Semantically Tied Paired Cycle Consistency for Zero-Shot Sketch-based Image Retrieval

Anjan Dutta
Computer Vision Center
Autonomous University of Barcelona
adutta@cvc.uab.es

Zeynep Akata
Amsterdam Machine Learning Lab
University of Amsterdam
z.akata@uva.nl

Abstract

Zero-shot sketch-based image retrieval (SBIR) is an emerging task in computer vision, allowing to retrieve natural images relevant to sketch queries that might not been seen in the training phase. Existing works either require aligned sketch-image pairs or inefficient memory fusion layer for mapping the visual information to a semantic space. In this work, we propose a semantically aligned paired cycle-consistent generative (SEM-PCYC) model for zero-shot SBIR, where each branch maps the visual information to a common semantic space via an adversarial training. Each of these branches maintains a cycle consistency that only requires supervision at category levels, and avoids the need of highly-priced aligned sketch-image pairs. A classification criteria on the generators' outputs ensures the visual to semantic space mapping to be discriminating. Furthermore, we propose to combine textual and hierarchical side information via a feature selection autoencoder that selects discriminating side information within a same end-to-end model. Our results demonstrate a significant boost in zero-shot SBIR performance over the state-of-the-art on the challenging Sketchy and TU-Berlin datasets.

1. Introduction

Matching natural images with free-hand sketches, *i.e.* *sketch-based image retrieval* (SBIR) [60, 58, 27, 33, 47, 43, 63, 7, 23] has received a lot of attention. Since sketches can effectively express shape, pose and fine-grained details of the target images, SBIR serves a favorable scenario complementary to the conventional text-image cross-modal retrieval or the classical content based image retrieval protocol. This is also because in some situations it may be hard to provide a textual description or a suitable image of the desired query, whereas, an user can easily draw a sketch of the desired object spontaneously on a touch screen.
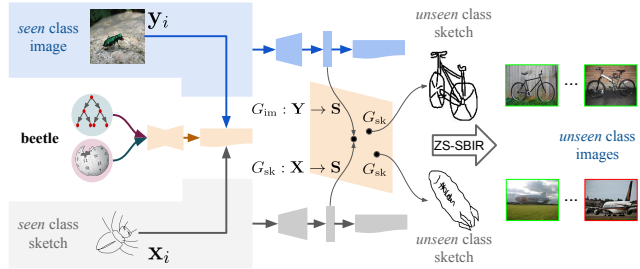

Figure 1. The proposed SEM-PCYC model learns to map visual information from sketch and image to a semantic space through an adversarial training based on the *seen* classes. During the testing phase the learned mappings are used for generating embeddings on the *unseen* classes for zero-shot SBIR.

As the visual information from all the classes gets explored by the system during training, with overlapping training and test classes, existing SBIR methods perform well [63]. Since in practice there is no guarantee that the training data would include all possible queries, a more realistic setting is *zero-shot* SBIR [43, 23] which combines zero-shot learning (ZSL) [25, 54] and SBIR as a single task, where the aim is an accurate class prediction and a competent retrieval performance. However, zero-shot SBIR is extremely challenging as it simultaneously deals with a significant domain gap, intra-class variability and limited knowledge about the *unseen* classes.

One of the major shortcomings of the prior work on ZS-SBIR is that sketch-image is retrieved after learning a mapping from an input sketch to an output image using a training set of labelled *aligned* pairs [23]. The supervision of paired correspondence is to enhance the correlation of multi-modal data (here, sketch-image) so that learning can be guided by semantics. However, for many realistic scenarios, obtaining paired (aligned) training data is either unavailable or very expensive. Furthermore, often a joint representation of two or more modalities is obtained by using a memory fusion layer [43], such as, tensor fusion [19], bilinear pooling [62] etc. These fusion layers are often expensive

in terms of memory [62], and extracting useful information from this high dimensional space could result in information loss [61].

To alleviate these shortcomings, we propose a semantically aligned paired cycle consistent generative (SEM-PCYC) model for zero-shot SBIR task, where each branch either maps sketch or image features to a common semantic space via an adversarial training. These two branches dealing with two different modalities (sketch and image) constitute an essential component for solving SBIR task. The cycle consistency constraint on each branch guarantees the mapping of sketch or image modality to a common semantic space and their translation back to the original modality, which further avoids the necessity of aligned sketch-image pairs. Imposing a classification loss on the semantically aligned outputs from the sketch and image space enforces the generated features in the semantic space to be discriminative which is very crucial for effective zero-shot SBIR. Furthermore, inspired by the previous works on label embedding [3], we propose to combine side information from text-based and hierarchical models via a feature selection auto-encoder [51] which selects discriminating side information based on intra and inter class covariance.

The main contributions of the paper are: (1) We propose the SEM-PCYC model for zero-shot SBIR task, that maps sketch and image features to a common semantic space with the help of adversarial training. The cycle consistency constraint on each branch of the SEM-PCYC model facilitates bypassing the requirement of aligned sketch image pairs. (2) Within a same end-to-end framework, we combine different side information via a feature selection guided auto-encoder which effectively choose side information that minimizes intra-class variance and maximizes inter-class variance. (3) We evaluate our model on two datasets (Sketchy and TU-Berlin) with varying difficulties and sizes, and provide an experimental comparison with latest models available for the same task, which further shows that our proposed model consistently improves the state-of-the-art results of zero-shot SBIR on both datasets.

2. Related Work

As our work belongs at the verge of sketch-based image retrieval and zero-shot learning task, we briefly review the relevant literature from both the fields.

Sketch Based Image Retrieval (SBIR). Attempts for solving SBIR task mostly focus on bridging the domain gap between sketch and image, which can roughly be grouped in *hand-crafted* and *cross-domain deep learning-based* methods [27]. Hand-crafted methods mostly work by extracting the edge map from natural image and then matching them with sketch using a Bag-of-Words model on top of some specifically designed SBIR features, *viz.*, gradient field HOG [20], histogram of oriented edges [40], learned key shapes [41] etc. However, the difficulty of reducing domain gap remained unresolved as it is extremely challenging to match edge maps with unaligned hand drawn sketch. This domain shift issue is further addressed by neural network models where domain transferable features from sketch to image are learned in an end-to-end manner. Majority of such models use variant of siamese networks [36, 42, 58, 46] that are suitable for cross-modal retrieval. These frameworks either use generic ranking losses, *viz.*, contrastive loss [9], triplet ranking loss [42] or more sophisticated HOLEF based loss [47]) for the same. Further to these discriminative losses, Pang *et al*. [33] introduced a discriminative-generative hybrid model for preserving all the domain invariant information useful for reducing the domain gap between sketch and image. Alternatively, some other works focus on learning cross-modal hash code for category level SBIR within an end-to-end deep model [27, 63]. In contrast, we propose a paired cycle consistent generative model where each branch either maps sketch or image features to a common semantic space via adversarial training, which we found to be effective for reducing the domain gap between sketch and image.

Zero-Shot Learning (ZSL). Zero-shot learning in computer vision refers to recognizing objects whose instances are not seen during the training phase; a comprehensive and detailed survey on ZSL is available in [54]. Early works on ZSL [25, 21, 5, 4] make use of attributes within a two-stage approach to infer the label of an image that belong to the *unseen* classes. However, the recent works [15, 39, 3, 2, 24] directly learn a mapping from image feature space to a semantic space. Many other ZSL approaches learn non-linear multi-modal embedding [45, 2, 53, 6, 64], where most of the methods focus to learn a non-linear mapping from the image space to the semantic space. Mapping both image and semantic features into another common intermediate space is another direction that ZSL approaches adapt [66, 16, 67, 1, 28]. Although, most of the deep neural network models in this domain are trained using a discriminative loss function, a few generative models also exist [52, 55, 8] that are used as a data augmentation mechanism. In ZSL, some form of side information is required, so that the knowledge learned from *seen* classes gets transferred to *unseen* classes. One popular form of side information is attributes [25] that, however, require costly expert annotation. Thus, there has been a large group of studies [29, 3, 53, 38, 37, 11] which utilize other auxiliary information, such as, text-based [30] or hierarchical model [32] for label embedding. In this work, we address zero-shot cross-modal (sketch to image) retrieval, for that, motivated by [3], we effectively combine different side information within an end-to-end framework, and map visual information to the semantic space through an adversarial training.

Zero-Shot Sketch-based Image Retrieval (ZS-SBIR).
Shen *et al.* [43] first combined zero-shot learning and sketch based image retrieval, and proposed a generative cross-modal hashing scheme for solving the zero-shot SBIR task, where they used a graph convolution network for aligning the sketch and image in the semantic space. Inspired by them, Yelamarthi *et al.* [23] proposed two similar autoencoder-based generative models for zero-shot SBIR, where they have used the aligned pairs of sketch and image for learning the semantics between them. In contrast, we propose a paired cycle-consistent generative model where each branch maps the visual information from sketch or image to a semantic space through an adversarial training with a common discriminator. The cycle consistency constraint on each branch allows supervision only at category level, and avoids the need of aligned sketch-image pairs.

3. SEM-PCYC Model

In this work, we propose the semantically aligned paired cycle consistent generative (SEM-PCYC) model for zero-shot sketch-based image retrieval. The sketch and image data from the *seen* categories are only used for training the underlying model. Our SEM-PCYC model encodes and matches the sketch and image categories that remain *unseen* during the training phase. The overall pipeline of our end-to-end deep architecture is shown in Figure 2.

Let $\mathcal{D}^s = \{\mathbf{X}^s, \mathbf{Y}^s\}$ be a collection of sketch and image data from the *seen* categories $\mathcal{C}^s$ that contain sketch images $\mathbf{X}^s = \{\mathbf{x}_i^s\}_{i=1}^N$ as well as natural images $\mathbf{Y}^s = \{\mathbf{y}_i^s\}_{i=1}^N$ for training, where N is the total number of sketch and image pairs that are not necessarily aligned. Without loss of generality, it can be assumed that sketch and image having the same index, say, i, share the same category label. Let $\mathbf{S}^s = \{\mathbf{s}_i^s\}_{i=1}^N$ be the set of side information useful for transferring the supervised knowledge to the *unseen* classes, which is an usual practice in ZSL methods. The main aim of our model is to learn two deep functions $G_{\text{sk}}(\cdot)$ and $G_{\text{im}}(\cdot)$ respectively for sketch and image for mapping them to a common semantic space where the learned knowledge can be applied to the *unseen* classes as well. Given a set of sketch-image data $\mathcal{D}^u = \{\mathbf{X}^u, \mathbf{Y}^u\}$ from the *unseen* categories $\mathcal{C}^u$ for test, the proposed deep functions $G_{\text{sk}} : \mathbb{R}^d \to \mathbb{R}^M$, $G_{\text{im}} : \mathbb{R}^d \to \mathbb{R}^M$ (d is the dimension of the original data and M is the targeted dimension of the common representation) map the sketch and natural image to a common semantic space where the retrieval is performed. Since the method considers SBIR in zero-shot setting, it is evident that the *seen* and *unseen* categories remain exclusive, *i.e.* $\mathcal{C}^s \cap \mathcal{C}^u = \varnothing$.

3.1. Paired Cycle Consistent Generative Model

For having the flexibility to handle sketch and image individually, *i.e.* even when they are not aligned sketch-image pairs, during training G_{sk} and G_{im}, we propose a cycle consistent generative model whose each branch is semantically aligned with a common discriminator. The cycle consistency constraint on each branch of the model ensures the mapping of sketch or image modality to a common semantic space, and their translation back to the original modality, which only requires supervision at category level. Imposing a classification loss on the output of G_{sk} and G_{im} allows generating highly discriminative features.

Our main goal is to learn two mappings G_{sk} and G_{im} that can respectively translate the unaligned sketch and natural image to a common semantic space. Zhu *et al.* [68] pointed out about the existence of underlying intrinsic relationship between modalities and domains, for example, sketch or image of same object category have the same semantic meaning, and possess that relationship. Even though, we lack visual supervision as we do not have access to aligned pairs, we can exploit semantic supervision at category levels. We train a mapping $G_{\text{sk}} : \mathbf{X} \to \mathbf{S}$ so that $\hat{\mathbf{s}}_i = G_{\text{sk}}(\mathbf{x}_i)$, where $\mathbf{s}_i \in \mathbf{S}$ is the corresponding side information and is indistinguishable from $\hat{\mathbf{s}}_i$ via an adversarial training that classifies $\hat{\mathbf{s}}_i$ different from $\mathbf{s}_i$. The optimal G_{sk} thereby translates the modality $\mathbf{X}$ into a modality $\hat{\mathbf{S}}$ which is identically distributed to $\mathbf{S}$. Similarly, another function $G_{\text{im}} : \mathbf{Y} \to \mathbf{S}$ can be trained via the same discriminator such that $\hat{\mathbf{s}}_i = G_{\text{im}}(\mathbf{y}_i)$.

Adversarial Loss. As shown in Figure 2, for mapping the sketch and image representation to a common semantic space, we introduce four generators $G_{\text{sk}} : \mathbf{X} \to \mathbf{S}$, $G_{\text{im}} : \mathbf{Y} \to \mathbf{S}$, $F_{\text{sk}} : \mathbf{S} \to \mathbf{X}$ and $F_{\text{im}} : \mathbf{S} \to \mathbf{Y}$. In addition, we bring in three adversarial discriminators: $D_{\text{se}}(\cdot)$, $D_{\text{sk}}(\cdot)$ and $D_{\text{im}}(\cdot)$, where D_{se} discriminates among original side information $\{\mathbf{s}\}$, sketch transformed to side information $\{G_{\text{sk}}(\mathbf{x})\}$ and image transformed to side information $\{G_{\text{im}}(\mathbf{y})\}$; likewise D_{sk} discriminates between original sketch representation $\{\mathbf{x}\}$ and side information transformed to sketch representation $\{F_{\text{sk}}(\mathbf{s})\}$; in a similar way D_{im} distinguishes between $\{\mathbf{y}\}$ and $\{F_{\text{im}}(\mathbf{s})\}$. For the generators G_{sk}, G_{im} and their common discriminator D_{se}, the objective is as follows:

$$\mathcal{L}_{\text{adv}}(G_{\text{sk}}, G_{\text{im}}, D_{\text{se}}, \mathbf{x}, \mathbf{y}, \mathbf{s}) = 2 \times \mathbb{E}\left[\log D_{\text{se}}(\mathbf{s})\right] \quad (1)$$
$$+ \mathbb{E}\left[\log(1 - D_{\text{se}}(G_{\text{sk}}(\mathbf{x})))\right] + \mathbb{E}\left[\log(1 - D_{\text{se}}(G_{\text{im}}(\mathbf{y})))\right]$$

where G_{sk} and G_{im} generate side information similar to the ones in $\mathbf{S}$ while D_{se} distinguishes between the generated and original side information. Here, G_{sk} and G_{im} minimize the objective against an opponent D_{se} that tries to maximize it, *i.e.* $\min_{G_{\text{sk}}, G_{\text{im}}} \max_{D_{\text{se}}} \mathcal{L}_{\text{adv}}(G_{\text{sk}}, G_{\text{im}}, D_{\text{se}}, \mathbf{x}, \mathbf{y}, \mathbf{s})$. In a similar way, for the generator F_{sk} and its discriminator D_{sk}, the objective is:

$$\mathcal{L}_{\text{adv}}(F_{\text{sk}}, D_{\text{sk}}, \mathbf{x}, \mathbf{s}) = \mathbb{E}\left[\log D_{\text{sk}}(\mathbf{x})\right] \\ + \mathbb{E}\left[\log(1 - D_{\text{sk}}(F_{\text{sk}}(\mathbf{s})))\right] \quad (2)$$
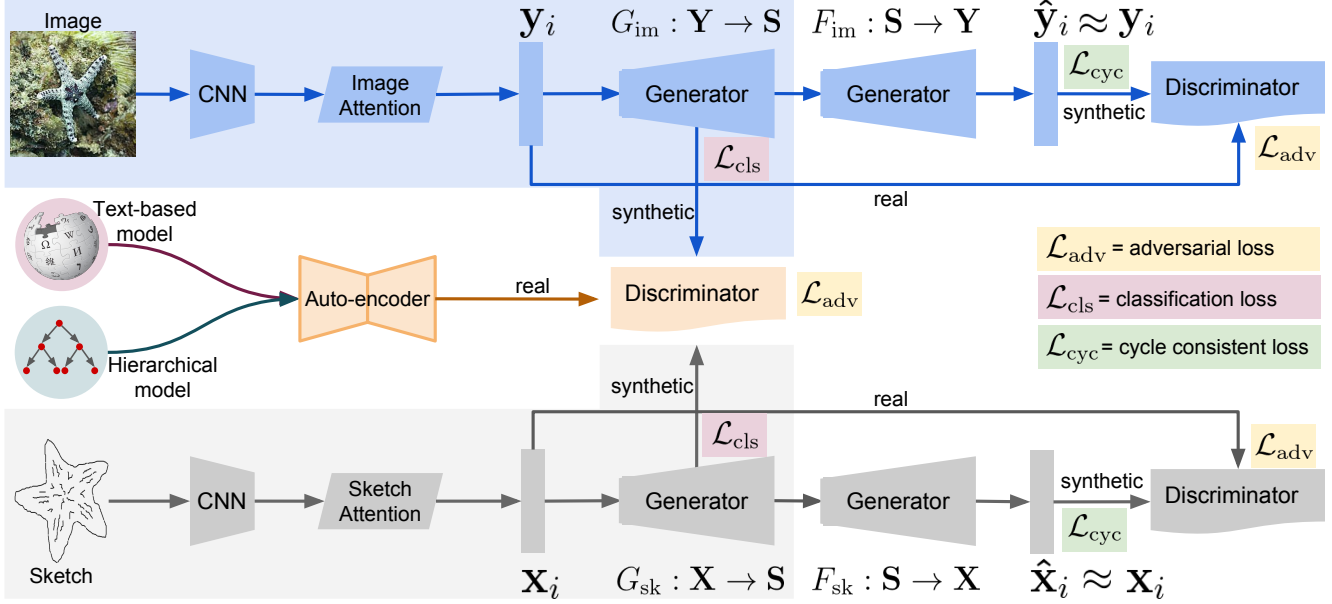
Figure 2. The deep network structure of SEM-PCYC. The sketch (in light gray) and image cycle consistent networks (in light blue) respectively map the sketch and image to the semantic space and then the original input space. An auto-encoder (light orange) combines the semantic information based on text and hierarchical model, and produces a compressed semantic representation which acts as a true example to the discriminator. During the test phase only the learned sketch (light gray region) and image (light blue region) encoders to the semantic space are used for generating embeddings on the *unseen* classes for zero-shot SBIR. (best viewed in color)

F_{sk} minimizes the objective and its adversary D_{sk} intends to maximize it, *i.e.* $\min_{F_{sk}} \max_{D_{sk}} \mathcal{L}_{adv}(F_{sk}, D_{sk}, \mathbf{x}, \mathbf{s})$. Similarly, another adversarial loss is introduced for the mapping F_{im} and its discriminator D_{im}, *i.e.*, $\min_{F_{im}} \max_{D_{im}} \mathcal{L}_{adv}(F_{im}, D_{im}, \mathbf{y}, \mathbf{s})$.

Cycle Consistency Loss. The adversarial mechanism effectively reduces the domain or modality gap, however, it is not guaranteed that an input $\mathbf{x}_i$ and an output $\mathbf{s}_i$ are matched well. To this end, we impose cycle consistency [68]. When we map the feature of a sketch of an object to the corresponding semantic space, and then further translate it back from the semantic space to the sketch feature space, we should reach back to the original sketch feature. This cycle consistency loss also assists in learning mappings across domains where paired or aligned examples are not available. Specifically, if we have a function $G_{sk} : \mathbf{X} \to \mathbf{S}$ and another mapping $F_{sk} : \mathbf{S} \to \mathbf{X}$, then both G_{sk} and F_{sk} are reverse of each other, and hence form a one-to-one correspondence or bijective mapping.

$$\mathcal{L}_{cyc}(G_{sk}, F_{sk}) = \mathbb{E}\left[\|F_{sk}(G_{sk}(\mathbf{x})) - \mathbf{x}\|_1\right] \\ + \mathbb{E}\left[\|G_{sk}(F_{sk}(\mathbf{s})) - \mathbf{s}\|_1\right] \quad (3)$$

Similarly, a cycle consistency loss is imposed for the mappings $G_{im} : \mathbf{Y} \to \mathbf{S}$ and $F_{im} : \mathbf{S} \to \mathbf{Y}$: $\mathcal{L}_{cyc}(G_{im}, F_{im})$. These consistent loss functions also behave as a regularizer to the adversarial training to assure that the learned function maps a specific input $\mathbf{x}_i$ to a desired output $\mathbf{s}_i$.

Classification Loss. On the other hand, adversarial training and cycle-consistency constraints do not explicitly ensure whether the generated features by the mappings G_{sk} and G_{im} are class discriminative, *i.e.* a requirement for the zero-shot sketch-based image retrieval task. We conjecture that this issue can be alleviated by introducing a discriminative classifier pre-trained on the input data. At this end we minimize a classification loss over the generated features.

$$\mathcal{L}_{cls}(G_{sk}) = -\mathbb{E}\left[\log P(c|G_{sk}(\mathbf{x}); \theta)\right] \quad (4)$$

where c is the category label of $\mathbf{x}$. Similarly, a classification loss $\mathcal{L}_{cls}(G_{im})$ is also imposed on the generator G_{im}.

3.2. Selection of Side Information

Motivated by attribute selection for zero-shot learning [18], indicating that a subset of discriminative attributes are more effective than the whole set of attributes for ZSL, we incorporate a joint learning framework integrating an auto-encoder to select side information. Let $\mathbf{s} \in \mathbb{R}^k$ be the side information with k as the original dimension. The loss function is:

$$\mathcal{L}_{aenc}(f, g) = \|\mathbf{s} - g(f(\mathbf{s}))\|_F^2 + \lambda \|W_1\|_{2,1} \quad (5)$$

where $f(\mathbf{s}) = \sigma(W_1\mathbf{s} + b_1)$, $g(f(\mathbf{s})) = \sigma(W_2 f(\mathbf{s}) + b_2)$, with $W_1 \in \mathbb{R}^{k \times m}$, $W_2 \in \mathbb{R}^{m \times k}$ and b_1, b_2 respectively as the weights and biases for the function f and g. Selecting side information reduces the dimensionality of embeddings,

which further improves retrieval time. Therefore, the training objective of our model:

$$\mathcal{L}(G_{sk}, G_{im}, F_{sk}, F_{im}, D_{se}, D_{sk}, D_{im}, f, g, \mathbf{x}, \mathbf{y}, \mathbf{s})$$
$$= \mathcal{L}_{adv}(G_{sk}, G_{im}, D_{se}, \mathbf{x}, \mathbf{y}, \mathbf{s}) + \mathcal{L}_{adv}(F_{sk}, D_{sk}, \mathbf{x}, \mathbf{s}) \quad (6)$$
$$+ \mathcal{L}_{adv}(F_{im}, D_{im}, \mathbf{y}, \mathbf{s}) + \mathcal{L}_{cyc}(G_{sk}, F_{sk}) + \mathcal{L}_{cyc}(G_{im}, F_{im})$$
$$+ \mathcal{L}_{cls}(G_{sk}) + \mathcal{L}_{cls}(G_{im}) + \mathcal{L}_{aenc}(f, g)$$

For obtaining the initial side information, we combine a text-based and a hierarchical model, which are complementary and robust [3]. Below, we provide a description of our text-based and hierarchical models for side information.

Text-based Model. We use two different text-based side information. (1) Word2Vec [31] is a two layered neural network that are trained to reconstruct linguistic contexts of words. During training, it takes a large corpus of text and creates a vector space of several hundred dimensions, with each unique word being assigned to a corresponding vector in that space. The model can be trained with a hierarchical softmax with either skip-gram or continuous bag-of-words formulation for target prediction. (2) GloVe [35] considers global word-word co-occurrence statistics that frequently appear in a corpus. Intuitively, co-occurrence statistics encode important semantic information. The objective is to learn word vectors such that their dot product equals to the probability of their co-occurrence.

Hierarchical Model. Semantic similarity between words can also be approximated by measuring their distance in a large ontology such as WordNet[1] of $\approx 100,000$ words in English. One can measure similarity using techniques such as path similarity and Jiang-Conrath [22]. For a set $\mathbb{S}$ of nodes in a dictionary $\mathbb{D}$, similarities between every class c and all the other nodes in $\mathbb{S}$ determine the entries of the class embedding vector [3]. $\mathbb{S}$ considers all the nodes on the path from each node in $\mathbb{D}$ to its highest level ancestor. The database of WordNet contains most of the classes of the Sketchy [42] and Tu-Berlin [13] datasets. Few exceptions are: *jack-o-lantern* which we replaced with *lantern* that appears higher in the hierarchy, similarly *human skeleton* with *skeleton*, and *octopus* with *octopods* etc. $|\mathbb{S}|$ for Sketchy and TU-Berlin datasets are respectively 354 and 664.

4. Experiments

Datasets. We experimentally validate our model on two popular SBIR benchmarks: Sketchy [42] and TU-Berlin [13], together with the extended images from [27].

The Sketchy Dataset [42] (Extended) is a large collection of sketch-photo pairs. The dataset consists of images from 125 different classes, with 100 photos each. Sketch images of the objects that appear in these 12, 500 images

are collected via crowd sourcing, which resulted in 75, 471 sketches. This dataset also contains a fine grained correspondence (aligned) between particular photos and sketches as well as various data augmentations for deep learning based methods. Liu *et al*. [27] extended the dataset by adding 60, 502 photos yielding in total 73, 002 images. We randomly pick 25 classes of sketches and images as the *unseen* test set for the zero-shot SBIR, and the data from remaining 100 *seen* classes are used for training.

The TU-Berlin Dataset [13] (Extended) contains 250 categories with a total of 20, 000 sketches extended by [27] with natural images corresponding to the sketch classes with a total size of 204, 489. 30 classes of sketches and images are randomly chosen to respectively form the query set and the retrieval gallery. The remaining 220 classes are utilized for training. We follow Shen *et al*. [43] and select classes with at least 400 images in the test set.

Implementation Details. We implemented the SEM-PCYC model using PyTorch [34] deep learning toolbox[2], which is trainable on a single TITAN Xp graphics card. We extract features from sketch and image from the VGG-16 [44] network model pre-trained on ImageNet [10] dataset (before the last pooling layer). Since in this work, we deal with single object retrieval and an object usually spans only on certain regions of a sketch or image, we apply an attention mechanism inspired by Song *et al*. [47] without the shortcut connection for extracting only the informative regions from sketch and image. The attended 512-D representation is obtained by a pooling operation guided by the attention model and fully connected (fc) layer. This entire model is fine tuned on our training set (100 classes for Sketchy and 220 classes for TU-Berlin). Both the generators G_{sk} and G_{im} are built with a fc layer followed by a ReLU non-linearity that accept 512-D vector and output M-D representation, whereas, the generators F_{sk} and F_{im} take M-D features and produce 512-D vector. Accordingly, all discriminators are designed to take the output of respective generators and produce a single dimensional output. The auto-encoder is designed by stacking two non-linear fc layers respectively as encoder and decoder for obtaining a compressed and encoded representation of dimension M.

While constructing the hierarchy for acquiring the class embedding, we only consider the *seen* classes belong to that dataset. In this way, the WordNet hierarchy or the knowledge graph for the Sketchy and TU-Berlin datasets respectively contain 354 and 664 nodes. Although our method does not produce binary hash code as a final representation for matching sketch and image, for the sake of comparison with some related works, such as, ZSH [56], ZSIH [43], GDH [63], that produce hash codes, we have used the iterative quantization (ITQ) [17] algorithm to obtain the binary

---

	Method	Sketchy (Extended)				TU-Berlin (Extended)			
		mAP @all	Precision @100	Feature Dimension	Retrieval Time (s)	mAP @all	Precision @100	Feature Dimension	Retrieval Time (s)
SBIR	Softmax Baseline	0.114	0.172	4096	3.5×10^{-1}	0.089	0.143	4096	4.3×10^{-1}
	Siamese CNN [36]	0.132	0.175	64	5.7×10^{-3}	0.109	0.141	64	5.9×10^{-3}
	SaN [59]	0.115	0.125	512	4.8×10^{-2}	0.089	0.108	512	5.5×10^{-2}
	GN Triplet [42]	0.204	0.296	1024	9.1×10^{-1}	0.175	0.253	1024	1.9×10^{-1}
	3D Shape [50]	0.067	0.078	64	7.8×10^{-3}	0.054	0.067	64	7.2×10^{-3}
	DSH (binary) [27]	0.171	0.231	64	6.1×10^{-5}	0.129	0.189	64	7.2×10^{-5}
	GDH (binary) [63]	0.187	0.259	64	7.8×10^{-5}	0.135	0.212	64	9.6×10^{-5}
ZSL	CMT [45]	0.087	0.102	300	2.8×10^{-2}	0.062	0.078	300	3.3×10^{-2}
	DeViSE [15]	0.067	0.077	300	3.6×10^{-2}	0.059	0.071	300	3.2×10^{-2}
	SSE [65]	0.116	0.161	100	1.3×10^{-2}	0.089	0.121	220	1.7×10^{-2}
	JLSE [67]	0.131	0.185	100	1.5×10^{-2}	0.109	0.155	220	1.4×10^{-2}
	SAE [24]	0.216	0.293	300	2.9×10^{-2}	0.167	0.221	300	3.2×10^{-2}
	FRWGAN [14]	0.127	0.169	512	3.2×10^{-2}	0.110	0.157	512	3.9×10^{-2}
	ZSH (binary) [57]	0.159	0.214	64	5.9×10^{-5}	0.141	0.177	64	7.6×10^{-5}
Zero-Shot SBIR	ZSIH (binary) [43]	0.258	0.342	64	6.7×10^{-5}	0.223	0.294	64	7.7×10^{-5}
	ZS-SBIR [23]	0.196	0.284	1024	9.6×10^{-2}	0.005	0.001	1024	1.2×10^{-1}
	SEM-PCYC	**0.349**	**0.463**	64	1.7×10^{-3}	**0.297**	**0.426**	64	1.9×10^{-3}
	SEM-PCYC (binary)	**0.344**	**0.399**	64	9.5×10^{-5}	**0.293**	**0.392**	64	9.3×10^{-4}
Generalized Zero-Shot SBIR	ZSIH (binary) [43]	0.219	0.296	64	6.7×10^{-5}	0.142	0.218	64	7.7×10^{-5}
	SEM-PCYC	**0.307**	**0.364**	64	1.7×10^{-3}	**0.192**	**0.298**	64	2.0×10^{-3}
	SEM-PCYC (binary)	**0.260**	**0.317**	64	9.4×10^{-5}	**0.174**	**0.267**	64	9.3×10^{-4}

Table 1. Zero-shot sketch-based image retrieval performance comparison with existing SBIR, ZSL, zero-shot SBIR and generalized zero-shot SBIR methods. Note: SBIR and ZSL methods are adapted to the Zero-Shot SBIR task, same *seen* and *unseen* classes are used for a fair comparison.
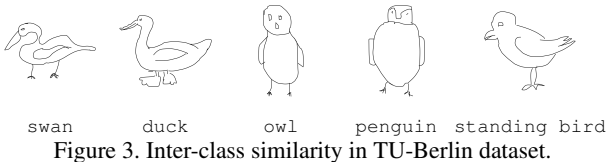

swan   duck   owl   penguin standing bird
Figure 3. Inter-class similarity in TU-Berlin dataset.

codes for sketch and image. We have used final representation of sketches and images from the train set to learn the optimized rotation which later used on our final representation for obtaining the binary codes.

4.1. Comparing with the State-of-the-Art

Apart from the two prior Zero-Shot SBIR works closest to ours, *i.e.* ZSIH [43] and ZS-SBIR [23], we adopt fourteen ZSL and SBIR models to the zero-shot SBIR task. The SBIR methods that we evaluate are SaN [60], 3D Shape [49], Siamese CNN [36], GN Triplet [42], DSH [27] and GDH [63]. A softmax baseline is also added, which is based on computing the 4096-D VGG-16 [44] feature vector pre-trained on the *seen* classes for nearest neighbour search. The ZSL methods that we evaluate are: CMT [45], DeViSE [15], SSE [66], JLSE [67], ZSH [56], SAE [24] and FRWGAN [14]. We use the same *seen-unseen* splits of categories for all the experiments for a fair comparison. We compute the mean average precision (mAP@all) and precision considering top 100 (Precision@100) [48, 43] retrievals for the performance evaluation and comparison.

Table 1 shows that most of the SBIR and ZSL methods perform worse than the zero-shot SBIR methods. Among them, the ZSL methods usually suffer from the domain gap that exist between the sketch and image modalities while SAE [24] reaches the best performance. The majority SBIR methods although have performed better than their ZSL counterparts, sustain the incapacity to generalize the learned representations to *unseen* classes. However, GN Triplet [42], DSH [27], GDH [63] have shown reasonable potential to generalize information only from object with common shape. As per the expectation, the specialized zero-shot SBIR methods have surpassed most of the ZSL and SBIR baselines as they possess both the ability of reducing the domain gap and generalizing the learned information for the *unseen* classes. ZS-SBIR learns to generalize between sketch and image from the aligned sketch-image pairs, as a result it performs well on the Sketchy dataset, but not on the TU-Berlin dataset, as in this case, aligned sketch-image pairs are not available. Our proposed method has consistently excelled the state-of-the-art method by 0.091 mAP@all on the Sketchy dataset and 0.074 mAP@all on the TU-Berlin dataset, which shows the effectiveness of our proposed SEM-PCYC model which gets benefited from (1) cycle consistency between sketch, image and semantic space, (2) compact and selected side information. In general, all the methods considered in Table 1 have performed worse on the TU-Berlin dataset, which might be due to the large number of classes, where many of them are visually similar and overlapping. These results are encouraging in that they show that the cycle consistency helps zero-shot SBIR task and our model sets the new state-of-the-art in this domain. The PR-curves of SEM-PCYC and considered baselines on Sketchy and TU-Berlin are respectively shown in Figure 5(a)-(b). We also conducted additional experi-

6

Figure 4. Top-10 zero-shot SBIR results obtained by our SEM-PCYC model on Sketchy (top four rows) and TU-Berlin (next four rows) are shown here according to the Euclidean distances, where the green ticks denote correctly retrieved candidates and the red crosses indicate wrong retrievals. (best viewed in color)

ments on generalized ZS-SBIR setting where search space contains *seen* and *unseen* classes. This task is significantly more challenging than ZS-SBIR as *seen* classes create distraction to the test queries. Our results in Table 1 (last two lines) show that our model significantly outperforms [43], due to the benefit of our cross-modal adversarial mechanism and heterogeneous side information.

Qualitative Results. Next, we analyze the retrieval performance of our proposed model qualitatively in Figure 4 (more qualitative results are available in [12]). Some notable examples are as follows. Sketch query of `tank` retrieves some examples of `motorcycle` probably because both of them have wheels in common. For having visual and semantic similarity, sketching `guitar` retrieves some `violins`. Querying `castle`, retrieves images having large portion of sky, because the images of its semantically similar classes, such as, `skyscraper`, `church`, are mostly captured with sky in background. In general, we observe that the wrongly retrieved candidates mostly have a closer visual and semantic relevance with the queried ones. This effect is more prominent in TU-Berlin dataset, which may be due to the inter-class similarity of sketches between different classes. As shown in Figure 3, the classes `swan`, `duck` and `owl`, `penguin` have substantial visual similar-

Text Embedding		Hierarchical Embedding			Sketchy	TU-Berlin
Glove	Word2Vec	Path	Lin [26]	Ji-Cn [22]	(Extended)	(Extended)
✓					0.284	0.228
	✓				0.330	0.232
		✓			0.314	0.224
			✓		0.248	0.169
				✓	0.308	0.227
✓		✓			0.338	0.276
✓			✓		0.299	0.253
✓				✓	0.285	0.243
	✓	✓			0.340	**0.297**
	✓		✓		0.288	0.264
	✓			✓	**0.349**	0.291

Table 2. Zero-shot SBIR mAP@all using different semantic embeddings (top) and their combinations (bottom).

ity, and all of them are `standing bird` which is a separate class of the same dataset. Therefore, for TU-Berlin dataset, it is challenging to generalize the *unseen* classes from the learned representation of *seen* classes.

4.2. Effect of Side-Information

In zero-shot learning, side information is as important as the visual information as it is the only means the model can discover similarities between classes. As the type of side information has a high effect in performance of any method, we analyze the effect of side-information and present zero-shot SBIR results by considering different side information and their combinations. We compare the effect of using GloVe [35] and Word2Vec [30] as text-based model, and three similarity measurements, i.e. path, Lin [26] and Jiang-Conrath [22] for constructing three different side information that are based on WordNet hierarchy. Table 2 contains the quantitative results on both Sketchy and TU-Berlin datasets with different side information mentioned and their combinations, where we set $M = 64$ (results with $M = 32, 128$ can be found in [12]). We have observed that in majority of cases combining different side information increases the performance by 1% to 3%.

On Sketchy, the combination of Word2vec and Jiang-Conrath hierarchical similarity reaches the highest mAP of 0.349 while on TU Berlin dataset, the combination of Word2Vec and path similarity leads with 0.297 mAP. We conclude from these experiments that indeed text-based and hierarchy-based class embeddings are complementary. Furthermore, Word2Vec captures semantic similarity between words better than GloVe for the task of zero-shot SBIR.

4.3. Model Ablations

The baselines of our ablation study are built by modifying some parts of the SEM-PCYC model and analyze the effect of different losses of our model. First, we train the model only with adversarial loss, and then alternatively add cycle consistency and classification loss for the training. Second, we train the model without the side information selection mechanism, for that, we only take the original text or hierarchical embedding or their combination as side
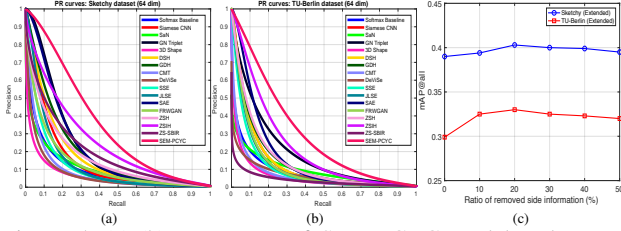
Figure 5. (a)-(b) PR curves of SEM-PCYC model and several SBIR, ZSL and zero-shot SBIR methods respectively on the Sketchy and TU-Berlin datasets, (c) Plot showing mAP@all wrt the ratio of removed side information. (best viewed in color)

Description	Sketchy	TU-Berlin
Only adversarial loss	0.128	0.109
Adversarial + cycle consistency loss	0.147	0.131
Adversarial + classification loss	0.140	0.127
Without selecting side information	0.382	0.299
Without regularizer in eqn. (5)	0.323	0.273
SEM-PCYC (full model)	**0.349**	**0.297**

Table 3. Ablation study on our 64-D model mAP@all results of several baselines are shown above.

information, which can give an idea on the advantage of selecting side information via the auto-encoder. Next, we experiment reducing the dimensionality of the class embedding to a percentage of the full dimensionality. Finally, to demonstrate the effectiveness of the regularizer used in the auto-encoder for selecting discriminative side information, we experiment by making $\lambda = 0$ in eqn. (5).

The mAP@all values obtained by respective baselines mentioned above are shown in Table 3. We consider the best side information setting according to Table 2 depending on the dataset. The assessed baselines have typically underperform the full SEM-PCYC model. Only with adversarial losses, the performance of our system drops significantly. We suspect that only adversarial training although maps sketch and image input to a semantic space, there is no guarantee that sketch-image pairs of same category are matched. This is because adversarial training only ensures the mapping of input modality to target modality that matches its empirical distribution [68], but does not guarantee an individual input and output are paired up. Imposition of cycle-consistency constraint ensures the one-to-one correspondence of sketch-image categories. However, the performance of our system does not improve substantially while the model is trained both with adversarial and cycle consistency loss. We speculate that this issue could be due to the lack of inter-category discriminating power of the learned embedding functions; for that, we set a classification criteria to train discriminating cross-modal embedding functions. We further observe that only imposing classification criteria together with adversarial loss, neither improves the retrieval results. We conjecture that in this case the learned embedding could be very discriminative but the two modalities might be matched in wrong way. Hence, it can

be concluded that all these three losses are complimentary to each other and absolutely essential for effective zero-shot SBIR. Next, we analyze the effect of side information and observe that without the encoded and compact side information, we achieve better mAP@all with a compromise on retrieval time, as the original dimension ($354 + 300 = 654d$ for Sketchy and $664 + 300 = 964d$ for TU-Berlin) of considered side information is much higher than the encoded ones (64d). We further investigate by reducing its dimension as a percentage of the original one (see Figure 5(c)), and we have observed that at the beginning, reducing a small part (mostly 5% to 30%) usually leads to a better performance, which reveals that not all the side information are necessary for effective zero-shot SBIR and some of them are even harmful. In fact, the first removed ones have low information content, and can be regarded as noise. We have also perceived that removing more side information (beyond 20% to 40%) deteriorates the performance of the system, which is quite justifiable because the compressing mechanism of auto-encoder progressively removes important and predictable side information. However, it can be observed that with highly compressed side information as well, our model provides a very good deal with performance and retrieval time. Without using the regularizer in eqn. (5), although our system performs reasonably, the mAP@all value is still lower than the best obtained performance. We explain this as a benefit of using ℓ_{21}-norm based regularizer that effectively select representative side information.

5. Conclusion

We proposed the SEM-PCYC model for the zero-shot SBIR task. Our SEM-PCYC is a semantically aligned paired cycle consistent generative model whose each branch either maps a sketch or an image to a common semantic space via adversarial training with a shared discriminator. Thanks to cycle consistency on both the branches our model does not require aligned sketch-image pairs. Moreover, it acts as a regularizer in the adversarial training. The classification losses on the generators guarantee the features to be discriminative. We show that combining heterogeneous side information through an auto-encoder, which encodes a compact side information useful for adversarial training, is effective. Our evaluation on two datasets has shown that our model consistently outperforms the existing methods in zero-shot SBIR task.

Acknowledgments

References

[1] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele. Multi-cue zero-shot learning with strong supervision. In *CVPR*, pages 59–68, 2016. 2

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE TPAMI*, 38(7):1425–1438, 2016. 2

[3] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 2, 5

[4] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*, pages 5975–5984, 2016. 2

[5] S. Changpinyo, W. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016. 2

[6] S. Changpinyo, W. Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*, pages 3496–3505, 2017. 2

[7] J. Chen and Y. Fang. Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In *ECCV*, pages 624–640, 2018. 1

[8] L. Chen, H. Zhang, J. Xiao, W. Liu, and S. Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, pages 1043–1052, 2018. 2

[9] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pages 539–546, 2005. 2

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248–255, 2009. 5

[11] Z. Ding, M. Shao, and Y. Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *CVPR*, pages 6005–6013, 2017. 2

[12] A. Dutta and Z. Akata. Supplementary material: Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019. 7

[13] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM TG*, 31(4):1–10, 2012. 5

[14] R. Felix, V. B. G. Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018. 6

[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 2, 6

[16] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, pages 2635–2644, 2015. 2

[17] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE TPAMI*, 35(12):2916–2929, 2013. 5

[18] Y. Guo, G. Ding, J. Han, and S. Tang. Zero-shot learning with attribute selection. In *AAAI*, pages 6870–6877, 2018. 4

[19] G. Hu, Y. Hua, Y. Yuan, Z. Zhang, Z. Lu, S. S. Mukherjee, T. M. Hospedales, N. M. Robertson, and Y. Yang. Attribute-enhanced face recognition with neural tensor fusion networks. In *ICCV*, pages 3764–3773, 2017. 1

[20] R. Hu and J. Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 117(7):790 – 806, 2013. 2

[21] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, pages 3464–3472, 2014. 2

[22] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING*, pages 19–33, 1997. 5, 7

[23] S. Kiran Yelamarthi, S. Krishna Reddy, A. Mishra, and A. Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, pages 316–333, 2018. 1, 3, 6

[24] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 4447–4456, 2017. 2, 6

[25] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2014. 1, 2

[26] D. Lin. An information-theoretic definition of similarity. In *ICML*, pages 296–304, 1998. 7

[27] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, pages 2298–2307, 2017. 1, 2, 5, 6

[28] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *CVPR*, pages 6165–6174, 2017. 2

[29] T. Mensink, E. Gavves, and C. G. M. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, pages 2441–2448, 2014. 2

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 2, 7

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013. 5

[32] G. A. Miller. Wordnet: A lexical database for english. *ACM*, 38(11):39–41, 1995. 2

[33] K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, pages 1–12, 2017. 1, 2

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017. 5

[35] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 5, 7

[36] Y. Qi, Y. Z. Song, H. Zhang, and J. Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, pages 2460–2464, 2016. 2, 6

[37] R. Qiao, L. Liu, C. Shen, and A. v. d. Hengel. Less is more: Zero-shot learning from online textual documents with noise suppression. In *CVPR*, pages 2249–2257, 2016. 2

[38] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016. 2

[39] B. Romera-Paredes and P. H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015. 2

[40] J. M. Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *ICIP*, pages 2998–3002, 2014. 2

[41] J. M. Saavedra and J. M. Barrios. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, pages 1–11, 2015. 2

[42] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM TOG*, 35(4):1–12, 2016. 2, 5, 6

[43] Y. Shen, L. Liu, F. Shen, and L. Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018. 1, 3, 5, 6, 7

[44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, abs/1409.1556, 2014. 5, 6

[45] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943, 2013. 2, 6

[46] J. Song, Y.-Z. Song, T. Xiang, and T. Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, pages 1–12, 2017. 2

[47] J. Song, Q. Yu, Y. Z. Song, T. Xiang, and T. M. Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, pages 5552–5561, 2017. 1, 2, 5

[48] W. Su, Y. Yuan, and M. Zhu. A relationship between the average precision and the area under the roc curve. In *ICTIR*, 2015. 6

[49] F. Wang, L. Kang, and Y. Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, pages 1875–1883, 2015. 6

[50] M. Wang, C. Wang, J. X. Yu, and J. Zhang. Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework. In *VLDB*, pages 998–1009, 2015. 6

[51] S. Wang, Z. Ding, and Y. Fu. Feature selection guided auto-encoder. In *AAAI*, pages 2725–2731, 2017. 2

[52] W. Wang, Y. Pu, V. K. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin. Zero-shot learning via class-conditioned deep generative models. In *AAAI*, 2018. 2

[53] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016. 2

[54] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, pages 1–14, 2018. 1, 2

[55] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 2

[56] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen. Zero-shot hashing via transferring supervised knowledge. In *ACM MM*, pages 1286–1295, 2016. 5, 6

[57] Z. Yang, W. W. Cohen, and R. Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *ICML*, pages 40–48, 2016. 6

[58] Q. Yu, F. Liu, Y. Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy. Sketch me that shoe. In *CVPR*, pages 799–807, 2016. 1, 2

[59] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Sketch-a-net: A deep neural network that beats humans. *IJCV*, pages 1–15, 2016. 6

[60] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales. Sketch-a-net that beats humans. In *BMVC*, pages 1–12, 2015. 1, 6

[61] T. Yu, J. Meng, and J. Yuan. Multi-view harmonized bilinear network for 3d object recognition. In *CVPR*, pages 186–194, 2018. 2

[62] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, pages 1839–1848, 2017. 1, 2

[63] J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. Tao Shen, and L. Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, pages 304–321, 2018. 1, 2, 5, 6

[64] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, pages 3010–3019, 2017. 2

[65] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE TIP*, 24(12):4766–4779, 2015. 6

[66] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015. 2, 6

[67] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, pages 6034–6042, 2016. 2, 6

[68] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017. 3, 4, 8