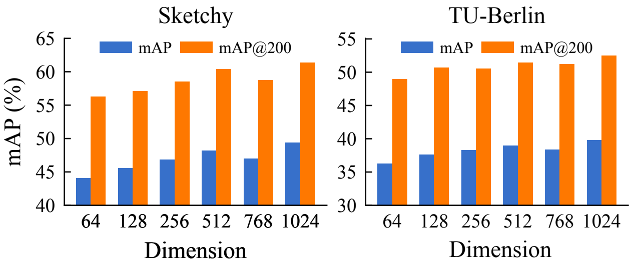


Figure 5: Effect analysis of embedding dimension.

erates into ordinary zero-shot learning. The *Ranking* module loses the ability to explore unlabeled data due to the linkage. In short, our model is a tight whole, and the components depend on each other to work properly. In addition, the adopted semi-heterogeneous backbone is more suitable for cross-modal learning than other alternatives. It can be speculated that the large number of learnable parameters in the heterogeneous structure increases the risk of overfitting.

Analysis

The recent trend of CNN-based deep retrieval models is to embed images into a high-dimensional feature space, e.g., 512D in *OCEAN* and 1024D in *SketchGCN*. We experimented with lower embedding dimensions and retrained our models for GZS-SBIR. The results of representative mAP and mAP@200 are shown in Fig. 5. Overall, the performance shows an upward trend with the increase of embedding dimension. It is noteworthy that our model is still competitive even if the embedding space is compressed to 64D. Our 64D model has achieved mAP 44.6% and mAP@200 56.3% on Sketchy, which are superior to the 512D *OCEAN* model in Tab. 1 (44.5%) and Tab. 2 (54.7%). Similarly, it also outperforms *OCEAN* on TU-Berlin with mAP 36.3% vs. 31.2% and mAP@200 49.0% vs. 36.9%.

Visualization

Fig. 4 visualizes several retrieval examples produced by our method. The trained model has the ability to retrieve relevant images with high semantic and visual similarity whether for simple (e.g., *flower* in TU-Berlin) or complex objects (e.g., *cabin* in Sketchy). For freehand sketches that may cause ambiguity, even if the retrieved images belong to wrong classes, they still show a high degree of visual similarity. Clues can be found from *dolphin*, *helicopter*, *book*, and *skyscraper* examples. Also note that several *shoe* images are retrieved for the *rollerblades* query. Although assessed as failures, they are indeed relevant in semantics and appearance. This reflects a inherent problem in TU-Berlin that the definition of category hierarchy is somewhat confusing: (i) classes from different WordNet (Miller 1995) levels are mixed (e.g., *mug* vs. *beer-mug*), and (ii) semantic concepts and attribute descriptions are used together (e.g., *seagull* vs. *flying bird*).

Conclusion

In this study, we have re-examined the challenging and realistic generalized zero-shot sketch-based image retrieval task. A novel semi-transductive learning paradigm is proposed to improve the generalization ability of zero-shot SBIR models on both seen and unseen classes. We propose to transfer cross-modal domain knowledge from unlabeled image data through a semi-heterogeneous feature encoder. A hybrid metric learning strategy with semantic alignment and distribution calibration is designed to adapt the retrieval model to the target domain. The proposed method has been evaluated on two large-scale benchmark datasets for the challenging GZS-SBIR task, and extensive experiments confirmed the superiority of our solution over various state-of-the-art competitors.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants

(62171057, 62201072, 62071067, 62101064, 62001054), China Postdoctoral Science Foundation under Grant 2022M710468, in part by the Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center, in part by BUPT innovation and entrepreneurship support program.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2016. Label-Embedding for Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7): 1425–1438.

Al-Halah, Z.; Tapaswi, M.; and Stiefelhagen, R. 2016. Recovering the Missing Link: Predicting Class-Attribute Associations for Unsupervised Zero-Shot Learning. In *Proc. IEEE Int. Conf. Comput. Vis.*, 5975–5984.

Changpinyo, S.; Chao, W.-L.; and Sha, F. 2017. Predicting Visual Exemplars of Unseen Classes for Zero-Shot Learning. In *Proc. IEEE Int. Conf. Comput. Vis.*, 3496–3505.

Chao, W.-L.; Changpinyo, S.; Gong, B.; and Sha, F. 2016. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In *Proc. 14th Eur. Conf. Comput. Vis.*, volume 9906, 52–68.

Chaudhuri, U.; Banerjee, B.; Bhattacharya, A.; and Datcu, M. 2020. A simplified framework for zero-shot cross-modal sketch data retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 699–706.

Chen, M.; Wang, C.; and Liu, L. 2020. Cross-domain retrieving sketch and shape using cycle CNNs. *Comput. Graph.*, 89: 50–58.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Kai Li; and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 20, 248–255.

Dey, S.; Riba, P.; Dutta, A.; Llados, J. L.; and Song, Y.-Z. 2019. Doodle to Search: Practical Zero-Shot Sketch-Based Image Retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, i, 2174–2183.

Dutta, A.; and Akata, Z. 2019. Semantically Tied Paired Cycle Consistency for Zero-Shot Sketch-Based Image Retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 5084–5093.

Dutta, T.; and Biswas, S. 2020. Style-guided zero-shot sketch-based image retrieval. In *Proc. Br. Mach. Vis. Conf.*

Eitz, M.; Hays, J.; and Alexa, M. 2012. How do humans sketch objects? *ACM Trans. Graph.*, 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.*, 34(5): 482–498.

Eitz, M.; Hildebrand, K.; Boubekeur, T.; and Alexa, M. 2011. Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors. *IEEE Trans. Vis. Comput. Graph.*, 17(11): 1624–1636.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M. A.; and Mikolov, T. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Proc. Annu. Conf. Neural Inf. Process. Syst.*, volume 26, 2121–2129.

Fu, Z.; Xiang, T. A.; Kodirov, E.; and Gong, S. 2015. Zero-shot object recognition by semantic manifold distance. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2635–2644.

Guo, L.; Liu, J.; Wang, Y.; Luo, Z.; Wen, W.; and Lu, H. 2017. Sketch-based Image Retrieval using Generative Adversarial Networks. In *Proc. 25th ACM Int. Conf. Multimed.*, 1267–1268.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.0: 1–9.

Huang, F.; and Canny, J. F. 2019. Sketchforme: Composing Sketched Scenes from Text Descriptions for Interactive Applications. In *Proc. 32nd Annu. ACM Symp. User Interface Softw. Technol.*, 209–220.

Huang, F.; Canny, J. F.; and Nichols, J. 2019. Swire: Sketch-based User Interface Retrieval Forrest. In *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.*, 1–10.

Huo, Q.; Wang, J.; Qi, Q.; Sun, H.; Ge, C.; and Zhao, Y. 2018. Users Personalized Sketch-Based Image Retrieval Using Deep Transfer Learning. In *Proc. 11th Int. Conf. Knowl. Sci. Eng. Manag.*, volume 3, 160–168.

Jiang, T.; Xia, G.-S.; and Lu, Q. 2017. Sketch-based aerial image retrieval. In *Proc. IEEE Int. Conf. Image Process.*, 3690–3694.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3): 453–465.

Li, C.; Zhou, Y.; and Yang, J. 2019. Sketch-Based Image Retrieval via a Semi-Heterogeneous Cross-Domain Network. In *Proc. 2019 IEEE Int. Conf. Multimed. Expo Work.*, 216–221.

Li, Y.; Hospedales, T. M.; Song, Y. Z.; and Gong, S. 2014. Fine-grained sketch-based image retrieval by matching deformable part models. In *Proc. Br. Mach. Vis. Conf.*

Liu, L.; Shen, F.; Shen, Y.; Liu, X.; and Shao, L. 2017. Deep Sketch Hashing: Fast Free-Hand Sketch-Based Image Retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2298–2307.

Liu, Q.; Xie, L.; Wang, H.; and Yuille, A. 2019. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proc. IEEE Int. Conf. Comput. Vis.*, 3661–3670.

Liu, S.; Long, M.; Wang, J.; and Jordan, M. I. 2018. Generalized zero-shot learning with deep calibration network. In *Proc. Annu. Conf. Neural Inf. Process. Syst.*, NeurIPS, 2005–2015.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. 1st Int. Conf. Learn. Represent.*, 1–12.

Miller, G. A. 1995. WordNet: a lexical database for English. *Commun. ACM*, 38(11): 39–41.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.;

Frome, A.; Corrado, G. S.; and Dean, J. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *Proc. 2nd Int. Conf. Learn. Represent.*, 1–9.

Pandey, A.; Mishra, A.; Verma, V. K.; Mittal, A.; and Murthy, H. A. 2020. Stacked adversarial network for zero-shot sketch based image retrieval. In *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2529–2538.

Pang, K.; Li, K.; Yang, Y.; Zhang, H.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2019. Generalising Fine-Grained Sketch-Based Image Retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 677–686.

Pang, K.; Yang, Y.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2020. Solving Mixed-Modal Jigsaw Puzzle for Fine-Grained Sketch-Based Image Retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 10344–10352.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 8024–8035.

Qi, Y.; Song, Y.-Z.; Zhang, H.; and Liu, J. 2016. Sketch-based image retrieval via Siamese convolutional neural network. In *Proc. IEEE Int. Conf. Image Process.*, 2460–2464.

Saavedra, J. M.; and Barrios, J. M. 2015. Sketch based Image Retrieval using Learned KeyShapes (LKS). In *Procedings Br. Mach. Vis. Conf.*, 164.1–164.11.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM Trans. Graph.*, 35(4): 1–12.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 11, 815–823.

Shen, Y.; Liu, L.; Shen, F.; and Shao, L. 2018. Zero-Shot Sketch-Image Hashing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 3598–3607.

Shrivastava, A.; Malisiewicz, T.; Gupta, A.; and Efros, A. A. 2011. Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph.*, 30(6).

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. 3rd Int. Conf. Learn. Represent.*, 1–14.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-Shot Learning Through Cross-Modal Transfer. In *Adv. Neural Inf. Process. Syst.*, volume 26.

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 *Proc. IEEE Int. Conf. Comput. Vis.*, 5552–5561.

Verma, V. K.; Mishra, A.; Mishra, A.; and Rai, P. 2019. Generative model for zero-shot sketch-based image retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, 704–713.

Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; and Schiele, B. 2016. Latent Embeddings for Zero-Shot Classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 69–77.

Xu, K.; Chen, K.; Fu, H.; Sun, W. L.; and Hu, S. M. 2013. Sketch2Scene: Sketch-based co-retrieval and co-placement of 3D models. *ACM Trans. Graph.*, 32(4): 1–12.

Xu, P.; Yin, Q.; Huang, Y.; Song, Y.-Z.; Ma, Z.; Wang, L.; Xiang, T.; Kleijn, W. B.; and Guo, J. 2018. Cross-modal subspace learning for fine-grained sketch-based image retrieval. *Neurocomputing*, 278: 75–86.

Yelamarthi, S. K.; Reddy, S. K.; Mishra, A.; and Mittal, A. 2018. A Zero-Shot Framework for Sketch Based Image Retrieval. In *Procedings 15th Eur. Conf. Comput. Vis.*, volume 11208, 316–333.

Yu, Q.; Liu, F.; Song, Y. Z.; Xiang, T.; Hospedales, T. M.; and Loy, C. C. 2016. Sketch me that shoe. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 799–807.

Zhang, H.; Liu, S.; Zhang, C.; Ren, W.; Wang, R.; and Cao, X. 2016. SketchNet: Sketch Classification with Web Images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1105–1113.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a Deep Embedding Model for Zero-Shot Learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 3010–3019.

Zhang, Z.; and Saligrama, V. 2015. Zero-Shot Learning via Semantic Similarity Embedding. In *Proc. IEEE Int. Conf. Comput. Vis.*, 4166–4174.

Zhang, Z.; and Saligrama, V. 2016. Zero-Shot Learning via Joint Latent Similarity Embedding. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 6034–6042.

Zhang, Z.; Zhang, Y.; Feng, R.; Zhang, T.; and Fan, W. 2020. Zero-Shot Sketch-based Image Retrieval via Graph Convolution Network. In *Proc. 24th AAAI Conf. Artif. Intell.*

Zhu, J.; Xu, X.; Shen, F.; Lee, R. K.-w.; Wang, Z.; and Shen, H. T. 2020. Ocean: A Dual Learning Approach For Generalized Zero-Shot Sketch-Based Image Retrieval. In *Proc. IEEE Int. Conf. Multimed. Expo*, 1–6.